Universität Zürich[UZH]

Master Thesis

An Empirical Analysis of the Formation of Sport Preferences in Switzerland; with a Focus on Inter- and Intragenerational Factors.

University of Zurich

Ich bewerte die Arbeit mit der Note
5.25

Submitted by:	Joffrey Anthony Mayer-Boutter
Matriculation Number:	13-757-034
Address:	Treppenweg 4 8634 Hombrechtikon
E-Mail:	joffreyanthony.mayer-boutter@uzh.ch
Date:	13.10.2020

Fall Semester 2020

Abstract

Today's western world, but - recently - also less developed countries all around the world are struggling with a significant increase in the proportion of overweight individuals (World Health Organization, 2020b). Switzerland is no exception and is trying to fight against these unhealthy habits through preventive measures, including exercise and sports promotion programs (Bundesamt für Gesundheit, 2018). The goal of this thesis is to investigate different reasons why individuals may or may not be motivated to do sport over their whole life, or lifecycle. By using modern discrete choice methods and a rich individual-level based Swiss dataset questioning entire families, this thesis finds that the belief of maintaining a better health-status in the future, as well as having chronical health problems from being overweight or obese motivate individual to not quit doing sport, whereas financial factors or other substitute leisure activities incline individuals to stop doing sport or to not even consider to start doing sport in their life, respectively. Furthermore - by looking at the dynamics of sport-preferences - this thesis finds evidence for a time-invariant intergenerational transmission-process.

Contents

List of Figures and Tables

List of Abbreviations

ASD	Absolute Standardized Difference
ATE	Average Treatment Effect
ATT	Average Treatment Effect on the Treated
BAG	Bundesamt für Gesundheit
BMI	Body Mass Index
CLM	Conditional Logit Model
DUM	Discounted Utility Model
HFH	Habit-Formation Hypothesis
HIH	Health-Investment Hypothesis
iid	Identically and Independently Distributed
LHS	Left-Hand Side
MNL	Multinomial Logit Model
OLS	Ordinary Least-Squared
PRF	Population Regression Function
RHS	Right-Hand Side
RUT	Random Utility Theory
RUM	Random Utility Model
SHP	Swiss Household Panel
SRF	Sample Regression Function
WARP	Weak Axiom of Revealed Preferences

WHO	World Health Organization

1. Introduction

According to recent panels, 14.8% of the female and 25.4% of the male population between 15 and 24 years in Switzerland is overweight or obese (Bundesamt für Statistik, 2018). However, not only Switzerland, but many other well-developed countries have seen an increase in the fraction of people with weight problems, with the United States as the most well-known country with 69% of its adult population being overweight or obese (Flegal et al., 2012). As the Bundesamt für Gesundheit (BAG, 2018) notes, these tendencies of an unhealthy lifestyle can lead to diverse forms of so called *non-communicable diseases* (NCDs) such as diabetes, cardiovascular diseases, difficulty in breathing or cancer. Nowadays, NCDs are the most common cause of death in our society and are responsible for about 70-80 % of direct health care costs (European Commission, 2014). To make things even worse, the Swiss business cycle research center (Köthenbürger & Anderes, 2019) estimates that the costs of health care should continue to rise in the following years, which is in line with the assumptions of the BAG (2018) that chronic diseases will continue to rise in the coming decades as a result of demographic developments and the ageing of the baby boom generation.

Seeing the growing *economic* but also health-related importance of NCDs, researchers began to attribute the cause of unhealthy lifestyles to the human behavior (Lopez et al., 2006). Based on recent findings, the Swiss government started to elaborate long-term policies in 2013 which targeted the attitudes towards health of parents and - especially - children, in order to initiate individuals into doing more physical activities on a regular basis to prevent future trends in obesity and, thus, lower the cases of NCDs. Nevertheless, even if the main objective is to slowly alter an individual's preferences towards a healthier lifestyle by motivating a person to do to more sports (BAG, 2018), empirical evidence on the formation of sport preferences remained rather rare before the start of the 20th century according to Woolger and Power (1993). Recently, there are studies that investigate the formation on sport preferences *empirically*, however focusing more on the fact that an individual stays physically active rather than specifically looking for the act of doing sport (Humphreys & Ruseski, 2011; Janke et al., 2013; Johannesson et al., 2010). Yet, whenever the literature addresses the theme of sport participation, then in a more *static* way, by analyzing sport as a consumption good, which means that those studies look at the direct effect of sport on utility (Cabane & Lechner, 2015). To my knowledge, nothing is known about the dynamic nature of sports preference. Questions such as what motivates people to stay sportively active *over their whole life*, while others prefer to be in a less active state, by only starting sport in adulthood, stay unanswered. Therefore, an

analysis of sport participation over an individual's whole *lifecycle* by using modern discrete choice models to uncover factors that explain why individuals prefer a certain sport-state over another, as well as dynamically simulate the change in the sport-state, by altering the model's parameters, may help to better understand how individuals may end up doing sport during their entire life, while others choose to quit sport at an older age and how sensitive those equilibria are, when changing some particular model parameters.

An aspect one needs to keep in mind when using individuals' sport activity over their whole lifecycle as the unit of observation in an empirical analysis, is that preferences need to be treated as endogenous, since some individuals will experience a preference reversal in their life, meaning that they first will start doing sport in childhood and later - because their tastes towards sport changed - decide to do the opposite and quit sport (and vice versa for individuals who did not do sport in childhood). While neoclassical economists always treated preferences as a "blackbox", things started to change during the 1980s with the rise of behavioral economics (Dohmen et al., 2012; Laibson & Zeckhauser, 1998). Recently, a growing branch of this behavioral literature endogenize individuals' preference endowments in order to show that there are some individual traits that can be "inherited" not solely by genetics, but also by interacting socially with others, typically from one generation to the next (Angrist, 1990; Campante & Yanazingawa-Drott, 2015; Currie & Moretti, 2003; Fernández et al., 2004). Estimating the power of *attitude transmission* is shown to be from great importance when analyzing decisions individuals make, since it can have long-lasting impacts on a wide range of socioeconomic outcomes such as income (Chetty et al., 2014), educational attainment (Goodman et al., 2019), wealth (Black et al., 2015) or health (Johnston et al., 2013). Given the importance of the *intergenerational* transmission of preferences in many fields, this *social factor* will be taken into consideration when analyzing the formation of sport preferences over a lifecycle. However, recent discoveries in other areas of behavioral economics suggests that human behavior is - besides intergenerational attitude transmission - driven by a vast amount of *other psychological phenomena* such as temptation (Laibson, 1997) or a preference for improvement (Chapman, 2000), leading to - sometimes - conflicting choice behaviors within a lifecycle of an individual. Therefore, focusing only on the aspect of intergenerational transmission of sports preferences may be insufficient and other psychological factors may also play a role.

Behavioral scientists are not the only community that try to explain preferences. In a more traditional economical approach, neoclassical theorists put the emphasis on *financial motives*, by assuming that individuals either have an exogenous preference towards sport solely because they "like" to be sportively active or because the choice of sport will bring them indirect utility

by maintaining their health and resulting in higher productivity and, thus, gaining more money (Becker, 1965; Cabane & Lechner, 2015; Colman & Dave, 2013).

In this thesis, I aim to design a model of preference formation of sport that is able to capture (at least partially) what Samuelson's (1937) discounted utility model (DUM) - the most simple but famous and widely applied intertemporal choice model (Frederick et al., 2002) - would call "irrational behavior" over the lifecycle of an individual by including important social factors such as the intergenerational transmission in the estimation. The research question I would like to examine, is to assess the importance of specifically selected social- as well as financial-factors during the formation of sport preferences within a lifecycle of an individual in Switzerland. More specifically, by using a multinomial logit model (MNL), it will be possible to analyze and assess the *relative importance* of each of the identified factors and explain, why individuals might choose - for example - to stop doing sport later in their lives (Crowson, 2020; Long & Freese, 2001). Secondly, by differentiating between the parents' sport activity, it will be possible to determine whether sport-preference transmission is *time-invariant* - as the usual intergenerational literature predicts it to be - or heterogeneous, that is non-constant over time.

To answer the research question in depth, I will first identify relevant factors related to sport preferences, primarily using the health-, behavioral- and intergenerational-literature as the main aid for the theoretical foundations of my model (Chapter 2). The objective of this chapter will be to formulate two main hypotheses, which will be based on the predictions of already existing theoretical models. The formulation of these two hypotheses will serve the purpose of differentiating between different types of individuals: the ones motivated intrinsically - possibly forming habitual behavior because of social factors like their parents (*habit-formation hypothesis* (HFH)) - and others mainly motivated extrinsically and gaining a preference in sport only later in their life, possibly because of health (*health-investment hypothesis* (HIH)). This procedure will justify why some individuals may expose the previously mentioned "irrational behavior", that is a change in their preferences over the course of their lifecycle.

After the qualitative work, the next chapter will revolve around the problem of finding the appropriate method(s) to estimate a model and test the previously formulated hypotheses. Next, the data used for the analysis will be presented, accompanied with some descriptive statistics. In the result-chapter, the focus will lie on presenting my findings. The last chapter of my work will summarize my findings, whether my research question could be answered definitively, what implications my findings may have, some short discussion about the general validity of my work and if further research would be necessary.

2. Theoretical Foundations

When trying to elaborate a choice model, it is important to keep in mind that - in economics - there is a difference between the concepts of utility, preferences, and choices. They are the core components when it comes to explaining the *reason why* someone chooses a particular option, or alternative (Edwards, 1954). By exploring the literature and giving formal, but also concrete real-life examples, it will be shown that each of those three are distinct elements that play a key role when researchers try to think about an individual's preference formation and choices.

After this general introduction, the next subsection will revolve around finding determinants in the formation of sport preferences by discussing the findings of the already existing literature, but also by adding my own thoughts and intuitions. This review of the literature will culminate in Figure 5 in Appendix A, which will visualize the most relevant factors for sport preferences across time and generations. The theoretical foundations having been laid out, Chapter 2 will end with the formulation of my hypotheses, which will ultimately be tested in the empirical model.

2.1 Preferences, Choices and the Concept of Utility in Economics

Decisions, or choices, are recurrent components in the daily life of every human being. While choosing between a glass of water or a cup of tea may seem anecdotal in contrast to decisions such as career choices or the desire to have children, each action taken by individuals - regardless of the degree of awareness - is coupled to some factors that *drive* their choices, known as tastes, attitudes or preferences (Arrow, 1958; Samuelson, 1948).

In order to better understand the difference between the concepts of preference, choices and utility the use of formal language can help simplifying and visualizing the relationships. Following Mas-Colell et al. (1995, Chapter 1), let's establish a simple set of mutually exclusive alternatives X from which an individual has the following options to choose from during his free-time:

$$X: \{Sport, No\ Sport\} \tag{1}$$

Note that the alternative *no sport* can be seen as an outside option, which can take on - for example - various different other leisure activities representing the opportunity costs of sport (Kjær, 2005; Train, 2003). Next, let's establish the relationship between preferences and choices: an individual will compare - in a pairwise fashion - the different alternatives in the set

X and establish a so called (strict) preference relation - denoted as $\succ$ - which will *cause* an individual to choose one alternative over the other (Arrow, 1959; Hansson, 1968):

$$Sport \succ No\ Sport \Rightarrow C(X) = C(\{Sport, No\ Sport\}) = \{Sport\} \qquad (2)$$

Where the left-hand side (LHS) of the above implication can be read as the individual *sport is strictly better than the alternative no sport* (Debreu, 1954; Hansson, 1968). It is to note that preferences in (2) are totally subjective to any individual and can be seen as a result of an ordering, or cognitive trade-off - here, sport versus not sport - between which a person will pick the alternative that he / she "likes" better (Edwards, 1954; Houthakker, 1965).

On the right-hand side (RHS) of statement (2), we have a choice correspondence *C(·)*, which is nothing else than a function that gives - as an output - the choice of an individual and - in our case, letting our set *X* entering *C(·)* as an input - results in the individual choosing the alternative *sport*. This definition of the choice correspondence is key, since *C(·)* will give us a *strictly* smaller subset: $\underbrace{C(X)}_{Outcome\ of\ C(\cdot)} \subset X$ (Hansson 1968; Houthakker, 1965). Why is this important? Because it shows that a choice - generally speaking - is nothing else than a concept that reduces a set of alternatives to a smaller set of alternatives.

Next, we need to define how the concept of preferences and choices are linked to utility. In the above statement (2), what we did is only a description of how a person with a set of alternatives, might end up with a strictly smaller set. The way this is achieved, is by using the function *C(·)*, but we did *not* specify by any means how this is done (Arrow, 1959). There is not yet a behavioral rule that denotes after which criteria *C(·)* should establish which of the alternatives in *X* should be chosen by an individual. This is where microeconomic theory comes into play and - with the standard concept of utility maximization - is able to give an individual a specific decision rule after which he will be able to make his / her choice (Edwards, 1954). Translated in formal language (Mas-Colell et al., 1995):

$$C(X) = arg\ max_{x \in X}\ u(x) \qquad (3)$$

Where the RHS of equation (3) translates as the utility maximization model's prediction, e.g. that the consumer will *choose* the alternative *x* in set *X*, which will give him the highest utility *u(x)*.

Next, we can combine (2) and (3) together (Mas-Colell et al., 1995):

$$Sport \succ NoSport \Rightarrow C(X) \overset{Specify\ C(\cdot)}{\widehat{=}} argmax_{x \in X} u(x) =$$

$$= \{x = Sport \in X \mid u(Sport) > u(y) \, \forall \, y \in X\} \Rightarrow C(X) = \{Sport\} \quad (4)$$

The key point in the statement (4) that one needs to be able to grasp, is that specifying *C(·)* with the concept of utility maximization will serve as the economic justification, why a person chooses the alternative to do *sport*. Therefore, because the alternative sport brings relatively more utility to the individual than the option not to do sport, it will *cause* him to choose sport. In contrast - *before* introducing the utility concept - an individual may have just "liked" sport, but the reason was not justified by economic foundations (Edwards, 1954; Houthakker, 1965; Samuelson, 1947). Thus, when thinking about the formation of sport preferences, the utility concept needs to stay as the invisible link between preferences and choices (Houthakker, 1950).

2.1.1 The Discounted Utility Model – the Standard Model of Intertemporal Choice

Now that the general theoretical components of a decision have been formalized and explored, the next aspect that will be in this work's focus are decisions, which require a person to use their ability to project themselves into the future and simulate multiple cost and benefit analyzes simultaneously in order to make a choice today. This particular type of decision is defined as an *intertemporal choice* (Frederick et al., 2002). There are many examples which fall into this category, for instance, the decision on how much money to spend today in order to have more / less savings for future retirement (Thaler, 1990) or - in my context - the decision to do sport today to gain - possibly - more health in the future.

If we take a more conceptual approach - similar to the one in the previous subsection 2.1 - intertemporal choices are nothing else than the cognitive trade-off individuals face off, when comparing instantaneous and delayed (future) utility of an action before they choose the option with the highest utility-value (Becker & Mulligan, 1997; Koopmans, 1960):

$$\underbrace{Sport_{tomorrow} \succ Sport_{today}}_{(Time)\ preference\ relation} \Rightarrow u(Sport_{tomorrow}) > u(Sport_{today}) \Rightarrow C(X) = \{Sport_{tomorrow}\} \quad (5)$$

Economists often use the term *time preference* to describe and explain the decisions taken in intertemporal choice situations (Becker & Mulligan, 1997; Frederick et al., 2002). Already at the beginning of the 19th century Rae (1905, pp. 55-59) – in a reprinted version of his book from 1834 - speculated that different *distinct psychological aspects* play a key role in the *formation* of preferences over time in the context of saving money, such as the desire to leave behind an inheritance in form of a *bequest*, as well as the ability to exert *self-control*. Almost 100 years later, the famous economist Irving Fisher (1930) extended Rae's (1905) approach by stressing the importance of psychological determinants like *foresight*, *habits*, *life expectancy*,

as well as a *fondness for prosperity* ("fashion", p. 50) when dealing with intertemporal choices. To stick with our savings example, the reason why an individual may weight the instantaneous payoff of consuming more today relatively stronger as shown on the LHS of statement (5), and thus, choose to have a less comfortable retirement in the future, is because of not being able to commit to a regular saving, since the person has little self-control (Laibson, 1994).

Why is this relevant in my context? The key point here is that - depending on how time preferences are handled in theoretical models, e.g. how they handle the very LHS in statement (5) - one can make vastly opposing predictions about the intertemporal choices of individuals. As we will see, the model I am about to present places too restrictive assumptions on the LHS of (5), leading to unrealistic forecasts when a researcher is interested in *empirically* predicting repeated choices - such as the decision to do sport - over a lifecycle of an individual.

When dealing with intertemporal choices, for a long time, the so called DUM from Samuelson (1937) was considered to be *the standard model* (Frederick et al., 2002). The three core assumptions in this model are:

1. The discount-rate is considered to be constant over time[1]. At first glance, this assumption may seem anecdotal, however, it imposes the restriction on an individual's behavior that its time preference will always stay the same or be *time-consistent* (Frederick et al., 2002). If we take the LHS of statement (5) as a starting point, the DUM would predict that an individual's revealed (time) preference - as a child - would be $Sport_{tomorrow} \succ Sport_{today}$ and, therefore he / she would choose not to do sport *today*. As time passes and the individual grows adult, the initial preference relation is assumed to *stay* $Sport_{tomorrow} \succ Sport_{today}$, which would still lead to the prediction of doing no sport when the individual grows older.

2. It assumes that an individual is *perfectly rational* and has *complete information*, which implies that - when a person is faced with a number of alternatives to choose from - a person will always be able to process all the relevant information flawlessly and predict the most optimal choice that brings the highest life-time utility (Edwards, 1954).

3. The discount-function does not depend on the *type* of consumption. That is, the exact *same* discount-rate applies to *all* forms of consumption, which implies that there is not such a thing as - for example - a coffee-consumption time preference or water-consumption time preference (Frederick et al., 2002).

[1] This assumption is often referred to as *stationary discounting* (Frederick et al., 2002).

In the context of *intertemporal choice of doing sport*, the first assumption of the above list has great importance, because - if an empirical model would be based on the DUM - it would *not* be possible to explain why some individuals only start / stop doing sport later in their lives, that is, why those individuals have *time-inconsistent preferences*. However - as Strotz (1955) argues - this kind of behavior seems to be reasonable to encounter in reality. As a child, an individual may have the preference relation of $Sport_{tomorrow} \succ Sport_{today}$, but - over the course of life - this preference relation might be *reversed* to $Sport_{today} \succ Sport_{tomorrow}$. As a consequence, such an individual would experience a change of mind and act the *opposite* way as an adult relative to when he / she was a child. Nevertheless, the DUM would *never* be able to predict this kind of behavior, because of the premise of a constant discount-factor. Thus, we see that - way that the DUM was constructed - makes preferences *exogenous* / unchangeable over time (Becker & Mulligan, 1997).

As one can imagine, the implied behavior from the *other* mentioned assumptions seem not to be realistic either. In fact, every point in the above list has been proven to be problematic when being tested empirically (Frederick et al., 2002). To be more concrete, here are two examples - derived from the behavioral literature - to illustrate the issues that lie within the DUM in certain choice-situations:

- For instance, we can think of a heavy smoker that *chooses to smoke* each day, even though he might have a *preference not to smoke*, because he knows that it can worsen his future health condition. Following the logic of a perfectly rational agent with time-consistent preferences in the DUM, this individual's behavior cannot be explained, since the act to smoke *should* maximize his utility and, therefore should be in line with his preferences. However, since we assume that the smoker has a preference not to smoke, this behavior is out of reach of the DUM's predictions.

- Another example is a situation where a couple is asked what liquor to drink in a bar. Again - if we assume two perfectly rational agents - the drinks chosen will be those that maximize each of their utilities. In this case, let's assume that individual *A* likes Bourbon and individual *B* Vodka the most. However - since the bar is very crowdy - the waiter is in a hurry and - as a consequence - the couple is forced to choose faster than expected. As a result, only individual *A* was able to read through some of the list of beverages. While doing so, individual *A* spotted individual *B*'s favorite drink, but did not see whether the bar offered some Bourbon. Even though there were other drinks that

individual *A* liked more than Vodka, the couple finally ordered two Vodkas from the waiter.

As can be seen from those examples, there seem to be various psychological factors - such as the *temptation* of lighting a cigarette in the case of the addicted smoker (O'Donoghue & Rabin, 1999) and *the presence of peers* (Grossman, 2000; Kremer & Levy, 2008), as well as *restricted information-processing* (Simon, 1972) for the couple in the bar - that might influence the behavior of individuals, for which the standard neoclassical model of choice like the DUM is not suited and - as Rae (1905) or Fisher (1930) pointed out - may explain why preferences are *endogenous* (Becker & Mulligan, 1997). This insight might also play a role in the formation of sport preferences, as will be shown in the later subsection 2.2.

2.1.2 The Random Utility Model – a Framework for Empirical Analysis

As we have seen, the concept of the DUM might not be able to grasp in enough dept every facet of an individual's decision process, especially when it comes to an altering behavior - so called *preference reversals* (Lichtenstein & Slovic, 1971) - over the course of a lifetime, as the above examples in section 2.1.1 demonstrated and which will be the relevant unit of observation when evaluating my research question. Therefore, an alternative modelling approach which takes endogenous preferences into account needs to be taken.

Before jumping into this more suited framework, there is another difficulty to keep in mind when it comes to *empirically* analyzing what the drivers of a person's choices are, namely that - in the real world - the concept of preferences and utility outlined in the above subsections will always stay on a *theoretical* level. In the data, a researcher can only identify a person's choice, but is rarely able to get access to the underlying tastes of a person[2] or figure out the true utility function of individuals which - according to economic theory - are the real cause of a decision (Morgenstern & Von Neumann, 1953, p. 8).

During the first half of the 20th century, many economists treated tastes as a "blackbox" and left (endogenous) preferences to the field of psychology (Arrow, 1958; Dohmen et al., 2012; Sen, 1973). Economists used preference-theories more as an analytical tool that stayed on a conceptual-level. Whenever it was *empirically* applied, then on *aggregated*- and market-level data, with a *representative agent* as the baseline. If results from such practical studies were found to go against the behavior of a representative agent[3], it was argued to be measurement

[2] There are some datasets which sometimes uncover so called *stated preferences*, that is questions about what a person might hypothetically choose as a particular option (Kjær, 2005).
[3] For example, *time-inconsistent* behavior, as demonstrated in the previous subsection 2.1.1.

error, rather to put the emphasis on some unobservable characteristics from individuals. However, with the rise of computational power, as well as *individual-level data*, things started to change in the 1970s (McFadden, 2001). During this time, Nobel-Prize laureate Daniel McFadden and other researchers - on the basis of preliminary work from Thurstone (1927) and Marschak (1960) - were able to come up with a *probabilistic approach* to empirically model an individual's utility function. The basic idea is to model the decision of an individual - in my case, whether an individual did sport or not - as a binary-dependent variable in order to approximate the (true) underlying utility function. Following StataCorp (2019, Intro 8):

$$U_{ij} = V(X_{ij}, \beta) + \varepsilon_{ij}, \tag{6}$$

$$Y_i = arg\,max_{\,j \in \{1,\dots,Z\}} U_{ij} \tag{7}$$

This is the random utility model (RUM) framework (Manski, 1977). Equation (6) can be thought of the population regression function (PRF) that researchers will try to estimate, where:

- U_{ij} reflects the *true* utility an individual *i* would get by choosing the particular alternative *j* within a set of alternatives *Z* (McFadden, 1974). In the RUM-framework, this is the variable of interest that we would like to estimate.
- $V(X_{ij}, \beta)$ represents the deterministic portion of the utility, that the researcher is able to observe for individual *i* for any of the possible alternatives *j* within a set *Z*. More precisely, *V(·)* is defined as a function of the observable vector of characteristics X_{ij} and parameters β (Cameron & Trivedi, 2005). For simplicity, I will denote the deterministic portion of the utility with V_{ij} for the remaining of this subsection.
- ε_{ij} reflects the error-term, which contains the random part of the utility which the researcher *cannot* observe (Cameron & Trivedi, 2005; Train, 2003).
- In equation (7), Y_i represents the choices that we observe in the data (Cameron & Trivedi, 2005). By assuming that the decision rule of an individual is the concept of utility maximization, the choice of an individual will be justified, since the individual will pick the alternative that brings the highest (true) utility U_{ij} (Walker & Ben-Akiva, 2002). To draw the parallel to the theoretical part in subsection 2.1, note that this is just the specification of the choice correspondence *C(·)*.

Depending on how we will specify and estimate our RUM, the deterministic / observable portion of the utility function V_{ij} can be disaggregated even further down, typically up until 3 different components (Cameron & Trivedi, 2005; Saberi, 2017):

$$V_{ij} = V(S_j, \beta) + V(X_i, \beta_j) + V(S_j, X_i, \alpha) \quad (8)$$

- $V(S_j, \beta)$ is defined as the part of the observable utility that contains a vector S_j of explanatory variables that vary by each alternative j within the set Z (Cameron & Trivedi, 2005). As an example, one could think of travel-costs when choosing between different transportation methods, like taking the train, the car or the bike (Train, 2003). Such variables are called *alternative-specific variables*. Importantly, note that the coefficients β do (usually) not vary over alternatives (StataCorp, 2019).
- $V(X_i, \beta_j)$ represents the part of the observable utility that contains a vector X_i of covariates that vary over the individual i, but do *not* vary over each alternative j within the set Z (Cameron & Trivedi, 2005). Therefore, I will refer to these variables as *individual-specific variables*. Examples for variables that would be included in the vector X_i would be the age of an individual, the gender or even the health-status. In contrast to alternative-specific variables, the coefficients β_j for each individual characteristic within X_i will vary over the alternatives (Train, 2003).
- Ultimately, the term $V(S_j, X_i, \alpha)$ is the part of the observable utility, which contains the *interaction* between some alternative- and individual-specific variable. To be consistent with the given example from $V(S_j, \beta)$ and $V(X_i, \beta_j)$, we could think of an interaction between the alternative-specific variable travel-cost and the individual-specific variable gender. These interactions also vary across alternatives and go under the name of *sociodemographics*. Like it was the case with alternative-specific variables, the coefficients α do not vary over the alternatives (Train, 2003).

As can be seen from the variable-examples given for $V(S_j, \beta)$ and $V(X_i, \beta_j)$, a researcher is thus able to personalize the behavior of individuals and adapt it to the available data. A further advantage of the RUM-approach is that the importance of each element $\begin{pmatrix} x_{j,1} \\ x_{j,2} \\ \vdots \end{pmatrix}$ in the vector X_i or $\begin{pmatrix} s_{1,i} \\ s_{2,i} \\ \vdots \end{pmatrix}$ in the vector S_j within an individual's decision-making process can be quantified by analyzing the marginal effects and sign of the beta coefficients. At this point, it is to note that RUMs are not estimated by ordinary least-squared (OLS), but rather by *maximum-likelihood estimation* (McFadden, 1974). Therefore, the coefficients must first be transformed, before being able to interpret the magnitude correctly (Stock & Watson, 2015). Formally, the estimated sample regression function (SRF) that tries to approximate the PRF, can be written as:

$$U_{ij} = \hat{V}_{ij} + \hat{\varepsilon}_{ij} \quad (9)$$

- Where $\hat{V}_{ij}$ is the observable part of the utility by the researcher that can be estimated by using a set of individual- and alternative-specific variables as explanatory variables (Cameron & Trivedi, 2005; StataCorp, 2019).
- $\hat{\varepsilon}_{ij}$ is the residual, which is the estimation of the error-term that we get (Train, 2003).

Lastly, by including sociodemographics, a researcher will be able to account for *systematic taste variation*, which will capture the heterogeneity in preferences across individuals by constructing an interaction term between an individual- and an alternative-specific-variable (Train, 2003).

Now that we know how to handle utilities, the remaining questions are, how this approach turns out to be probabilistic and how it is connected to utilities? One key aspect of the RUMs they are built around the fact that a researcher cannot observe every aspect of an individual's choice, be it because not all relevant characteristics of an alternative and / or the individual are observable or because of the problem that the researcher is most likely unable to detect the specific context why each individual may choose an option *p* over an alternative *q*. This deficit in information is the reason why a *probabilistic approach* is justified (Manski, 1977). A researcher can never know an individual's choice with 100 percent certainty. Therefore, RUMs assign a probability to every possible alternative of an individual's choice, denoted as *choice-probabilities*. According to McFadden (1974) this gives us:

$$\overbrace{P_{ij}}^{choice\ probability} = Pr(U_{ij} > U_{iq}, \forall\, j \neq q) \quad (10)$$

Where I assume that individual *i* will choose alternative *j* instead of all other options *q*, since it brings him higher utility U_{ij} relative to any U_{iq}. Plugging in equation (6) the expression (10) results in:

$$P_{ij} = Pr(V_{ij} + \varepsilon_{ij} > V_{iq} + \varepsilon_{iq}, \forall\, j \neq q) \quad (11)$$

$$\Leftrightarrow P_{ij} = Pr(\varepsilon_{iq} - \varepsilon_{ij} < V_{ij} - V_{iq}, \forall\, j \neq q) \quad (12)$$

According to Train (2003), P_{ij} is defined as a cumulative distribution. By using the density function of the error-term f(ε_i) - where $\varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iZ} \end{pmatrix}$ is a vector that collects all the error-terms from each alternative within choice set *Z* for a particular individual *i* - we can rewrite the choice probability in equation (12) into an integral:

$$P_{ij} = \int \underbrace{I(\varepsilon_{iq} - \varepsilon_{ij} < V_{ij} - V_{iq}, \forall j \neq q)}_{indicator-function} f(\varepsilon_i)\, d\varepsilon_i \qquad (13)$$

Where *I(·)* is a so-called indicator-function which takes the value of 1 if the expression within the parentheses is true and 0, if the input is false (Train, 2003). This equation will be the starting point in the later subsection 3.2.2.

Now that the *generic* empirical guidelines for my further analysis have been laid out and the relationship between choices, utility and probabilities clarified, the last important question that remains is, if we can justify to set preferences and choices as equal to one another, since - in reality - only choices are observable while tastes stay latent and only on a theoretical level? To answer this question, one needs to take the theoretical framework that form the foundations of random utility models - known as *random utility theory* (RUT) - into account. The main goal of RUT is to uncover the preferences that drive a person's choice (Kjær, 2005). The advantage of RUT is that it was proven to be consistent with the economic concept of utility maximization (Lancaster, 1966; Manski, 1977). As Mas-Colell et al. (1995) point out, a choice correspondence *C(·)* that is specified by utility maximization has a utility representation, which in turn implies that we can infer latent preferences from the choices individuals make (Sen, 1973). To get a more formal approach on why this holds, I refer to Appendix D to get a deeper insight on this.

2.2 Construction of a Theoretical Model: Why do Individuals do Sport?

After assessing the suitability of commonly used models when dealing with preferences and choices for my later empirical analysis, this next section will revolve around *finding determinants* that explain, why people have a preference for sport or not. By exploring the *sport literature*, but also by using my own intuition, the goal will be to construct a theoretical model on my own, which will serve as a base for the formulation of my hypotheses in the next subsection, but also for the empirical modelling in Chapter 3.

2.2.1 Motivation for being Sportively Active: Neoclassical VS. Behavioral Literature

In sports economics, there are two opposing theoretical frameworks that try to determine, why people tend to do more or less sport in their lives. On the one hand, we have the *neoclassical theories* with the Becker model (1965) being of main interest, when it comes to explain how individuals allocate their time spent on leisure and work. The choice made by an individual within this model is explained to result from a trade-off between the number of working-hours and leisure-hours, in order to determine the income that maximizes an

individual's utility. Thus, the two main drivers of an individual's decision are derived from *financial incentives*, according to neoclassical theorists. Within these general conditions, there are - accordingly - 2 main reasons why individuals engage in sports activities: either having an *(exogeneous) preference for sport*, or to *remain healthy*, in order to be able to be more productive and - as a consequence - to acquire more income. It should be noted that - depending on which of the two neoclassical motives we focus on - the link between utility and sport changes. While individuals having an exogeneous preference for sport gain utility *directly* by consuming / doing the sport they love, individuals practicing a physical activity because they want to stay healthier, gain utility *indirectly* - via the health-channel - from doing sport (Cabane & Lechner, 2015).

On the other hand, there is the *faction of behavioral theorists* which explore alternative reasons why individuals choose to do sport. The key difference between behavioral and neoclassical theorists is how they deal with preferences. For neoclassical theorists, the exact reasons why a person might have a preference is secondary, or even exogeneous (Sen, 1973), while their behavioral counterparts endogenize them. When it comes to endogenizing preferences in the context of sport, behavioral scientists mostly take inspiration from other scientific fields, especially from psychology and sociology. In contrast to the neoclassical theory, these "heterodox" theories put the *main emphasis on various social relations* rather than financial aspects. It should be mentioned here that behavioral scientists do not simply disregard financial incentives, yet they lose weight and become secondary, compared to social incentives (Cabane & Lechner, 2015). For instance, the intergenerational literature - one branch of the vast behavioral literature - hypothesizes that there are some individual traits that can be "inherited" not solely by genetics, but also by interacting socially with others, typically from one generation to the next, or with close friends and family-members within the entourage (Angrist, 1990; Campante & Yanazingawa-Drott, 2015; Currie & Moretti, 2003; Fernández et al., 2004; Goodman et al., 2019; Johnston et al., 2013; Kremer & Levy, 2008). Consequently, individuals experiencing social interactions with more physically active peers that - initially - may not have had a preference for sport, could - over time - develop a liking towards sport later in their lives (BAG, 2018; Downward & Riordan, 2007). By endogenizing the preferences of individuals, explaining so called preference reversals becomes a possibility.

2.2.2 Determinants of Sports over an Individual's whole Lifecycle

Now that the main differences in ideologies between neoclassical and behavioral theorists have been made clear and on which factors they put their focus, the question remains: who is in the right? By taking a first peak at my data, the answer would be: probably both.

Table 1: Sample Frequencies and Proportions of Individuals' Lifecycle Choices for doing Weekly Sport

Lifecycle Sport-Choices	1) Always Sport	2) Sport only as an Adult	3) Sport as a Child	4) Never Sport	Total
Frequency	1001	120	330	128	1579
Proportion	63.39%	7.60%	20.90%	8.11%	100.00%

As we can see from Table 1 above, sorting by every individuals' sport choices throughout their lifecycle results in 4 categories: there are some agents that decide to always do sport, some that only start to do sport later in their life, some that stop when they become older and others that never even considered to do sport. Thus, there are individuals that - consistent with the DUM of the neoclassical economists - always stick to their choices (categories 1 and 4), while others experience a preference reversal and switch choices (categories 2 and 3), which is in line with the endogenized preferences originated from the behavioral scientists. As a consequence, modelling the entire lifecycle choices of an individual within the RUM-framework will require me to think about the financial aspects from neoclassical theories, as well as why individual suddenly change their minds and begin / end to do sport, that is from a more behavioral perspective where social interactions play a key role by shaping one's preferences according to the sport literature. Since - to my knowledge - there were no studies available that analyzed the sport decision-process over a lifecycle, the following theoretical framework and flowchart which can be found in Appendix A (see Figure 5) were mostly built with the help of the health-, intergenerational- and other branches of the behavioral-literature, but also extended with my own intuitions if needed. The goal was to build and visualize a theoretical framework, that would summarize all the important determinants within the decision of doing sport or not over the whole lifecycle, which will later serve as a base for my empirical work.

2.3 Hypotheses

As we have seen, individuals may have varying tastes when we look at sport preferences over a lifetime. Now that we have solid *theoretical* grounds, let's hypothesize and - later - test *empirically* with the RUM-framework, why individuals may stick to their choices to always / never do sport and why others may decide to start / stop doing sport as an adult. Note that for each of the 4 possible lifecycle sequences, I will formulate one hypothesis:

1. Doing sport in childhood and adulthood: When looking at motives that will influence people to stick with their choice to *always do sport*, models which consider social interaction across generations seem to be the way to go. More specifically, the influence of the parents will be the focus of this first hypothesis, since the intergenerational literature finds that attitude transmission across generations have strong effects and implications on the subsequent generation (Campante & Yanazingawa-Drott, 2015; Dohmen et al., 2012; Fernández et al., 2004). The main intuition would be that - when individuals are still in their childhood phase - parents play a big role in shaping the tastes of their offsprings. Similarly to the theoretical model of Bisin and Verdier (2000), parents that have a preference for doing sports themselves, will desire to pass down their passion to their children. This can be thought of a sort of bequest that is passed down from one generation to the next, similarly to one of the motives mentioned by Rae (1905), or - more recently - formalized by Currie (2009) in the context of a child's health production. Consequently, the initiation of doing sport as a child because of the parents will then lead to the development of a habitual behavior of doing sport, which should increase the probability of the individual to continue doing sport, also as an adult (Heckman, 1981; Train, 2003). This hypothesis why people might choose to always do sport, will be referred to as *HFH*.

2. Starting sport as an adult: When looking at the motives, why individuals might *start to do sport only later in their lives*, I took inspiration from the neoclassical theories, especially from the theoretical model of Grossman (1972), which represents a dynamic optimization problem of an individual with regard to his so called "health capital". To get a better grasp of what the abstract concept of a "health capital" is, one can think of it as an endowment, that each individual receives from God / Nature at the beginning of their lives and which starts to depreciate, until it expires and the individual dies. Another aspect of this health capital is that - the older the individual gets - the stronger the depreciation of an individual's health (Grossman, 2000). By deciding to do sport, this rate of depreciation can be reduced, so that less "health-capital" is depreciated in the next period / future relative to an individual that decides not to do sport. Thus, an individual would experience a gain in utility if he / she is able to maintain his / her health capital at a higher level with sport (because the live-span would be extended) relative to the counterfactual decision not to do sport. Furthermore - and taking inspiration from Grossman (2000) - I suspect that the

depreciation-rate of "health capital" tends to be lower at the beginning of a child's life: as a consequence, the utility-gain for a child (if he or she does sport) will be relatively lower, compared to adults who will probably have more incentives to reduce their health capital depreciation-rate and prolong life through sport. Thus, an adult - relative to a child - would receive much more benefit from sport than a child, because these individuals also have a different reference-point in life. Hence, children should be characterized by possessing a high "health capital" with low depreciation, while adults would demonstrate a lower health capital with high depreciation (the marginal rate of substitution between "sport" and "no sport" changes with age) and therefore be more motivated to start doing sport later in life, since the indirect utility gains due to sport for adult are high. This idea of maintaining its *future health* due to *today's* choice of doing sport at a later age - similarly to the motive of having a preference for improvement discussed by Chapman (2000) - is my second hypothesis and will be referred to as the *HIH*.

3. Stop doing sport in adulthood: When looking at motives for why an individual might *cease to do sport in adulthood*, the financial incentives gaining more importance in a later phase of an individual's life - as described by the neoclassical theorists (Becker, 1965) - may be one of the main reasons. The number of hours worked - in contrast to sport, which is an activity that is executed during one's free-time - will be the hypothesis that explains, why people stop doing sport when they are older.

4. No sport as a child, neither in adulthood: Last but not least, the individuals which *never do sport* may have never considered the choice of doing sport to be available, possibly because they prefer *other leisure-activities* over the sport-activity (Cawley, 2004).

As we can see with the formulation of those hypotheses, *inter- as well as intragenerational* factors may have an influence on the underlying lifecycle preferences of individuals in the context of sport. Ultimately, the formulation of these hypotheses serves the purpose of differentiating between different types of individuals: the ones motivated intrinsically - possibly forming habitual behavior because of their parents (HFH) - and others mainly motivated extrinsically and gaining a preference in sport only later in their life, possibly because of health (HIH). This will be helpful when specifying the exact model later. Hence, let us proceed to the empirical analysis to determine the importance of each of those determinants.

3. Empirical Analysis

This new chapter marks the beginning of the quantitative analysis. The previously presented qualitative work can be seen as the necessary preliminary knowledge that will be used as a reference and basis for the subsequent chapters. The main objective in the remaining part of this thesis will be to test the formulated hypotheses and - ultimately - answer my research question stated during the introduction.

In this chapter, I will first outline the general constraints imposed by the dataset I selected, by focusing mainly on the aspects of variable-availability and the dataset's general structure, since those limitations will impact the methods that will be implemented to test my hypotheses. After these preparations, the actual model will be estimated using the identified methods (Chapter 3.2.1 and 3.2.2).

3.1 Data

The data-source used for this analysis is based on the questionnaire of the Swiss Household Panel (SHP). Every year - starting in 1999 until 2018 - each individual older than 13 within a household that participated in the survey was interviewed - mainly by phone - to answer a vast amount of questions, covering most aspects of an individual's personal life: work, income, education, personal health, leisure activities, as well as some information about the household. The data was collected *randomly* across the 7 main regions in Switzerland, thus guaranteeing that the sample will be representative of the Swiss population (Voorpostel et al., 2018).

Another relevant piece of information concerning the *generic* structure of the dataset is how each individual within a household, or family, is answering to the SHP. Depending on the person's age or availability, the subjects of the survey are left to choose between three possible types of questionnaires[4] (Voorpostel et al., 2018):

1. Fill out an *individual-questionnaire*, that contains all the personal information about a person mentioned in the first paragraph of this subsection. Ideally, this is the type of questionnaire which should be completed by an individual (Voorpostel et al., 2018).
2. Fill out a *proxy-questionnaire*, which mostly contains some information about work and health-status, but - relative to an individual questionnaire - offers only very little information on the individual-level (Voorpostel et al., 2018).

[4] Assuming that they are willing to participate in the SHP.

3. Fill out a *household-questionnaire*, which gives information on the composition of an individual's household, but does *not* include any personal-level data that can be used for the analysis (Voorpostel et al., 2018).

Besides the wide range of variables, what makes this dataset especially suitable for my analysis is that - within the year 2013, or *wave 15*, to speak in terms of the SHP-terminology - a set of detailed questions about an individual's sport-activity were asked. Most importantly - within this set of sport-questions - one is of particular interest, because it questions the individuals about their sport participation when they were *"young"*, where young is defined as being 12 years old (Voorpostel et al., 2018). Since the main unit of observation and dependent variable in my analysis is the sport-activity within a lifecycle of a person, having individuals answering a question whether they participated in sport at a very early stage of their lives is crucial, otherwise the dynamics of sport preferences could only be achieved over a much shorter time-span.

At this point, it is to note that, while most of the questions stay the same during each wave of the questionnaire, there are some *variables that were only asked within one particular wave*, or - sometimes - in very irregular survey-intervals (Voorpostel et al., 2018). The sport-variables are an example of such non-recurring questions. Even though the set of questions about the current sport-activities reappears in 2015 (wave 17), *the question whether an individual did sport in his youth is only asked in the year 2013*. Hence, the analysis will focus on those individuals, which were interviewed in 2013. Although an analysis using the wave 17 is also possible, I am bound to the individuals that initially answered the questionnaire in 2013. Since there is the possibility that individuals decide to leave the panel after each year that passes, using the sport-data of wave 17 would result in a sample with less observations. Furthermore, because I matched each parent-pair to the children in order to account for social-interaction within the family, the risks of attrition and missing data become even greater. Hence, I would have had to deal with fewer observations and, therefore, also less statistical power if the data in 2015, rather than in 2013, was used. Therefore, I base my analysis on the year of 2013.

3.1.1 Cleaning

After this crucial initial decision, I start by *cleaning* the dataset and creating the variables needed based on the theoretical model I created in section 2.2.1 that resulted in Figure 5 (see Appendix A). Initially, wave 15 contained 20'451 observations. However - as we saw within subsection 2.2 - social interactions play a key role when accounting for endogenous preferences in theory, especially the parental variables, which will be needed to test my HFH. Therefore,

only those individuals were taken into the final sample, which had *at least one parent* responding to the sport questions. From a technical point of view, this requires matching the parent and children over their respective IDs and move back and forth between short- and wide-formatted datasets. The code for this can be found in the file '*ddset-construction.R*'. With those restrictions, the ultimate dataset focuses on 1579 individuals in total. Overall, I use 37 variables which are able to cover most of the factors identified in the literature. They are specified in Table 7 and 8 which can be found in Appendix B.

Before continuing forward with the methodologies chosen, two important points need to be addressed:

- From a technical point of view, the dataset used is a *cross-section*, since only one wave can be considered, due to data-limitations, within this empirical analysis. As we could see with sport variables, some questions in the SHP are framed in a way that they ask an individual about their past, especially events that occurred a year before. For instance, one health-related question asks about the number of doctor consultations during the last 12 months (Voorpostel et al., 2018). This allows for the inclusion of so called *lagged-variables* in my analysis, a type that will become relevant later in the methodological-part.

- Concerning the sport-variables of the parents that should proxy the intergenerational transmission process, empirical papers within the behavioral literature point out that the effects of the parents on the children's preferences are *distinct* from one another (Dohmen et al., 2012). Therefore, using each parent separately in my empirical analysis would have been ideal. However - in many cases - I only have the information on one of the two parents[5]. At this point, it is to note that the way in which the dataset is constructed *assures to always have the information for at least one of the child's parents*. To counteract this issue, I constructed a new categorical variable named *Parents doing Sport*, as well as *Parents Sport when Young*, which measures whether *at least one of the two parents* practices sports or not. This will allow me to measure the effect of the parents on the child, by containing the information of the father *and* mother into a single variable. The advantage of this is that it will allow me to keep as many observations as possible during the model estimation, a - oftentimes - required necessity when using certain choice models as an analysis-tool (Petrucci, 2009). The reasons behind why one of the parents is frequently missing can have two different origins:

[5] To be precise, I only have the complete information on *both* parents in 957 out of 1579 times.

either the mothers / fathers filled out the proxy or grid questionnaire, *or* because the parents separated in an earlier wave and thus one of the two parents is not able to participate in wave 15, since the family does not live within the same household anymore (Voorpostel et al., 2018).

3.2 Methodology

In this section, I will use the concept of RUMs laid out in subsection 2.1.2 for my empirical analysis, which can be considered today's state-of-the-art method to uncover the sport-preferences of individuals, sometimes also called *discrete choice modelling* (Kjær, 2005; Saberi, 2017). Before talking about pinning down the final model, one major issue regarding the *relationship between health and sport* needs to be addressed first.

There is plenty of evidence which hints at a *mutual dependence* between being physically active and personal health-status. More specifically, many studies support the view that doing sport will result in being healthier (Humphreys et al., 2014; Reiner et al., 2013; Sari, 2014). However, the opposite is also true, with a higher (perceived) health-status being associated with more sport (Bauman et al., 2002; Cawley, 2004; García et al., 2011). Hence, there would be a *reverse causality problem* if I wanted to include a health-variable as a regressor in my estimation in order to test my HIH. Therefore, a methodology, or identification strategy, trying to remove the endogeneity from the health-variable *before* including it into the final discrete choice model must be found first.

3.2.1 Estimation of the Health Variable

My final discrete choice model should - ideally - use a health-variable that only reflects the *health gain / loss due to sport*. In this *desirable* scenario, there also would be no endogeneity problem between sport and health. However, there is no such variable in my data set. Therefore, I need to *construct* a health-variable myself to estimate only the effect of sport on health. To build this variable, I therefore searched within the empirical sports-economics literature to implement a suitable identification strategy that fitted my needs. Ultimately, I use a *propensity score matching method*, similarly to Schüttoff et al. (2018), since the set of variables they used in their empirical analysis is extremely similar to mine[6] and matching allows us to estimate the change in health *due to sport*.

[6] This is not a coincidence, since the dataset used by the researchers is the German Socio Economic Panel (GSOEP), which is explicitly designed to - oftentimes - ask the same questions as the SHP for the construction of the same variables, enabling the development of cross-country studies (Voorpostel et al., 2018).

At this point, it is to note that matching will not necessarily solve the problem of reverse causality mentioned before. However, the SHP contains a health-variable defined as *health-improvement* which asks respondent whether - relative to the previous year - their health-status increased or decreased[7] (Voorpostel et al., 2018). I argue that - if I take the *health improvement* variable from the *subsequent* wave - that is one year later (*t+1*, where *t* is my reference year 2013 / wave 15) - the mutual dependence should vanish. While *today's* choice of sport is then able to influence next year's health-improvement, the *reverse* should not be true.

The basic idea behind matching is to compare individuals who do not differ in any observable characteristics, *except* the treatment (= doing sports). The technique by which this is accomplished is by matching individuals displaying similar *propensity scores*, which is a sophisticated term that defines the conditional probability of doing sports (Rosenbaum & Rubin, 1983). Following Dehejia and Wahba (1999), this is done in two steps:

1. In a first stage, the *main goal* is to estimate the propensity scores by using the *treatment-variable* of currently doing sport as the *dependent variable* which will be regressed - in a logistic regression - on all observables (= x-variables) that potentially influence the probability to do sports. Since the y-variable is binary (doing sport or not) in this case, it implies that the estimation-results will reflect the individual's *predicted current (conditional) probability of doing sport*, which represent the propensity scores that we need. Expressed formally, this means:

$$\widehat{P(Sport_i \mid x_1}, x_2 \dots x_J) \; = \; \beta_0 \; + \; \sum_{j=1}^{J} \beta_j * x_j \qquad (14)$$

   Where $Sport_i$ is the binary dependent variable whether individual *i* is currently practicing sport on a weekly basis and β_j is the *j-th* coefficient on the *j-th* explanatory variable x_j in the estimated logistic regression in (14).
2. Then - in a second stage and similarly to Schüttoff et al. (2018) - the effect of the treatment-variable $Sport_i$ on the health variable $Health_i$ is estimated:

$$\widehat{Health}_i \; = \; \beta_0 \; + \; \beta_1 * \; Sport_i \qquad (15)$$

At this point, it is to note that - *before* the estimation of (15) takes place - an algorithm will match individuals from the treatment- and control-group (= those that do no sport) over similar propensity scores that we previously estimated in (14) (Sekhon, 2011; Stuart, 2010). The reason

[7] Originally, the scale of this variable goes from 0 to 10, where 5 denotes an unchanged health-status relative to the previous year (Voorpostel et al., 2018). If individuals give a value below 5, it means that their health-status decreased (and vice versa).

why this is done, is because - when using an *observational dataset* like it is the case with the SHP - individuals doing sport may be very different from those from those who choose to do no sport (Rubin, 1997). Thus, by running a regression *without* the matching-methodology - that is, with the *initial* treatment- and control-group - we would get a biased *average treatment effect* (ATE) on the coefficient β_1 in (15) and would not be able to determine, whether the effect of sport on health is really due to the fact that individuals decided to participate in sport. This problem is called *selection-bias* and happens - in contrast to experiments - because assignment to treatment was *not* random (Rubin, 1974).

By using a matching-algorithm, we are - literally - able to "toss away" all the individuals from the sample, which are too different from - typically - the treatment-group. More precisely, matching will create a *new control-group*, based on the propensity scores (Rubin, 1977, 2007). Note that - when creating this new control-group - the algorithm is usually allowed to match a particular individual from the control-group to *more than one* individual from the treatment-group and / or even match one individual from the control group multiple times ("with replacement"), which can help reduce bias. From a more technical point of view, matching *with replacement* will require a weighting scheme, since the observations within the control group will typically not be independent from one another anymore (Sekhon, 2011; Stuart, 2010). When creating this new control-group, a researcher is free to choose between many different matching-algorithms, such as nearest neighbor, kernel- or caliper-matching, to name a few (Heckman et al., 1997).

After having constructed this new control-group - which should, in a best-case scenario, only differ in their treatment-status - the regression in (15) will be performed. Ultimately and ideally - if all the assumptions that the matching-methodology requires were fulfilled - the estimate of the treatment-variable β_1 will be a *causal estimate* and interpreted as the *average health gain / loss due to sport* (Rubin, 1974). Note that, since we *adapt the control-group to the treatment group*, what we end up with is - usually - an *average treatment effect on the treated* (ATT). However, the matching-method with its flexible weighting-scheme is also able to find the ATE. Hence - depending on whether we weight the estimate β_1 in equation (15) by the number of treatment- or control-observation - it is possible to get estimates for different types of treatment effects (ATT or ATE) (Stuart, 2010).

Now that we know how matching works, the last question would be, how this relates to my work. The final objective of this thesis will be to answer the formulated hypotheses in order to answer my research question. As we have seen, I hypothesized that the reason why individuals

may *choose* to start with sport in their adulthood, may be because they have a preference for improving their future health (Chapman, 2000). The problem is that there exists *no* variable in my dataset, that *only* reflects a person's *expected change in health because of sports*. As we established earlier in this chapter, there is a variable in the SHP denoted as *health-improvement* which seems to be valid candidate for my HIH, yet the change in health may *not necessarily* be due to sport. However, this is exactly where matching comes into play since it allows me to estimate the specific effect of sport on health-improvement (by matching propensity-scores). Thus, the generic regression in (15) would become:

$$Health - \widehat{Improvement\ in\ 2014}_i = \beta_0 + \beta_1 * Sport\ in\ 2013_i \qquad (16)$$

Finally, my ultimate *idea* is to use the estimate β_1 from the matching-regression (16) in order to build a (short-term) *Health-Expectations due to Sport*-variable. The key point that a reader needs to understand here, is that β_1 can be interpreted as "(additional) future-health *due to sport*"[8]. Thus, by taking the *magnitude* of the coefficient β_1, I should be able to proxy a person's anticipated (short-term) gain in health, if he chooses to do sport *today*, hence reflecting his / her preferences for improving their future health in the final discrete choice model. To summarize, β_1 in (15) reflects a person's health-expectation due to sport.

Of course, estimating regression (16) for each individual separately would be *ideal*. However, this would require a *panel dataset*, which is not possible with the current SHP-structure since the questions relating to sport must have been asked in each year. Nevertheless, estimating a *single* matching regression to construct this *health-expectation*-variable would be insufficient, since it is - firstly - unrealistic to expect that *all* individuals have the *same* β_1 health-expectations due to sport and - secondly - might lead to a *perfect multicollinearity problem* if I were to construct the same health expectations for all individuals in my discrete choice model, because the *Health-Expectations due to Sport*-variable would basically be a constant. Nevertheless, since matching allows for the estimation of different types of treatment effects - the ATT & ATE in particular - it is possible to add variation into the constructed *Health-Expectations due to Sport*-variable by allocating the ATT to individuals that are *currently* doing sport, whereas the ATE is assigned to those who currently do no sport. Additionally, to reduce the perfect multicollinearity problem even further and bring more variation into the *Health-Expectations due to Sport*-variable, I will build four *sub-samples* to run the matching-regression (16) on:

[8] Note that in this statement I implicitly assume that the coefficient β_1 in (15) will be positive, since - intuitively - it seems reasonable to assume that more sport will lead to better health.

Table 2: Sample-Selection for Matching

1. Children and adolescents under 21 years of age who are classified as normal weight by the *world health organization* (WHO, 2020a).
2. Adults over 20 years old who are classified as normal weight by the WHO (2020a).
3. Individuals categorized as underweight by the WHO (2020a).
4. Individuals classified by the WHO (2020a) as overweight, obese or severely obese[9].

The above subsample-division in Table 2 is the most reasonable approach I could achieve with the given data. Intuitively it seems to make sense that - for example - individuals with overweight have different health-expectations relative to normally weighted children when they do sports, similarly to my argumentation in subsection 2.3 concerning the HIH, where I justified why children and adults may have different health-expectations. This is in line with empirical findings: Deforche et al. (2006) found that adolescents categorized as overweight or obese displayed a more negative attitude to engage in physical activity relative to normally weighted adolescents. Therefore, building various sub-samples in order to estimate different β_1 in (16) seems to be meaningful. In total, there will be 8 possible β_1 matching-estimators if we run regression (16) with the 4 specification of Table 2 and depending on whether the individual currently does sport (then we allocate the *ATT* from equation (16)) or not (then we allocate the *ATE*).

Now that every individual has their "expected improvement in health due to sport" being allocated correctly, the next step is to incorporate their reference point into this expectation variable. Taking inspiration from the behavioral literature, the main idea here is that all individuals have a different "anchor" for their health, which is given by their current, subjective health-satisfaction-status (Tversky & Kahneman, 1974). For instance, it makes sense to think of individuals currently being on a lower level of health-satisfaction, probably expecting a higher gain in health in the next year, if they choose to do sport today. Thus, I thought of implementing a ratio into my regression, instead of the plain coefficients β_1 of the matching regression in (16) by using the *health-satisfaction*-variable in the *current* year from the SHP-questionnaire as a reference point to build the expected (short-term) gain in health-satisfaction, which is the final-form of my *Health-Expectations due to Sport*-variable :

[9] *Normal-weight* is defined as having a body mass index (BMI) between 18.50 and 24.99. In contrast, *underweight* individuals exhibit a BMI of below 18.5, whereas *overweight* individuals have a BMI between 25.00 and 29.99, while obesity begins at a BMI of 30.00 (WHO, 2020a). According to the WHO (2020b), the BMI is calculated as: $BMI = \frac{weight\ (in\ kg)}{height^2\ (in\ m)}$.

$$Health - Expectations\ due\ to\ Sport_{\ i} = \{\frac{(hsat + \beta_{1i})}{hsat} - 1\} * 100 \quad (17)$$

Where *hsat* is the health-satisfaction variable in the current year of 2013 and β_{1i} one of the 8 possible matching-estimators that will be estimated with matching-regression (16). This variable will have the advantage of giving those on the lower scale of *hsat* to weight the gains stronger, than someone who has already a high scale of *hsat*. For example, if a person with current *hsat* of 7 (from a scale from 0 to 10) expects a gain due to sport of 0.3 - according to equation (17) - this results in (Voorpostel et al., 2018):

$$Health - Expectations\ due\ to\ Sport_{\ i} = \{\frac{(7 + 0.3)}{7} - 1\} * 100 = 4.29\% \quad (18)$$

In contrast, a person that expect the *same* gain in health due to sport of 0.3 but has a *higher* initial health-satisfaction of - for example - 8, will have a lower expected gain in health-satisfaction:

$$Health - Expectations\ due\ to\ Sport_{\ i} = \{\frac{(8 + 0.3)}{8} - 1\} * 100 = 3.75\% \quad (19)$$

This is a 14.4% lower expected gain in health, relative to someone lower on the health-satisfaction scale. Thus, building this ratio as a variable will allow me to implement the idea that a person on the lower scale of the health-satisfaction will have a different valuation of gaining 0.3 points in health relative to a person that is already at a higher level of current health-satisfaction, which enables me to incorporate a reference-point for health (Grossman, 2000). Note that I also allow individuals already having the maximum health-satisfaction value of 10 to have an expected gain due to sport, since - in the Grossman (1972) model - it is assumed that someone's "health capital" is steadily decreasing over time. Thus, even individuals on the highest possible scale would still expect some health-improvements.

3.2.2 Discrete Choice Modelling

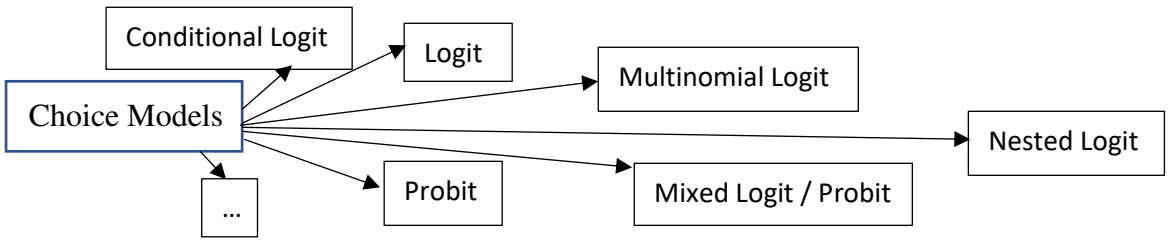
As we have seen in the subsection 2.1.2 that introduced RUMs in a general way, the ultimate goal is to estimate the PRF in equation (6). This is achieved by using SRF-equation (9) as an approximation of the true data generating process in (6), where the researcher can use individual-, as well as alternative-specific characteristics to achieve this. Once the utility $\hat{V}_{ij}$ is estimated, the last step in RUMs is to compute the choice probability P_{ij} that a particular alternative *j* will be chosen by an individual. This is done by solving the integral in equation (13), namely $\int I(\varepsilon_{iq} - \varepsilon_{ij} < V_{ij} - V_{iq}, \forall\, j \neq q)\, f(\varepsilon_i)\, d\varepsilon_i$. At this point, it is to note that - depending on how we specify the random distribution $f(\varepsilon_i)$ - the integral in expression (13)

can be evaluated differently (Train, 2003). As seen in (6), the random error can be calculated by subtracting the observable part of the utility V_i from the true utility U_i:

$$\varepsilon_i = U_i - V_i \tag{20}$$

Starting from definition (20), we see that the characteristics of ε_i, more specifically its distribution *f(ε_i)*, will heavily depend on the observable part of utility V_i which is just a function of characteristics and its parameters $V(X_{ij}, \beta)$, as described in (6) from subsection 2.1.2. Hence, the question that needs to be answered in this chapter is which exact model within the RUM-framework is best suited to estimate the observable part of the utility $\hat{V}_{ij}$ in the SRF-equation (9), since - ultimately - this will allow to make assumptions about *f(ε_i)* and to solve the integral that allows to figure out the choice probabilities of each alternative (Train, 2003).

Figure 1: Modelling Possibilities within the RUM-Framework



As we can see with Figure 1, there are various types of possible models that can be estimated within the RUM-framework, based on how the researcher specifies *f(ε_i)* (Train, 2003). However, some of the above models are just extensions of other models. The main idea here is to give readers a spectrum of possibilities. Ultimately, RUMs can be differentiated depending on whether the researcher uses individual- or alternative-specific characteristics (Cameron & Trivedi, 2005). Therefore, the two possible primary model-classes are:

- Either using a *conditional logit model* (CLM), where *only alternative-specific characteristics* X_j are used and - by definition - the function $V(X_{ij}, \beta) = V(X_j, \beta) = X_j\beta$ (Cameron & Trivedi, 2005).

- Or using a *MNL*, where *only a set of K individual-specific variables* are used and - by definition - the function $V(X_{ij}, \beta) = V(X_i, \beta) = X_i\beta_j$. Importantly, note that - in an MNL - one of the choices *j* within the set of alternatives *Z* will be *normalized to zero*, that is let alternative j = 1 be the *reference-category*, then the vector of coefficients in the MNL is $\beta = \begin{pmatrix} \beta_2 \\ \vdots \\ \beta_Z \end{pmatrix}$, where $\beta_1 = \mathbf{0}_{K\,x\,1}$ (Cameron & Trivedi, 2005; Long & Freese, 2001).

By looking at all the variables I use with the help of Table 7 and 8 in Appendix B, we see that - without exception - my regressors are all *individual-specific variables*. Therefore, the *MNL will be used for the estimation*. Now that I pinned down the model I will estimate, the next question will be, how exactly I will specify the different alternatives within the MNL, which reflect the *dependent variables* in this model. As the research-question already hints at and the way my hypotheses were formulated, the observable utilities will reflect an individual's *sequence of choice*. More specifically, it will reflect the sport-choices over the whole lifecycle of a person. Therefore, the *dependent variable in my MNL* will be specified as follows:

Table 3: Description of Dependent Variables

$$y_{i1} = Doing\ sport\ when\ young\ and\ in\ adulthood = always\ sport$$

$$y_{i2} = Doing\ no\ sport\ when\ young, but\ doing\ sport\ as\ an\ adult = sport\ as\ adult$$

$$y_{i3} = Doing\ sport\ when\ young\ but\ not\ in\ adulthood = sport\ as\ a\ child$$

$$y_{i4} = Neither\ doing\ sport\ when\ young, nor\ in\ aduldhood = never\ sport$$

These choice sequences described in Table 3 will reflect an individual's deterministic portion of utility V_i for one of the four particular (lifecycle) outcomes. The estimation of $\hat{V}_{ij}$ will be performed by maximum-likelihood and the set of individual-specific variables in Table 7 and Table 8 will be used. The choice-sequence which will be normalized and serve as the *base category* will be the alternative *always sport*.

Last but not least, it is important to talk about the *coefficient-interpretation*. In contrast to the CLM, the coefficients β_j of each individual-specific variable within an MNL will *vary* with each alternative. For instance, let us take the 4 lifecycle choices as I defined them in Table 3. As we have seen, one of the alternatives must be normalized in an MNL, in this case let it be *always sport*. As a consequence, we will have only 3 y-variables in the MNL, while the missing one is the reference-category. The normalization *also* has an implication on how the coefficients will be interpreted. If, for example, the factor "health" is inserted as an x-variable in the MNL, then - because the coefficients vary for each alternative - the estimation will result in 3 different health-coefficients for each of the 3 defined outcomes. These would then - depending on the *sign* of the coefficient - be interpreted as a *relative gain (positive coefficient-sign) / loss (negative coefficient-sign) of utility*[10] *for an additional unit of "health", relative to the selected reference category* (Crowson, 2020; Kwak & Clayton-Matthews, 2002). As we have seen with

[10] To be precise, it is the change in the log odds, not utility (Crowson, 2020). However, because equation (9) in subsection 2.1.2 refers to it as $\hat{V}_{it}$, I refer and interpret it as the relative change in utility.

equation (10) in subsection 2.1.2, gaining relatively more utility for a particular option within the RUM-framework is *equivalent* to saying that an individual will be more *likely* to choose the alternative with the higher utility. Hence, I can also interpret the coefficient as changes in predicted probabilities due to a unit change in the independent variable, which is more intuitive and represents the marginal effect[11] interpretation (Hanmer & Ozan Kalkan, 2013).

Now that the coefficient-interpretation has been established, the remaining question is, how it relates to the test of my hypotheses? As we have established in the theoretical foundations, all individuals experience their own "history" in life, every human has a different status quo and, thus, puts a different weight on different factors: individuals doing sport only as adults may attribute a higher importance on health-related variables, whereas, for individuals with habitual behavior, the influence of the parents is relatively stronger (Cawley, 2004; Glanz et al., 2008). This is where the framework of the MNL comes in as a handy methodological tool, since the coefficient-interpretation allows for a relative comparison of different alternatives / choices by looking at the signs of the coefficients.

Obviously, with my data, is not possible to compare the relative valuation of a regressor for each specific individual. However - by “nesting” individuals according to their types of lifecycle choices - it will be possible to group individuals with a similar history - and thus, similar preference - together, reducing bias *across* the different groups and get a more precise idea, how strong the relative valuation of factors such as health are for a particular individual *type* - for example the type of individual doing sport since childhood - relative to other types (Train, 2003). This methodology will allow me to test *all* my formulated hypotheses. For instance, if individuals have a preference for improving their future health - as I speculated in my HIH - then I should expect the relative valuation of the factor health for the type of individuals who start to do sport later in their life being the highest, relative to the other categories, or - at least - higher than for individuals doing always sport.

To be more precise, here is a concrete example: remembering that I defined the individuals always doing sport as the reference category, then - if the coefficient on health results in having a *positive* sign in my MNL - it would give empirical evidence[12] for documented psychological phenomena related to health (Chapman, 2000; Frederick et al., 2002). The other hypotheses can be checked accordingly to the given example, by looking at the variable of interest as defined in the subsection 2.3.

[11] Note that the marginal effect in an MNL is non-constant (Hanmer & Ozan Kalkan, 2013).
[12] Of course, assuming that the coefficient is statistically significant.

4. Results

This chapter will apply the strategies defined in the methodological part on my data. Before moving on to the actual results, Table 4 summarizes some important variables, such as the sport-activity from both parents and children, but also some health-related variables that play a major role in the matching regression or the subsequent MNL-estimation. For more detailed summary statistics, consult Appendix C Table 10 and 11.

Table 4: Summary Statistics

Sample Summary Statistics	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Age	1,579	22.267	6.651	14	17	26	58
Gender	1,579	0.495	0.500	0	0	1	1
Health-Satisfaction	1,579	8.211	1.516	0	8	9	10
Health-Improvement	1,579	0.362	1.185	-5	0	0	5
BMI	1,543	22.297	3.768	13.590	19.869	23.875	54.785
Monthly Sport	1,579	0.777	0.416	0	1	1	1
Weekly Sport	1,579	0.710	0.454	0	0	1	1
Youth Sport	1,579	0.843	0.364	0	1	1	1
Mother Sport	1,435	0.755	0.430	0.000	1.000	1.000	1.000
Mother Youth Sport	1,435	0.615	0.487	0.000	0.000	1.000	1.000
Father Sport	1,101	0.718	0.450	0.000	0.000	1.000	1.000
Father Youth Sport	1,101	0.760	0.427	0.000	1.000	1.000	1.000
Parent Sport	1,493	0.861	0.346	0.000	1.000	1.000	1.000
Expect. Health due to Sport (%)	1,541	2.799	1.517	0.207	1.799	3.849	17.319

As we can see from Table 4, the average age of the individuals in my sample is rather young. This is - however - not surprising, since I only used observations that have at least one of their parents answering the individual-questionnaire to account for the intergenerational process. Because the surveyed individuals are mainly young adult, it is also rather unsurprising that the individual's health-satisfaction - a subjective health-measurement - is over 8, on a scale that goes from 0 to 10. The more objective health-measure of the BMI confirms this healthiness of an average individual in my sample, since a BMI of 22.3 is considered to be a person of normal-weight (WHO, 2020a). Moving on to the sport-variables, we see that - regardless, which generation we look at - most of the individuals seem to be active in sport. While fathers and the individuals from the subsequent generation experience a drop in the average sport activity when becoming older, we can witness an increase in sport activity for mothers. Another thing to note is the much lower number of observations that I have at my disposal for the parents. I loose around 10% of data for mothers and 30% for the fathers. The reasons for this loss of information are multiple. Sometimes, the parents separated and - as a consequence - only one of the children's parents can be interviewed, while at other times, only one of the parents answered

the individual-questionnaire, where the sport-questions were asked (Voorpostel et al., 2018). As already mentioned in the methodological part, I therefore compromised and constructed the variable *Parent Sport*, that accounts for both parents at the same time, by using the information, whether *at least* one of the two parents practices sport or not (the same strategy is used for the parent's youth-question). Hence, around 95% of the observations can be conserved, which is a drastic improvement. Nevertheless, we need to keep in mind that *other* parental-variable suffer from the same problem and lead to a loss of information when performing the analysis. Finally, the *Expected Health due to Sport* variable is based on the results of the matching-estimates and reflects the final-form of the equation (16) that will be used to test my HIH in the MNL. The main concern expressed in subsection 3.2 was that of multicollinearity: Yet the allocation of the different matching-estimates, as well as the inclusion of the current health-satisfaction as a reference point for each individual seem to have induced enough variation for this variable to be utilized in the final MNL.

Next, let's perform a standard 1 to 1 matching for the construction of *Expected Health due to Sport* variable by using the following model as a baseline to estimate the propensity scores:

$$sport_{iw} = \beta_0 + \beta_1 age_{iw} + \beta_2 m\_age_{iw} + \beta_3 sex_{iw} + \beta_4 educ_{iw} + \beta_5 m\_educ_{iw} + \beta_6 work_{iw} + \beta_7 p\_sep_{iw} + \beta_8 leisure\_index_{iw} + \beta_9 leisure\_screen_{iw} + \beta_{10} doct\_cons_{iw} + \beta_{11} hsat_{iw} + \beta_{12} h\_impr\_lagged_{iw} + \beta_{13} nber\_sib_{iw} + \beta_{14} ysport_{iw} + \beta_{15} p\_sport_{iw} + \beta_{16} sib\_sport_{iw} + \beta_{17} p\_unhealthy_{iw} + \beta_{18} hh\_inc_{iw} + \beta_{19} region_{iw} \quad (21)$$

Where $sport_{iw}$ denotes if individual i with the - in Table 2 - defined weight-categories $w \in \{underweight, normal\ weight\ \&\ young, normal\ weight\ \&\ adult, overweight\ or\ obsese\}$ practices sport. The exact specification of all abbreviated independent variables in (21) can be found in Appendix B. Since there is no general consensus on how matching should be executed (Sekhon, 2011), I thought to use the model of Schüttoff et al. (2018) as a reference for the covariates to control on, since the GSOEP has very similar variables with the SHP and they also used matching within the context of sport-economics. Another criterium for the selection of the regressors is based on my theoretical model I formulated back in section 2.2.2. The estimation-results can be found in Table 12 Appendix E, since equation (21) has only the main purpose of estimating propensity scores to match on. In general, when it comes to the model specification for matching in order to get a causal effect, a researcher needs to include covariates that could influence the assignment to treatment and also variables that could influence potential

outcomes of an individual (Rubin 1977, 2005). All in all, most of the results are in line with the findings of the paper mentioned above, even though the underlying populations are different[13]:

- The age of an individual is negatively associated with sports-participation across all models.
- Being a female is associated with practicing less sport across in 3 out of 4 models, almost like in the baseline model of Schüttoff et al. (2018).
- An individual doing sport in the *past* is positively associated with doing sport today for all 4 models. Interestingly, the coefficient is only statistically significant for *young* individuals. As time passes, the family seems to have less weight, as the magnitude of the coefficient for doing sport as a child decreases for normally weighted adults and the other environmental factors seem to affect the individual stronger, as can be seen through the positive and significant association between other leisure-activities and sports-participation in model 1.
- Next, by looking at the personal health-related variables like the current health satisfaction or the health-improvement variable, we see that the state of being healthy seems to be less relevant than the gain of health in the last 12 months, regardless of the individual's being young, old, overweight or underweight. Furthermore, improvement in health in the past 12 months is always positively associated with greater sport activity, while feeling healthy will generally be negatively associated with being involved in weekly sports activities. This would imply that the motivation of doing sport comes not from being in a healthy state, but rather from the feeling of gaining more health. Lastly, it is to note that the coefficient for the young on health improvement may be positive, but the magnitude is - relative to the other subsamples - much smaller. This would hint to the fact that children would not be as motivated by the extrinsic motivation of gaining health, as other types of individuals.
- Financial incentive - claimed to be of main interest for doing sport or not according to the neoclassical theory - seem to play a secondary role, since the coefficients are not significant. Interestingly, income has a different effect, depending on the chronic health-status of an individual. While having no long-term health issues (normal weight) is positively associated with income, over- or underweight individuals exhibit negative associations with sports participation.

[13] Schüttoff et al. (2018) use adolescents aged 18 to 19, which would be similar to my second model (see appendix), even though I used individuals of the age 14 to 20.

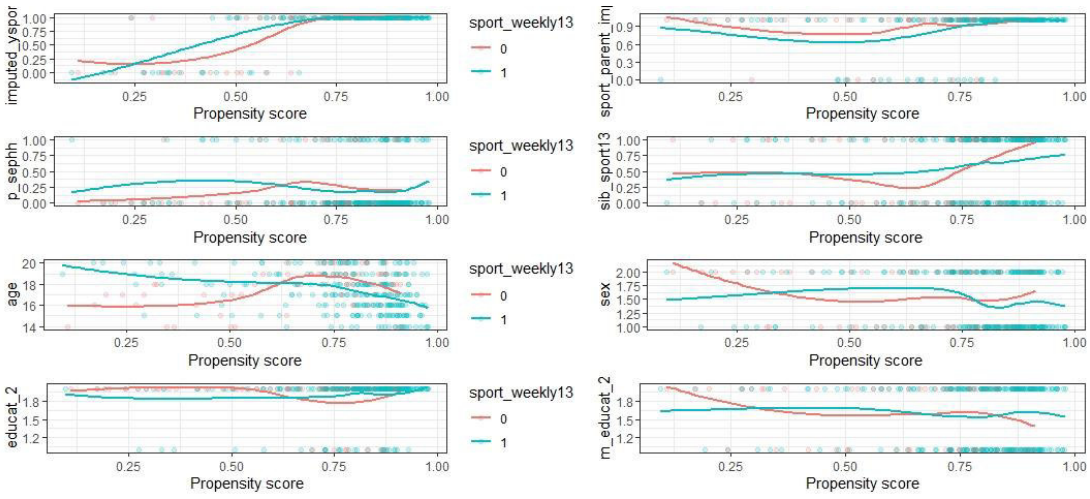
Table 5: Average Treatment Effect on the Treated (ATT) and Average Treatment Effect (ATE) Matching-Estimators

Matching-Estimators	**estimated ATT**	p-value ATT	**estimated ATE**	p-value ATE
Normal Adult	0.16192	0.28858	0.085047	0.56102
Normal Young	0.34638	0.16527	0.30288	0.17509
Overweighted	0.19847	0.40868	0.020725	0.92109
Underweighted	0.31915	0.6103	0.15517	0.78541

In order to get to the estimates in Table 5 above, the propensity scores were estimated through the logit-regressions we saw earlier in equation (16), namely $Health - Improvement\ in\ 2014_i = \beta_0 + \beta_1 * Weekly\ Sport\ in\ 2013_i$, *after* the matching-algorithm determined which treatment-observation is matched with the control-group member.

When looking at the signs of the coefficients in Table 6, we see that doing sports is *always* associated with gaining more health, independently if the individuals are young, old, overweight or underweight, but none is statistically significant. Interestingly, the gains in health due to sport are the highest for young individuals that are 20 or younger. Since we saw earlier in the logit regressions that children are the group of individuals that are the least motivated by health improvements, this result seems rather counter-intuitive. One explanation for this may be found when looking back at the logit-regression in Table 12, Appendix E. There, model 2 had the most amount of significant observables compared to all the other models. Since matching works better when having many good predictors, it is less surprising to find the lowest p-values on the matching-estimators of the young individuals at the p-values (Stuart, 2010).

Figure 2: Balance Test by Visual Inspection of the Propensity Scores for some Covariates in Matching-Regression



Note: The package used for the above Figure 2 is MatchIt, whereas the matching results are from the Matching-package. Therefore, the 1:1 matching results from the visual check may differ, because the algorithms behind the functions are programmed differently. Nevertheless, the basic idea of checking covariate balance stays valid and this is a way to verify it.

To correctly assess whether the matching was successful, one needs to check how balanced the covariates get *after* the matching-algorithm occurred. This balancing has the purpose of creating similar treatment- and control-groups such that - ideally - both only differ in whether they got treated or not and, thus, uncover the causal effect of sport on health-improvement. There are different possibilities to check the covariate balance after matching. The visualization as showed in Figure 2 is one possible way on determining this. Basically, if the treatment and control group have very similar means for each x-variable at each propensity-score value then we can say that matching was applied successfully (Ejdemyr, 2020). The above graph is just an extract, but - for all 4 models built - the red and blue line's means are only satisfactory for less than half of the regressors[14]. The calculated absolute standardized differences (ASD) in Appendix J, Table 15 - another way to assess the balance *after* matching - back up this conclusion: while - *on average* - normally weighted adults and young individuals have acceptable ASDs below a certain threshold of 0.1 (Normand et al., 2001), this is not the case for overweight and underweight individuals. Ideally ASDs indicate a successful matching when their values are close to zero, which is not the case (Rubin, 2001; Stuart, 2010). To get better results for the matching-estimators, I therefore change the model-specifications and build different logit-regressions for each sub-sample by including interaction- and squared-terms, but also other variables. Besides finding a positive and significant coefficient, the remaining models did not achieve substantially better results than the baseline model, which is why I use the results of it for the construction of my *Health-Expectation due to Sport*-variable - as described in subsection 3.2.1 - and proceed to the final MNL-model. All other models can be found in Appendix G.

Table 6: Estimation of Multinomial-Logistic-Regression.

Dependent variable: lifecycle choice-sequence for weekly Sport			
	Sport only as Adult	Sport only as a Child	Never Sport
	(1)	(2)	(3)
Parents doing Sport	1.000	-0.318	-0.495
	(0.626)	(0.291)	(0.349)
Parents Sport when Young	0.641	0.131	-0.373
	(0.567)	(0.399)	(0.462)
Mother: Secondary Education	-0.617	-0.207	-0.448
	(0.599)	(0.426)	(0.477)
Mother: Tertiary Education	-0.925	-0.288	-0.181
	(0.649)	(0.463)	(0.519)
Father: Secondary Education	1.043	-0.398	0.147
	(1.075)	(0.494)	(0.654)

[14] Check Appendix I, Figure 19 for the remaining covariates for the normally weighted young individuals.

Father: Tertiary Education	0.953	-0.547	-0.594
	(1.080)	(0.502)	(0.677)
2 People living in Hh	-0.609	-0.068	0.021
	(0.549)	(0.436)	(0.838)
3 People living in Hh	-0.322	0.533	0.973
	(0.592)	(0.483)	(0.852)
4 People living in Hh	-0.407	0.318	0.496
	(0.581)	(0.484)	(0.862)
5 People living in Hh	-0.111	0.574	0.609
	(0.618)	(0.519)	(0.886)
6 or more People living in Hh	-0.657	0.794	0.650
	(0.820)	(0.594)	(0.959)
Age	0.079***	-0.053*	-0.072*
	(0.030)	(0.028)	(0.043)
Gender: Female	0.554**	0.471**	0.561**
	(0.271)	(0.197)	(0.254)
Number of Doctor Consultations: Last 12 Months	-0.047	0.089***	0.003
	(0.044)	(0.018)	(0.037)
Being Underweighted	-0.209	-0.563	-0.747*
	(0.475)	(0.369)	(0.435)
Being Overweighted	0.136	-1.410***	-1.680***
	(0.363)	(0.326)	(0.496)
Being Obese or Severly Obese	-1.205	-1.587***	-1.144
	(1.050)	(0.605)	(0.750)
Hours worked per Week	0.0002	0.016***	0.012
	(0.007)	(0.006)	(0.007)
Main Income Contributor in the Hh	-0.293	0.073	-0.318
	(0.464)	(0.353)	(0.601)
Personal Education: Secondary Educ	0.245	-0.760***	-1.465***
	(0.367)	(0.282)	(0.383)
Personal Education: Tertiary Educ	0.045	-1.357***	-1.484***
	(0.517)	(0.392)	(0.556)
Number of Hours on a Screen per Day	-0.096	0.045	0.099**
	(0.077)	(0.040)	(0.040)
Feeling Unhealthy: Last Year	0.115	0.160	0.610*
	(0.399)	(0.298)	(0.331)
Health-Expectation due to Sport	-0.082	-1.513***	-1.183***
	(0.122)	(0.127)	(0.161)
Constant	-5.871***	3.488***	2.953*
	(1.766)	(1.151)	(1.587)
Akaike Inf. Crit.	1,834.112	1,834.112	1,834.112
Observations		1068	
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 6 shows the estimated results from the MNL. From the subsection 3.2.2 we know that the *signs* of the coefficient will hint at the relative valuation of a covariate, that is whether an

individual is more / less likely to choose a particular choice sequence. In combination with the p-values, the MNL-estimates will ultimately allow me to test my four hypotheses.

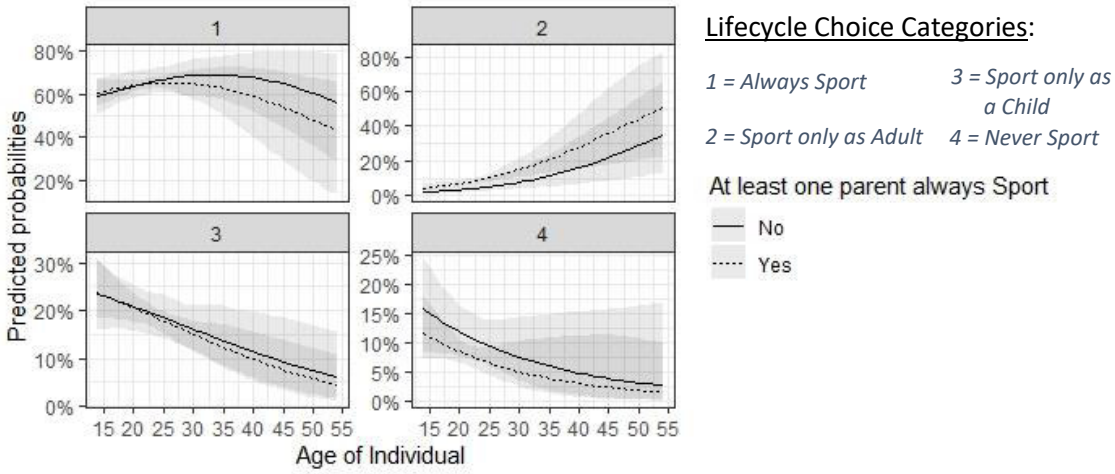
As we can see from the MNL-results, both coefficients on the parents' sport activity are not significant on any choice sequence. Hence, there is *not enough evidence* to confirm my *HFH*, that children form a habitual preference for sport, due to their parents' preference for sport. However, the signs of the coefficients hint at an existing intergenerational process. For instance, the *Parents doing Sport* is *positive* for individuals that started doing sport as an adult, meaning that these types of individuals gain - on average - relatively more utility from their parents' current sport activity relative to individual that have a habit of doing sport over their whole life, ceteris paribus. The lagged variable *Parents Sport when Young* seem to confirm the tendence of an intergenerational process. As Schüttoff et al. (2018) note in their study, including lagged variables into a regression allows to control for unobservable behavioral attributes, such as the *inherent preferences* of the parents for sports. Having a parent that did sport in their youth will increase the likelihood of an individual choosing to do sport in their childhood. Thus, social factors such as the previous generation's current or past sport activity seem to motivate individuals from the next generation into doing a weekly sport-activity sooner or later in their lifecycles.

Concerning the *HIH*, the sign of the variable *Health-Expectation due to Sport* is of main interest. We can see that individuals having always done sport throughout their whole life gain relatively more utility compared to all other defined categories, even more than individuals starting with sport later in their lives, which I speculated to have a positive sign. This result would hint at the fact that individuals who developed a habitual behavior to always do sport during their lives, are also sensitive to extrinsic factors. The coefficients are even statistically significant for individual's that stop doing sport or never did sport, which would *partially confirm my HIH* that individuals are motivated by the expected gains of doing sport in the future. However, as we saw with the matching results, this variable that directly tests the health-expectations of individuals is based on rather bad matching results. To overcome this issue, *other* health-variables can be used such as a short-term illness, proxied by doctor consultations. In this case, an individual is more likely to stop sport when older. Secondly, having a chronical disease - proxied by being obese, underweight or overweight - will result in individuals being less likely to never choose *or* to quit doing sport, relative to always doing sport and certeris paribus. Those statistically significant results *and* the fact that individuals with higher personal educational-status are less likely to never do *or* quit sport when being an adult - certeris paribus and relative to individuals that are always doing sport - *give additional evidence for my HIH*.

For reasons why an individual may stop doing sport, financial incentives - proxied by the amount of hours worked per week - have a statistically significant impact on an individual, being more likely to *stop doing sport* when being an adult, certeris paribus. Additionally, for individuals that *never did sport* in their lives, alternative leisure activities working as a substitute - proxied by the amount of hours watching TV, being on the internet or gaming - have a positive and statistically significant impact on this type of individuals, certeris paribus. Hence, this *confirms the last two hypotheses* I formulated initially in subsection 2.3.

Finally, it is to be mentioned that - even if my hypotheses could be answered by looking at the direction of the signs and statistical significance of the coefficient - oftentimes, the literature points out that solely looking at coefficients in choice models is not sufficient, especially because the estimates are interpreted as a change in the odds of the dependent variable, which is not always intuitive (Hanmer & Ozan Kalkan, 2013; Long & Freese, 2001). However, as demonstrated in subsection 2.1.2, an individual's choice can also be expressed as a probability. This can be advantageous if a researcher is - for example - interested whether there is a difference in the (marginal) effect of having a parent always doing sport on the lifecycle sport-choice over the individual's whole range of ages, reflecting how sport preferences can dynamically change over time and across groups (Neumann, 2020).

Figure 3: Predicted Lifecycle Sport-Choice-Probabilities for Individuals with at least one Parent always doing Sport VS. No Parent doing Sport during their Life across Ages



Note: Results are based on a slightly different model, where I simply modified the MNL in Table 6 by merging the variables Parents doing Sport and Parents Sport when Young into a single variable, now representing the lifecycle-choice of the parents. See also Appendix K, Table 18 for the estimation of this model. The gray areas around the black lines represent the 95% confidence intervals. For the same graph but using the original MNL in Table 6 as a baseline and the parents' current or past sport choice, see Appendix L, Figure 20 and Figure 22 respectively.

As Figure 3 demonstrates, there is a difference in the choice probability of selecting a certain alternative depending on an individual's age and the lifecycle-choice of the previous generation. As shown in Figure 26 from Appendix L, those differences in the choice-probabilities are - for

some lifecycle-choices - statistically significant on the 10% significance level, which would bring evidence that the transmission of sport-preferences across generations is *not constant over time*. More precisely, it seems that individuals may be more inclined to do sport later in life, if at least one parent did always sport. Also, this would *speak against my HFH*, since we see that there is no statistical significance when the individuals are being young.

5. Discussion and Conclusion

My study finds evidence that both, financial- as well as social-factors are important during an individual's sport-preference formation over an individual's lifecycle in Switzerland. While parents have been shown *not* to be responsible for an individual forming a habitual behavior for always doing sport (rejection of HFH), the sports attitude transmission across generations increases over time for children starting to do sport as adults, as opposed to individuals that always do sport, for which it decreases. Thus, the intergenerational transmission process is shown to be heterogeneous, or time-invariant. Considering financial aspects, the number of hours worked per week are shown to increase the likelihood of quitting weekly sport activity later in life, showing the relative importance of monetary incentives in an individual's decision-making process when quitting sport in adulthood (confirming my third hypothesis). For the aspect of maintaining a higher health-status, I find that higher expectations of gaining health, as well as having long-lasting chronical weight problems increases the likelihood to stop doing sport (confirming my HIH). At this point, it is to note that - while maintaining higher health was introduced to be a neoclassical motive, since an investment in health would imply higher productivity and more time at work gaining money - there is also a psychological aspect of maintaining a good health-status, because an individual may have a preference for improvement (Chapman, 2000). Hence, the health aspect reflects both a financial as well as a behavioral factor. Lastly, the number of hours passed watching TV, being on the internet or gaming, confirms the relative importance of other leisure activities acting as substitutes to increase the likelihood to never do sport (confirming my fourth hypothesis).

To assess the reliability of my final results, one needs to remember that the choice probabilities in RUMs depend on how the unobservable part of the utility is specified. First, ε_{ij} in (6) is assumed to be identically and independently (extreme value[15]) distributed (iid) in the MNL, meaning that error-term is assumed to be uncorrelated across the alternatives of the

[15] An extreme value distribution is defined as having fatter tails relative to a normal distribution. Assuming an extreme value distribution enables a slightly better modelling, relative to a normal distribution (Train, 2003, Chapter 3, p. 39).

choice set, while also having the same variance across alternatives. Additionally, because I analyze a lifecycle and observe choices over multiple periods in time, my dependent variable is technically a *sequence of choice*. Therefore, I also need the independence of choices over time for my error-terms, otherwise the model estimates will be biased. Thus, one needs to carefully think about the correlation of the error-terms across the different alternatives and over time (Train, 2003). An example for such a disturbing factor that would bias my results is the "worrying about future-health" factor. As it was argued during the description of my HIH in subsection 2.3, the presence of the endowed "health capital" starts to depreciate with higher speed as an individual gets older. Since the "worrying factor" increases over time, the inclusion of a constant in the MNL would not be enough to absorb it, because I pooled all individuals together for each of the four lifecycle choices and, thus, would not account for their heterogeneity in this "unobserved" factor over time. Another example that affect the iid assumption would be measurement errors. It is imaginable that some *parents* answer sport-question in their childhood imprecisely, typically leading to a bias towards zero for my estimates (Hyslop & Imbens, 2001). However, there are extended versions of the MNL, like the mixed logit, nested logit or a probit that allow for correlation patterns[16] across alternatives and time, enabling a less restrictive modelling approach by relaxing the iid assumption and thus, diminishing bias (Train, 2003). Alternatively, a re-specification of the model with the inclusion of interactions, such as sociodemographics could also be used in order to avoid the violation of the iid-assumption (Train, 2003). However, since this dataset did not contain any alternative-specific variables to interact with, the inclusion of sociodemographics to eliminate some random taste-variation could not be implemented. Using a score-test to check for heteroscedasticity, we see that my MNL violates a part of the iid assumption (Appendix M).

Secondly - and being a specific property of the MNL - choice probabilities between two particular alternatives *i* and *j* are assumed to be unaffected by the reduction / introduction of a new alternative *z*. This assumption is called the independence of irrelevant alternatives (IIA). If the IIA is violated, I will end up with biased probabilities. Ultimately, the main goal when constructing a choice model is to try and achieve an error term that is “white noise”, meaning that a scientist should be able to specify the observable utility V_{ij} in a way that the error-term becomes uncorrelated across alternative and time (Train, 2003). Only then it will be possible to capture the true dynamics behind the repeated choice process of doing sport and make precise predictions on how individuals behave in the context of sports participation over their lifecycle.

[16] In the case of mixed logit models, this is done by including random coefficients, which account for random taste-variation (Train, 2003, Chapter 3, p. 48).

As explained, the MNL is likely to not fulfill the required assumptions and there is room for alternative modelling strategies. Extensions of the MNL, such as a mixed logit model or a probit model, could *theoretically* fix the issues talked about in this conclusion. Unfortunately, their implementation requires alternative-specific variables, which I do not have (StataCorp, 2019).

The poor matching results I found also need some justifications. We saw that the balance *after* matching was insufficient for many covariates, making it difficult to assess whether the gains in health were in fact due to doing sport. In addition, the *conditional independence assumption* - the key assumption when matching is used - is likely to be violated (Heckman et al., 1997). For instance, we can think about variables such as the sport-activity of the individual's partner or cultural background, which could influence the assignment to treatment or potential outcomes and - ultimately - resulting in me not being able to make a causal statement. Nevertheless, since the purpose of my matching-strategy was never to do causal statements, but rather, to create a variable that is able to measure the improvement in health due to sport in order to answer the HIH originally formulated in subsection 2.3, simultaneously capture an individual's reference point (see equation (22)), as well as eliminating the mutual dependence problem between sport and health, the efforts seem to be justified.

Finally, the lacking amount of total observations in my dataset is another problem that is limiting this analysis. For example, the sport literature points out that children are significantly more physically active than young adults (between 20-30) and that physical activity starts to increase again from around 33 years of age (García et al., 2011; Humphreys & Ruseski, 2010; Stamatakis & Chaudhury, 2008). Wicker et al. (2009) even recommend classifying individuals according to age groups to get better estimates. By running an MNL with a subsample of individuals aged between 20 and 30 (see Appendix K, Table 17), the magnitude of the coefficients such as 13.977 or 14.553 turned out hinting towards a perfect *prediction problem*, meaning that some independent variables have none / too little variation for some categories of the dependent variable (Long & Freese, 2001, p. 145). Thus, even if such an analysis would seem to lead to better results *in theory*, this dataset is not able to support such an analysis in practice. Hence, a bigger sample would allow for more precise estimates, as well as more statistical power.

All in all, this thesis showed what capabilities lie within discrete choice modelling to determine individuals' formation of sport-preferences. To fully unlock their potential and augment internal validity, extended modelling approaches as well as a bigger sample size would help to increase the insights in this domain and is up to further research.

6. References

Angrist, J. D. (1990). Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *The American Economic Review*, *80*(3), 313–336.

Arrow, K. J. (1958). Utilities, Attitudes, Choices: A Review Note. *Econometrica: Journal of the Econometric Society*, *26*(1), 1–23.

Arrow, K. J. (1959). Rational Choice Functions and Orderings. *Economica*, *26*(102), 121–127.

Bauman, A. E., Sallis, J. F., Dzewaltowski, D. A., & Owen, N. (2002). Toward a Better Understanding of the Influences on Physical Activity: The Role of Determinants, Correlates, Causal Variables, Mediators, Moderators, and Confounders. *American Journal of Preventive Medicine*, *23*(2), 5–14.

Becker, G. S. (1965). A Theory of the Allocation of Time. *The Economic Journal*, *75*(299), 493–517.

Becker, G. S., & Mulligan, C. B. (1997). The Endogenous Determination of Time Preference. *The Quarterly Journal of Economics*, *112*(3), 729–758.

Bisin, A., & Verdier, T. (2000). "Beyond the Melting Pot": Cultural Transmission, Marriage, and the Evolution of Ethnic and Religious Traits. *The Quarterly Journal of Economics*, *115*(3), 955–988.

Black, S. E., Devereux, P. J., Lundborg, P., & Majlesi, K. (2015). Poor Little Rich Kids? The Determinants of the Intergenerational Transmission of Wealth. *2015/6*.

Bundesamt für Gesundheit. (2018). *Gesundheitsförderung und Prävention in der frühen Kindheit*. Retrieved June 01, 2020, from https://www.npg-rsp.ch/fileadmin/npg-rsp/Themen/Fachthemen/BAG_2018_fruehe-Kindheit.pdf

Bundesamt für Statistik. (2018). *Schweizerische Gesundheitsbefragung 2017*. BFS: Neuchâtel, Switzerland, page 13. Retrieved September 02, 2020, from https://www.bfs.admin.ch/bfs/de/home/statistiken/gesundheit/erhebungen/sgb.assetdetail.6426300.html

Cabane, C., & Lechner, M. (2015). Physical Activity of Adults: A Survey of Correlates, Determinants, and Effects. *Jahrbücher für Nationalökonomie Und Statistik*, *235*(4-5), 376–402.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Campante, F., & Yanagizawa-Drott, D. (2015). *The Intergenerational Transmission of War* (No. w21371). National Bureau of Economic Research.

Cawley, J. (2004). An Economic Framework for Understanding Physical Activity and Eating Behaviors. *American Journal of Preventive Medicine*, *27*(3), 117–125.

Chapman, G. B. (2000). Preferences for Improving and Declining Sequences of Health Outcomes. *Journal of Behavioral Decision Making*, *13*(2), 203–218.

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1553–1623.

Colman, G., & Dave, D. (2013). Exercise, Physical Activity, and Exertion over the Business Cycle. *Social Science & Medicine (1982)*, *93*, 11–20.

Crowson, H. M. (2020). *Multinomial Logistic Regression using Stata: Covid-19 threat perceptions from Pew data*. [Powerpoint slides]. Retrieved September 08, 2020, from https://drive.google.com/open?id=15kGpOdKZeIxpkkgZokevv8AZasoMOByq

Currie, J. (2009). Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development. *Journal of Economic Literature*, *47*(1), 87–122.

Currie, J., & Moretti, E. (2003). Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings. *The Quarterly Journal of Economics*, *118*(4), 1495–1532.

Debreu, G. (1954). Representation of a Preference Ordering by a Numerical Function. *Decision Processes*, *3*, 159–165.

Deforche, B. I., De Bourdeaudhuij, I. M., & Tanghe, A. P. (2006). Attitude Toward Physical Activity in Normal-weight, Overweight and Obese Adolescents. *Journal of Adolescent Health*, *38*(5), 560–568.

Dehejia, R. H., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062.

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2012). The Intergenerational Transmission of Risk and Trust Attitudes. *The Review of Economic Studies*, *79*(2), 645–677.

Downward, P., & Riordan, J. (2007). Social Interactions and the Demand for Sport: An Economic Analysis. *Contemporary Economic Policy*, *25*(4), 518–537.

Edwards, W. (1954). The Theory of Decision Making. *Psychological Bulletin*, *51*(4), 380-417.

Ejdemyr, S. (2020). *R Tutorial 8: Propensity Score Matching*. Retrieved September 02, 2020, from https://sejdemyr.github.io/r-tutorials/statistics/tutorial8.html

European Commission (2014). *The 2014 EU Summit on Chronic Diseases. Brussels, 3 and 4 April 2014, Conference Conclusions*. Retrieved September 02, 2020, from https://ec.europa.eu/health//sites/health/files/major_chronic_diseases/docs/ev_20140403_mi_en.pdf

Fernández, R., Fogli, A., & Olivetti, C. (2004). Mothers and Sons: Preference Formation and Female Labor Force Dynamics. *The Quarterly Journal of Economics*, *119*(4), 1249– 1299.

Fisher, I. (1930). *The Theory of Interest: As Determined by Impatience to Spend Income and Opportunity to Invest it*. New York: Macmillan.

Flegal, K. M., Carroll, M. D., Kit, B. K., & Ogden, C. L. (2012). Prevalence of Obesity and Trends in the Distribution of Body Mass Index Among US Adults, 1999-2010. *JAMA : The Journal of the American Medical Association*, *307*(5), 491–497.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, *40*(2), 351–401.

García, J., Lera-López, F., & Súarez, M. J. (2011). Estimation of a Structural Model of the Determinants of the Time Spent on Physical Activity and Sport: Evidence for Spain. *Journal of Sports Economics*, *12*(5), 515–537.

Glanz, K., Rimer, B. K., & Viswanath, K. (2008). *Health behavior and health education: Theory, research, and practice*. John Wiley & Sons.

Goodman, J., Hurwitz, M., Mulhern, C., & Smith, J. (2019). *O Brother, Where Start Thou? Sibling Spillovers in College Enrollment* (No. w26502). National Bureau of Economic Research.

Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, *80*(2), 223–255.

Grossman, M. (2000). The Human Capital Model. In Culyer, A. J., & Newhouse, J. P. (Eds.), *Handbook of Health Economics* (Vol. 1, pp. 347-408). Elsevier.

Hanmer, M. J., & Ozan Kalkan, K. (2013). Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models. *American Journal of Political Science*, *57*(1), 263–277.

Hansson, B. (1968). Choice Structures and Preference Relations. *Synthese*, *18*(4), 443–458.

Heckman, J.J. (1981). Heterogeneity and State Dependence. In Rosen, S. (Ed.), *Studies in Labor Markets* (pp. 91-140). University of Chicago Press.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of Economic Studies*, *64*(4), 605–654.

Houthakker, H. S. (1950). Revealed Preference and the Utility Function. *Economica*, *17*(66), 159–174.

Houthakker, H. S. (1965). On the Logic of Preference and Choice. In Tymieniecka, A.-T. (Ed.), *Contributions to Logic and Methodology in Honour of J.M. Bochenski* (pp. 193-207). Amsterdam.

Humphreys, B. R., McLeod, L., & Ruseski, J. E. (2014). Physical Activity and Health Outcomes: Evidence from Canada. *Health Economics*, *23*(1), 33–54.

Humphreys, B. R., & Ruseski, J. E. (2010). *The Economic Choice of Participation and Time Spent in Physical Activity and Sport in Canada* (No. 2010-14). University of Alberta, Department of Economics.

Humphreys, B. R., & Ruseski, J. E. (2011). An Economic Analysis of Participation and Time Spent in Physical Activity. *The B.E. Journal of Economic Analysis & Policy*, *11*(1).

Hyslop, D. R., & Imbens, G. W. (2001). Bias from Classical and other Forms of Measurement Error. *Journal of Business & Economic Statistics*, *19*(4), 475–481.

Janke, K., Propper, C., & Shields, M. A. (2013). Does Violent Crime Deter Physical Activity? *IZA Discussion Paper*, IZA Discussion Paper.

Johannesson, M., Östling, R., & Ranehill, E. (2010). The Effect of Competition on Physical Activity: A Randomized Trial. *The B.E. Journal of Economic Analysis & Policy*, *10*(1), 2555.

Johnston, D. W., Schurer, S., & Shields, M. A. (2013). Exploring the Intergenerational Persistence of Mental Health: Evidence from three Generations. *Journal of Health Economics*, *32*(6), 1077–1089.

Kjær, T. (2005). *A Review of the Discrete Choice Experiment - with Emphasis on its Application in Health Care*. Syddansk Universitet. Health Economics Papers, No. 1.

Koopmans, T. C. (1960). Stationary Ordinal Utility and Impatience. *Econometrica: Journal of the Econometric Society*, *28*(2), 287–309.

Köthenbürger, M., & Anderes, M. (2019). *KOF Prognose der Gesundheitsausgaben: Herbst 2019* (No. 141). KOF Studien. Retrieved September 11, 2020, from https://ethz.ch/content/dam/ethz/special-interest/dual/kof-dam/documents/Publications/No_141_Gesundheitsausgabenprog_Herbst_2019.pdf

Kremer, M., & Levy, D. (2008). Peer Effects and Alcohol Use among College Students. *Journal of Economic perspectives*, *22*(3), 189–206.

Kwak, C., & Clayton-Matthews, A. (2002). Multinomial Logistic Regression. *Nursing research*, *51*(6), 404–410.

Laibson, D. (1994). Self-Control and Saving. *Massachusets Institute of Technology mimeo*.

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics*, *112*(2), 443–478.

Laibson, D., & Zeckhauser, R. (1998). Amos Tversky and the Ascent of Behavioral Economics. *Journal of Risk and Uncertainty*, *16*(1), 7–47.

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, *74*(2), 132–157.

Lichtenstein, S., & Slovic, P. (1971). Reversals of Preference Between Bids and Choices in Gambling Decisions. *Journal of Experimental Psychology*, *89*(1), 46-55.

Long, J. S., & Freese, J. (2001). *Regression Models for Categorical Dependent Variables using Stata*. College Station, Texas: Stata press.

Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., & Murray, C. J. (2006). Global and Regional Burden of Disease and Risk Factors, 2001: Systematic Analysis of Population Health Data. *The Lancet (British Edition)*, *367*(9524), 1747–1757.

Manski, C. F. (1977). The Structure of Random Utility Models. *Theory and Decision*, *8*(3), 229-254.

Marschak, J. (1960). Binary Choice Constraints on Random Utility Indicators. In Arrow, K. (Ed.), *Stanford Synopsium on Mathematical Methods in the Social Sciences*. CA: Stanford University Press. Retrieved September 11, 2020, from https://cowles.yale.edu/sites/default/files/files/pub/d00/d0074.pdf

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic Theory*. New York, N.Y: Oxford University Press.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In Zarembka, P. (Ed.), *Frontiers in Econometrics* (pp. 105-142). New York: Academic Press.

McFadden, D. (2001). Economic Choices. *American Economic Review*, *91*(3), 351–378.

Morgenstern, O., & Von Neumann, J. (1953). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Neumann, M. (2020). *MNLpred - Simulated Predictions From Multinomial Logistic Models*. Retrieved September 09, 2020, from https://github.com/ManuelNeumann/MNLpred

Normand, S., Landrum, M., Guadagnoli, E., Ayanian, J., Ryan, T., Cleary, P., & McNeil, B. (2001). Validating Recommendations for Coronary Angiography following Acute Myocardial Infarction in the Elderly: A Matched Analysis using Propensity Scores. *Journal of Clinical Epidemiology*, *54*(4), 387-398.

O'Donoghue, T., & Rabin, M. (1999). Doing it Now or Later. *American Economic Review*, *89*(1), 103–124.

Petrucci, C. J. (2009). A Primer for Social Worker Researchers on How to Conduct a Multinomial Logistic Regression. *Journal of Social Service Research*, *35*(2), 193–205.

Rae, J. (1905). *The Sociological Theory of Capital: Being a Complete Reprint of the New Principles of Political Economy, 1834*. New York.

Reiner, M., Niermann, C., Jekauc, D., & Woll, A. (2013). Long-term Health Benefits of Physical Activity – a Systematic Review of Longitudinal Studies. *BMC Public Health*, *13*(1), 1–9.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, *66*(5), 688-701.

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*, *2*(1), 1–26.

Rubin, D. B. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine*, *127*(8 Pt 2), 757–763.

Rubin, D. B. (2001). Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology*, *2*(3), 169–188.

Rubin, D. B. (2005). Causal Inference using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Rubin, D. B. (2007). The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials. *Statistics in Medicine*, *26*(1), 20–36.

Saberi, M. (2017). *Discrete Choice Modelling: Basics and Theory*. [Video File]. Retrieved September 24, 2020, from https://www.youtube.com/watch?v=aPdedMTiFzo

Samuelson, P. A. (1937). A Note on Measurement of Utility. *The Review of Economic Studies*, *4*(2), 155–161.

Samuelson, P. A. (1947). *Foundations of Economic Analysis* (Vol. 80, Harvard Economic Studies). Cambridge: Harvard University Press.

Samuelson, P. A. (1948). Consumption Theory in Terms of Revealed Preference. *Economica*, *15*(60), 243–253.

Sari, N. (2014). Sports, Exercise, and Length of Stay in Hospitals: is there a Differential Effect for the Chronically Ill People? *Contemporary Economic Policy*, *32*(2), 247–260.

Schüttoff, U., Pawlowski, T., Downward, P., & Lechner, M. (2018). Sports Participation and Social Capital Formation During Adolescence. *Social Science Quarterly*, *99*(2), 683–698.

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.

Sen, A. (1973). Behaviour and the Concept of Preference. *Economica*, *40*(159), 241–259.

Simon, H. A. (1972). Theories of Bounded Rationality. *Decision and Organization*, *1*(1), 161–176.

Stamatakis, E., & Chaudhury, M. (2008). Temporal Trends in Adults' Sports Participation Patterns in England between 1997 and 2006: The Health Survey for England. *British Journal of Sports Medicine*, *42*(11), 901–908.

StataCorp. (2019). *Stata Choice Models Reference Manual, Release 16*. College Station, TX: Stata Press. Retrieved September 24, 2020, from https://www.stata.com/manuals/cm.pdf

Stock, J. H., & Watson, M. W. (2015). Introduction to Econometrics (Updated 3rd Edition, Global ed., The Pearson Series in Economics). Harlow: Pearson Education.

Strotz, R. H. (1955). Myopia and Inconsistency in Dynamic Utility Maximization. *The Review of Economic Studies*, *23*(3), 165–180.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A review and a Look Forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1-21.

Thaler, R. H. (1990). Anomalies: Saving, Fungibility, and Mental Accounts. *Journal of Economic Perspectives*, *4*(1), 193–205.

Thurstone, L. L. (1927). A Law of Comparative Judgment. *Psychological Review*, *34*(4), 273-286.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

Voorpostel, M., Tillmann, R., Lebert, F., Kuhn, U., Lipps, O., Ryser, V.-A., Antal, E., Monsch, G.-A., Dasoki, N., & Wernli, B. (2018). *Swiss Household Panel Userguide (1999-2017), Wave 19, December 2018*. Lausanne: FORS.

Walker, J., & Ben-Akiva, M. (2002). Generalized Random Utility Model. *Mathematical Social Sciences*, *43*(3), 303–343.

Wicker, P., Breuer, C., & Pawlowski, T. (2009). Promoting Sport for All to Age-specific Target Groups: The Impact of Sport Infrastructure. *European Sport Management Quarterly*, *9*(2), 103–118.

Woolger, C., & Power, T. G. (1993). Parent and Sport Socialization: Views from the Achievement Literature. *Journal of Sport Behavior*, *16*(3), 171.

World Health Organization. (2020a). *Body Mass Index – BMI*. Retrieved September 10, 2020, from https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi

World Health Organization. (2020b). *Obesity and Overweight*. Retrieved September 10, 2020, from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

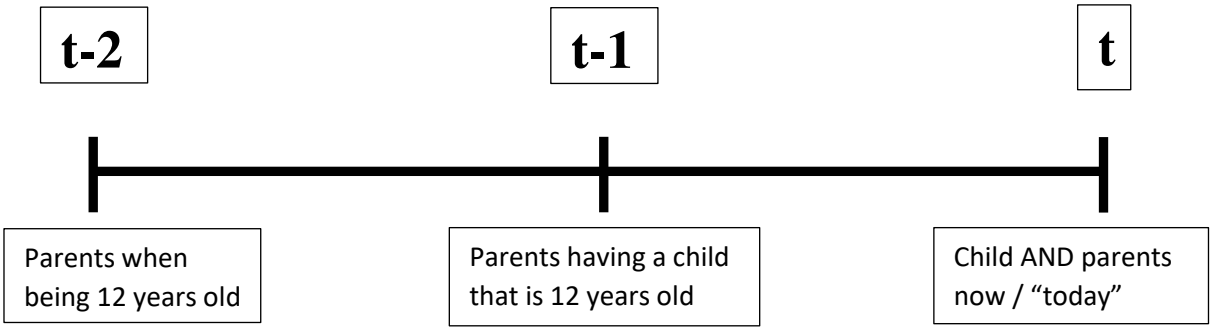
Appendix

A) Theoretical Model: Why do Individual do Sport?

The following Figure 5 is a flowchart, which reflects how an individual's sport preference formation is affected by different factors, across time and generations. The main inspiration for the construction of this flowchart was the model of Bisin and Verdier (2000), in which preferences of the *next* generation are influenced by both, the parents - i.e. the *previous* generation - and by *external factors* such as friends, the region or - in the sport context and following Grossman (1972, 2000) - by the prospect of gaining additional health in the future.

The graph can be split into *two parts*. In the *first part* - marked by various shades of *blue* within the flowchart - the parents' preferences are visualized and then transitions - in the *second part*, marked by various shades of *orange* - into the (next) children-generation and shows their evolution of preferences. In order to better link the theory with the empirical part in section 3 of this work, the flowchart was divided into *three periods*, for which I had concrete data:

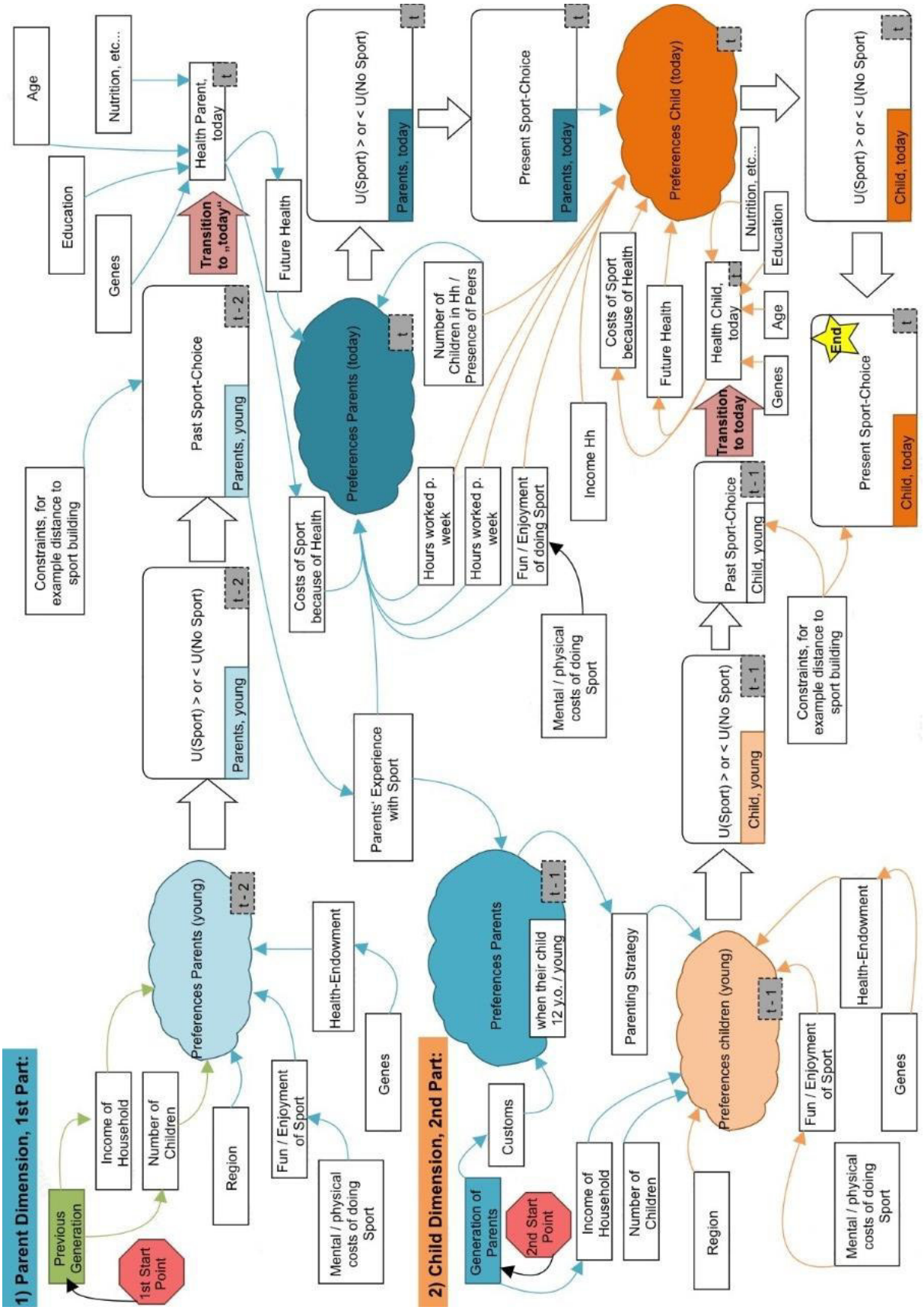
Figure 4:The three Periods of References for the Flowchart in Figure 5



The reading direction begins in the top-left corner (see the *red* octagon with the inscription "1st Start Point"[17]) and ends at the bottom right corner (marked by the *yellow star*). Furthermore, I used different shades of the colors blue and orange to symbolize that there is a transition from the past to the present for both, parents and their children. Note that - of course - not all factors that may influence an individual's sport-preferences could possibly be included in Figure 5. The purpose of this graph is to give the reader a better overview on how the sport-preference formation of an individual is formed and dynamically evolves over time.

[17] There is also a second starting point in the bottom-left corner (see the second *red* octagon with the inscription "2nd Start Point"), if the reader wishes to concentrate solely on the (next) children-generation.

Figure 5:Flowchart on the Formation of Preferences across Generations and Time



Note: This Figure was constructed with the help of the app Flowdia.

B) List of Variables

Table 7: Detailed List of Variables used for the Logit- & Matching-Regressions

Full Name	Description	Abbreviation
Weekly Sport	Dummy, that takes the value of "1" if an individual does sport on a weekly basis currently (in 2013). My dependent variable in the logit-regression.	sport_weekly13 / $sport_{iw}$ in equation (21), (25), (26) and (27)
Age	This variable informs about the age of the individual in the current year (2013) (Voorpostel et al., 2018).	age
Mother Age	This variable informs about the age of the individual's mother in the current year (2013) (Voorpostel et al., 2018).	imput_mage13_2 / m_age_{iw} in equation (21)
Gender	Gender of the Individual. The reference category is being a male (Voorpostel et al., 2018).	sex
Education of Mother	Similarly to Schüttoff et al. (2018), this dummy-variable gives information whether an individual reached an educational-level that allows him to access tertiary education or not (A-level). This criterion is fulfilled if an individual has at least a maturity-degree. The educational attainment is based on the ISCED-classification. The reference category is when the mother has a degree higher than maturity (Voorpostel et al., 2018).	m_educat_2 / m_educ_{iw} in equation (21), (25), (26) and (27)
Personal Education	Similarly to Schüttoff et al. (2018), this dummy-variable gives information whether an individual's mother reached an educational-level that allows her to access tertiary education or not (A-level). This criterion is fulfilled if the mother has at least a maturity-degree. The educational attainment is based on the ISCED-classification. The reference category is when the degree is higher than maturity (Voorpostel et al., 2018).	educat_2 / $educ_{iw}$ in equation (21), (25), (26) and (27)
Work	This dummy-variable denotes whether the individual is currently working or not. Not working is the reference category.	work
Parents Separated	This variable only denotes if the child's original parents are separated or not. This variable does ***not*** account for the current civil-status of the parents. It can very well be, that the mother / father is re-married to another husband / wife in the current year of 2013.	p_sephh / p_sep_{iw} in equation (21)
Leisure-Activity Index	For each individual, this variable aggregates over all the leisure-activities that an individual does on a weekly basis in the current year (2013) besides sport. Those leisure activities include: meeting friends, going to the restaurant / bar, playing music, gardening, going out clubbing, attending to sport events, going to the cinema, reading books, going to the theater, attending to the opera, drawing / sculpturing, doing photography and going to the museum (Voorpostel et al., 2018).	act_index13 / $leisure\_index_{iw}$ in equation (21) and (26)
Number of hours on a screen per day (Internet, TV, Gaming)	This variable aggregates the screen time passed on TV, video games and on the internet per day in hours (Voorpostel et al., 2018).	aggr_hrs_screentime_day13 / $leisure\_screen_{iw}$ in equation (21)
Number of Doctor Consultations: Last 12 Months	This variable indicates how many times an individual went to the doctor in the last 12 months (Voorpostel et al., 2018). This is a lagged variable.	year_doct2 / $doct\_cons_{iw}$ in equation (21), (25), (26) and (27)
Current Health-Satisfaction	Current health-satisfaction on a scale from 0 to 10 (Voorpostel et al., 2018).	hsat
Health-Improvement 2013: Last 12 Months	This variable gives - on a scale from -5 to 5, where "0" is the reference-point on the scale - if an individual experienced a health improvement in the last 12 months or not (and how strong) (Voorpostel et al., 2018). This is the lagged-outcome variable.	h_impr12 / $h\_impr\_lagged_{iw}$ in equation (21), (25), (26) and (27)
Health-Improvement 2014: Last 12 Months	This variable gives - on a scale from -5 to 5, where "0" is the reference-point on the scale - if an individual experienced a health improvement or not (and how strong) in 2014 relative to 2013. This is the final dependent variable for which I try to explain how much of the health-improvement is due to sports (Voorpostel et al., 2018).	h_impr13
Number of Siblings	This variable denotes the number of siblings currently present in the household, that is in 2013. This variable does not give any information	numb_sib / $nber\_sib_{iw}$ in equation (21)

	about siblings that already left the household and live separately on their own!	
Sport as a Child	This dummy-variable denotes whether the individual did sport in his youth or not. Doing no sport in the youth is the reference category (Voorpostel et al., 2018).	imputed_ysport2 / $ysport_{iw}$ in equation (21)
Parents doing Sport	This dummy-variable denotes whether *at least one* of the parents does sport currently (2013). The reference category is when none of the parents do sports.	sport_parent_imput2 / p_sport_{iw} in equation (21)
Siblings doing Sport	This dummy-variable denotes whether at least one of the siblings of an individual that is also analyzed in this dataset is doing sport or not. I need this variable for the SUTVA to hold (Rubin, 2005).	sib_sport13
Unhealthy Parents: last Year	This dummy-variable denotes whether at least one of the parents was being unhealthy in 2012. This is a lagged variable.	imput_unhealthy_parents / $p\_unhealthy_{iw}$ in equation (21), (25), (26) and (27)
Hh-Income	This is the *natural logarithm* of the household-income in 2013.	imputed_hhinc2_ln / hh_inc_{iw} in equation (21), (25), (26) and (27)
Region-Dummies	This is a categorical variable which denotes the region in which the individual lives in. In total, there are 7 categories: Lake Geneva region, Middleland, North-west Switzerland, Zurich, East Switzerland, Central Switzerland, and Ticino (Voorpostel et al., 2018).	region

Table 8: Detailed List of Additional Variables used for the Multinomial-Logit-Regression

Full Name	Description	Abbreviation
Parents Sport when Young	This dummy-variable denotes whether *at least one* of the parents did sport as a child, that is, when being 12 years old (Voorpostel et al., 2018). Thereference category is when none of the parents did sport as a child.	sport_parent_young / p_ysport_{iw} in equation (25), (26) and (27)
Education of Father	This is a categorical variable that indicates whether the individual's father highest educational attainment is either primary, secondary or tertiary education (Voorpostel et al., 2018). The reference category is primary education.	f_educat_3
Education of Mother	This is a categorical variable that indicates whether the individual's mother highest educational attainment is either primary, secondary or tertiary education (Voorpostel et al., 2018). The reference category is primary education.	m_educat_3
Personal Education	This is a categorical variable that indicates whether the individual's highest educational attainment is either primary, secondary or tertiary education (Voorpostel et al., 2018). The reference category is primary education.	educat_3
2 People living in Hh	This is a dummy-variable indicating whether the individual lives in a Household with 2 people. The reference category is a single-household.	nbpers_2
3 People living in Hh	This is a dummy-variable indicating whether the individual lives in a Household with 3 people. The reference category is a single-household.	nbpers_3
4 People living in Hh	This is a dummy-variable indicating whether the individual lives in a Household with 4 people. The reference category is a single-household.	nbpers_4
5 People living in Hh	This is a dummy-variable indicating whether the individual lives in a Household with 5 people. The reference category is a single-household.	nbpers_5
6 or more People living in Hh	This is a dummy-variable indicating whether the individual lives in a Household with 6 or people. The reference category is a single-household.	nbpers_6_or_more
Being Underweighted	This is a dummy created on the basis of the BMI, which I calculated through the weights & height of an individual. If BMI < 18.50, then the individual is considered underweighted. The reference category is having a normal weight, which is - according to WHO (2020a) - attained if the BMI is between 18.51 and 24.99.	underweight_imp
Being Overweighted	This is a dummy created on the basis of the BMI. If the BMI is between 25.00 and 29.99, then the individual is considered overweighted. The reference category is having a normal weight, which is - according to the WHO (2020a) - attained if the BMI is between 18.51 and 24.99.	overweight_imp
Being Obese or Severely Obese	This is a dummy created on the basis of the BMI. If the BMI is greater of equal than 30.00, then the individual is falls under the category of being obese, while a BMI over 34.99 is considered to be severely obese. The reference category is having a normal weight, which is - according to the WHO (2020a) - attained if the BMI is between 18.51 and 24.99.	obesity_imp

Number of Hours Worked per Week	This variable gives the number of hours per week that an individual works (Voorpostel et al., 2018).	imputed_hrsweek_2
Main Income Contributor	This dummy-variable reflects all individuals, which are earning 50% or more of the total's houshold-income.	main_inc_contrib_imput
Feeling Unhealthy: last Year	This variable indicates whether an individual felt unhealthy in the year before (2012). This is a lagged variable.	unhealthy12
Health-Expectations due to Sport	This variable are the different matching-estimators from the matching-regression. It reflects the expected short-term gain in health due to sport from an individual.	h_expectation

Table 9: Detailed List of Additional Variables used for further Models in Appendix

Full Name	**Description**	**Abbreviation**
Education of Father	Similarly to Schüttoff et al. (2018), this variable gives information whether the individual's father reached an educational-level that allows him to access tertiary education or not. This criterion is fulfilled if the father has at least a maturity-degree. The educational attainment is based on the ISCED-classification (Voorpostel et al., 2018).	f_educat_2 / f_educ_{iw} in equation (25), (26) and (27)
Leisure-Activity Index in 2010	For each individual, this variable aggregates over all the leisure-activities that an individual did on a weekly basis in the *year of 2010* besides sport, which is 3 years *before* the current year of analysis (2013). Those leisure activities include: meeting friends, going to the restaurant / bar, playing music, gardening, going out clubbing, attending to sport events, going to the cinema, reading books, going to the theater, attending to the opera, drawing / sculpturing, doing photography and going to the museum (Voorpostel et al., 2018).	act_index10 / $leisure\_index10_{iw}$ in equation (27)
Growth Rate of BMI	This variable is the calculated BMI-growth rate that an individual experienced from 2012 to 2013. Positive growth rates can be interpreted as a gain in body mass, whereas a negative growth rate can be seen as a loss in body mass.	BMI_growth
First difference in Health-Status	This variable is simply the (first) difference in Health-Status of 2013 relative to 2012.	D_hsat / $\Delta hsat_{iw}$ in equation (26)

C) Detailed Summary Statistics

Table 10: Detailed Summary Statistics, Part 1

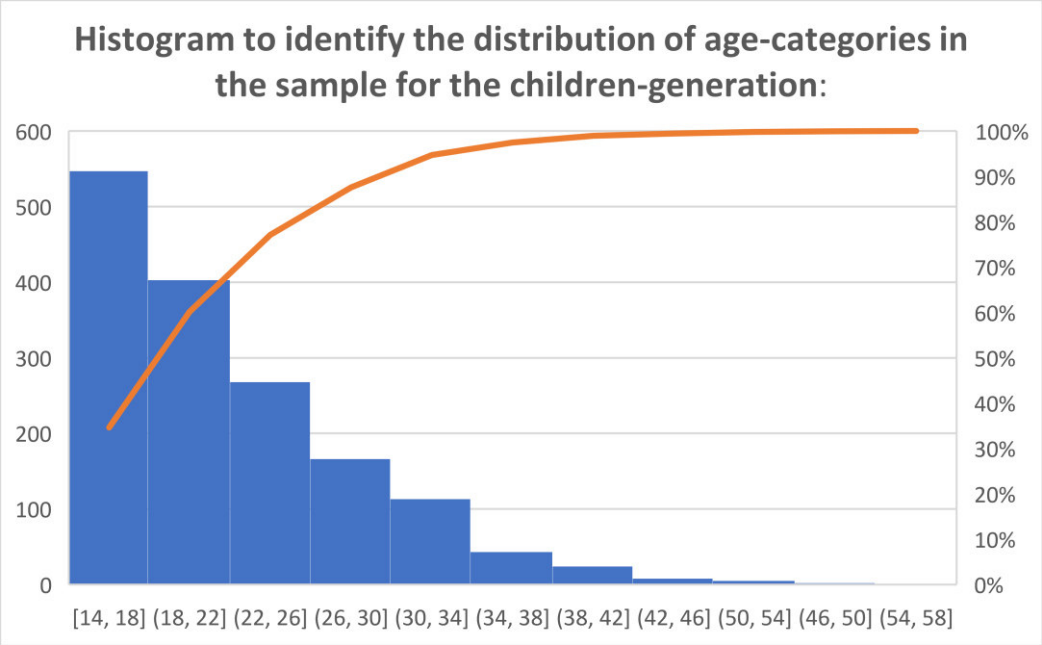
Sample Summary Statistics	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Age	1,579	22.267	6.651	14	17	26	58
Gender	1,579	0.495	0.500	0	0	1	1
Separated Parents	1,579	0.244	0.430	0	0	0	1
Weekly Sport	1,579	0.710	0.454	0	0	1	1
Mother Age	1,557	52.285	7.091	36.000	47.000	56.000	91.000
Mother Education	1,557	0.598	0.490	0	0	1	1
Personal Education	1,579	0.661	0.474	0	0	1	1
Working	1,579	0.697	0.460	0	0	1	1
Leisure Activity Index 2013	1,579	3.270	1.419	0	2	4	8
Cumulated Number of Hours watching TV, Internet & Gaming per Day	1,578	2.870	2.537	0.000	1.400	3.600	26.000
Doctor Consultations: Last 12 Months	1,573	2.753	5.377	0.000	0.000	3.000	65.000
Current Health-Satisfaction	1,579	8.211	1.516	0	8	9	10
Health Improvement 2013: Last 12 Months	1,579	0.362	1.185	-5	0	0	5
Health Improvement 2014: Last 12 Months	1,332	0.364	1.211	-5.000	0.000	1.000	5.000
Number of Siblings	1,579	0.901	1.137	0	0	1	8
Sport as a Child	1,579	0.843	0.364	0	1	1	1
Parents doing Sport	1,493	0.861	0.346	0.000	1.000	1.000	1.000
Parents Sport when Young	1,351	0.943	0.232	0.000	1.000	1.000	1.000
Siblings doing Sport	1,579	0.381	0.486	0	0	1	1
Unhealthy Parents: last Year	1,569	0.240	0.427	0.000	0.000	0.000	1.000
Household Income (Log)	1,553	11.798	0.568	8.132	11.513	12.169	13.373
Number of People within a Household	1,579	3.528	1.424	1	2	4	11
Being of Normal Weight	1,543	0.714	0.452	0.000	0.000	1.000	1.000
Being Underweighted	1,543	0.110	0.312	0.000	0.000	0.000	1.000
Being Overweighted	1,543	0.139	0.346	0.000	0.000	0.000	1.000
Being Obese or severely Obese Individual	1,543	0.038	0.190	0.000	0.000	0.000	1.000
Number of Hours Worked per Week	1,579	24.007	20.227	0	0	42	96
Main Income Contributor	1,504	0.220	0.414	0.000	0.000	0.000	1.000
Feeling Unhealthy: last Year	1,486	0.159	0.366	0.000	0.000	0.000	1.000
Health-Expectations due to Sport (%)	1,541	2.799	1.517	0.207	1.799	3.849	17.319

Table 11: Detailed Summary Statistics, Part 2

Categorical Variables: Personal Education	1) Primary Education	2) Secondary Education	3) Tertiary Education	Total
Frequency	682	588	309	1579
Proportion	43.19%	37.24%	19.57%	100.00%

Categorical Variables: Education of Mother	1) Primary Education	2) Secondary Education	3) Tertiary Education	Total
Frequency	102	1040	415	1557
Proportion	6.55%	66.80%	26.65%	100.00%
Categorical Variables: Education of Father	1) Primary Education	2) Secondary Education	3) Tertiary Education	Total
Frequency	51	611	763	1425
Proportion	3.58%	42.88%	53.54%	100.00%

Figure 6: Distribution of Individuals' Age-Categories within the Sample



D) Consequences of Assuming a Rational Preference Relation

Microeconomic theory assumes an individual to be *rational* when analyzing his / her behavior (Edwards, 1954). By specifying the choice correspondence *C(·)* according to the model of utility maximization in statement (4) from subsection 2.1, we automatically impose some constraints on the preference relation $Sport \succ NoSport$. Specifically, *only if* a preference relation $\succsim$ is rational[18] (and continuous[19]), can a preference relation be represented by a utility function *u(x)* (and vice versa). Formally, this is (Houthakker, 1950; Mas-Colell et al., 1995, Chapter 1):

$$Sport \overset{rational}{\succsim} No\ Sport \Leftrightarrow u(Sport) \geq u(No\ Sport) \qquad (22)$$

Additionally, another constraint is that, *not all* decision rules that can be used to specify *C(·)* have a utility representation[20]. To be precise, the fulfillment of the so called **W**eak **A**xiom of **R**evealed **P**references (WARP) is a sufficient and necessary condition for a choice correspondence *C(·)* to be represented by a utility function (Arrow, 1959; Mas-Colell et al., 1995; Samuelson, 1947, Chapter 5):

$$WARP\ is\ fulfilled \Leftrightarrow C(\cdot)\ \ has\ a\ utility\ representation\ u(\cdot) \qquad (23)$$

When using the economic concept of utility maximization to specify *C(·)*, we were able to *give C(·) a utility representation* and thus, by using (23), are able to fulfill the WARP-condition (Mas-Colell et al., 1995). This closes the circle that is pictured in Figure 7 further below, because - by definition of (23) - if WARP is satisfied, it will have a utility representation *u(·)* and - by (22) - we can conclude that the preference relation will have the properties of rationality (Mas-Colell et al., 1995):

$$WARP\ is\ fullfilled \Rightarrow preferences\ are\ rational \qquad (24)$$

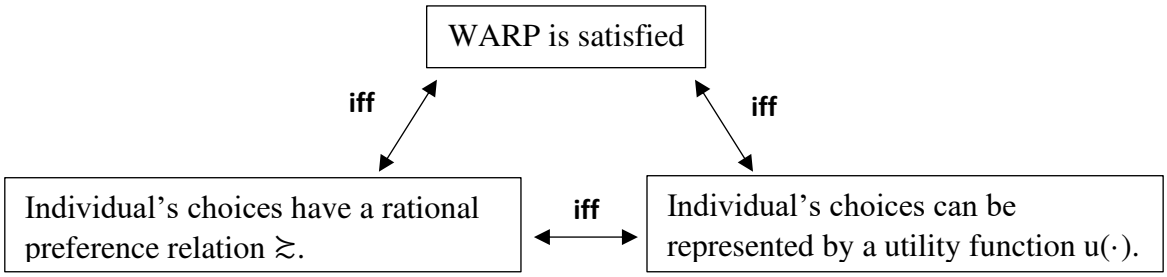
Therefore, *rational* preferences have the required properties to be represented by a utility function and, hence, the theoretical foundations on preferences are in line with utility maximization. This can be visualized by:

[18] Rational preferences are implied by the properties of a preference relation $\succsim$ being *complete* and *transitive*. By definition, a preference relation is *complete*, if all the items within a set can be compared. On the other hand, *transitive* preferences are - generically - characterized as: if $x \succsim y \cap y \succsim z \Rightarrow x \succsim z$ (Mas-Colell et al., 1995).

[19] Additionally, the assumption of a continuous preference relation $\succsim$ is necessary for equation (23) to hold (Mas-Colell et al., 1995).

[20] Typically, there is no utility representation when individuals have - for example - lexicographic preferences (Mas-Colell et al., 1995).

Figure 7: Relationship between Preferences, Utility Functions and Choices



E) Matching-Estimations: Logit-Regression

Table 12: Logistic Regressions for the Estimations of Propensity Scores

	Dependent variable: Weekly Sport			
	Treatment of doing Sport:			
	(1) Normal (adult)	(2) Normal (young)	(3) Over-weight	(4) Under-weight
Age	-0.008	-0.177*	-0.016	-0.081
	(0.037)	(0.099)	(0.047)	(0.130)
Mother Age	0.021	0.035	-0.013	0.060
	(0.030)	(0.036)	(0.039)	(0.104)
Gender	0.010	-0.224	-0.127	-1.900*
	(0.256)	(0.276)	(0.399)	(1.024)
Personal Education: A-level	-0.303	-0.171	-0.685	-0.389
	(0.276)	(0.499)	(0.435)	(1.053)
Mother Education: A-level	0.173	0.118	-0.517	-1.456
	(0.257)	(0.291)	(0.393)	(0.891)
Work	-0.495	-0.338	0.019	0.195
	(0.406)	(0.313)	(0.551)	(0.912)
Parents Separated	0.233	0.061	-0.059	1.060
	(0.322)	(0.338)	(0.422)	(1.232)
Leisure-Activity Index	0.192**	0.064	0.012	0.188
	(0.096)	(0.097)	(0.126)	(0.283)
Number of Hours on a Screen per Day	-0.036	-0.067*	-0.148	-0.106
	(0.061)	(0.041)	(0.101)	(0.189)
Number of Doctor Consultations: Last 12 Months	-0.082**	-0.047**	-0.086*	0.084
	(0.037)	(0.023)	(0.048)	(0.109)
Current Health-Satisfaction	0.142	-0.109	-0.089	-0.091
	(0.095)	(0.107)	(0.121)	(0.340)
Health-Improvement: Last 12 Months	0.283*	0.065	0.315*	0.434
	(0.156)	(0.112)	(0.161)	(0.459)
Number of Siblings	0.276	-0.072	0.321	3.119**
	(0.281)	(0.142)	(0.371)	(1.220)
Sport as a Child	0.464	2.334***	0.502	2.303
	(0.322)	(0.358)	(0.577)	(1.422)
Parents doing Sport	0.383	0.877**	0.746	1.287
	(0.369)	(0.422)	(0.495)	(1.171)
Siblings doing Sport	-0.145	0.518	-0.478	-2.775*

	(0.499)	(0.327)	(0.651)	(1.440)
Unhealhty Parents last Year	-0.043	0.521	-0.320	-0.679
	(0.293)	(0.344)	(0.391)	(1.007)
Hh-income (log)	0.123	0.020	-0.072	-0.450
	(0.217)	(0.273)	(0.359)	(0.820)
Constant	-3.493	1.046	4.774	4.564
	(2.988)	(4.110)	(4.741)	(10.770)
Region Fixed Effects	Yes	Yes	Yes	Yes
Observations	389	387	193	116
Log Likelihood	-209.029	-182.000	-104.351	-30.364
Akaike Inf. Crit.	468.058	414.000	258.701	110.728
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 13: Logistic Regressions for the Estimations of Propensity Scores, including Sport Participation of Parents when Young

	Binary Dependent variable: Weekly Sport			
	Treatment of doing Sport for:			
	(1)	(2)	(3)	(4)
	Normal (adult)	Normal (young)	Overweight	Underweight
Age	-0.008	-0.211**	0.003	-0.124
	(0.036)	(0.093)	(0.054)	(0.125)
Mother Age	0.010	0.037	0.035	0.009
	(0.029)	(0.035)	(0.044)	(0.091)
Gender	-0.087	-0.470*	-0.265	-1.050
	(0.246)	(0.275)	(0.410)	(0.837)
Personal Education	-0.328	0.005	-0.417	-1.260
	(0.276)	(0.473)	(0.488)	(1.107)
Mother Education	-0.049	0.262	-0.739*	-1.114
	(0.256)	(0.287)	(0.430)	(0.783)
Work	0.024	-0.381	-0.165	-0.906
	(0.362)	(0.308)	(0.560)	(0.800)
Parents Separated	0.011	0.144	0.444	0.480
	(0.322)	(0.362)	(0.495)	(0.935)
Leisure-Activity Index	0.246***	0.125	0.013	0.147
	(0.093)	(0.098)	(0.133)	(0.229)
Number of Hours on a Screen per Day	-0.048	-0.051	-0.053	0.035
	(0.056)	(0.043)	(0.102)	(0.117)

Number of Doctor Consultations: Last 12 Months	-0.063**	-0.041*	-0.075*	0.048
	(0.028)	(0.023)	(0.042)	(0.085)
Current Health-Satisfaction	0.009	-0.105	-0.019	-0.027
	(0.086)	(0.104)	(0.121)	(0.275)
Health-Improvement: Last 12 Months	0.229*	0.038	0.212	0.275
	(0.138)	(0.110)	(0.171)	(0.321)
Number of Siblings	-0.077	-0.061	0.360	1.114**
	(0.210)	(0.146)	(0.372)	(0.551)
Sport as a Child	0.500	2.685***	0.685	1.496
	(0.314)	(0.366)	(0.590)	(1.064)
Parents doing Sport	0.479	0.913**	0.058	0.561
	(0.371)	(0.416)	(0.512)	(1.129)
Parents Sport when Young	0.393	-0.258	0.735	-17.063
	(0.439)	(0.552)	(0.811)	(1,476.064)
Siblings doing Sport	0.215	0.516	0.223	-1.058
	(0.431)	(0.316)	(0.638)	(0.931)
Unhealhty Parents last Year	0.133	0.648*	-0.458	0.017
	(0.290)	(0.343)	(0.403)	(0.821)
Hh-income (log)	0.095	0.198	0.117	-1.008
	(0.221)	(0.298)	(0.377)	(0.653)
Constant	-2.488	-0.931	-1.502	32.464
	(3.057)	(4.332)	(5.215)	(1,476.093)
Region Fixed Effects	Yes	Yes	Yes	Yes
Observations	396	410	188	115
Log Likelihood	-218.572	-188.444	-97.051	-36.152
Akaike Inf. Crit.	489.144	428.888	246.103	124.304
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 13 has a slightly different specification relative to Table 12. More specifically, I just added one control variable in the form of the sport participation of the parent, when they were young. Two new insights can be drawn from these regressions:

- I explained in section 4 that parents seem to be from greater importance for younger individuals, since only for this group the sport-attitude of the previous generation was statistically significant, as we can see from Table 12. Now, with Table 13, another parental variable *Unhealthy Parents last Year* becomes statistically significant for

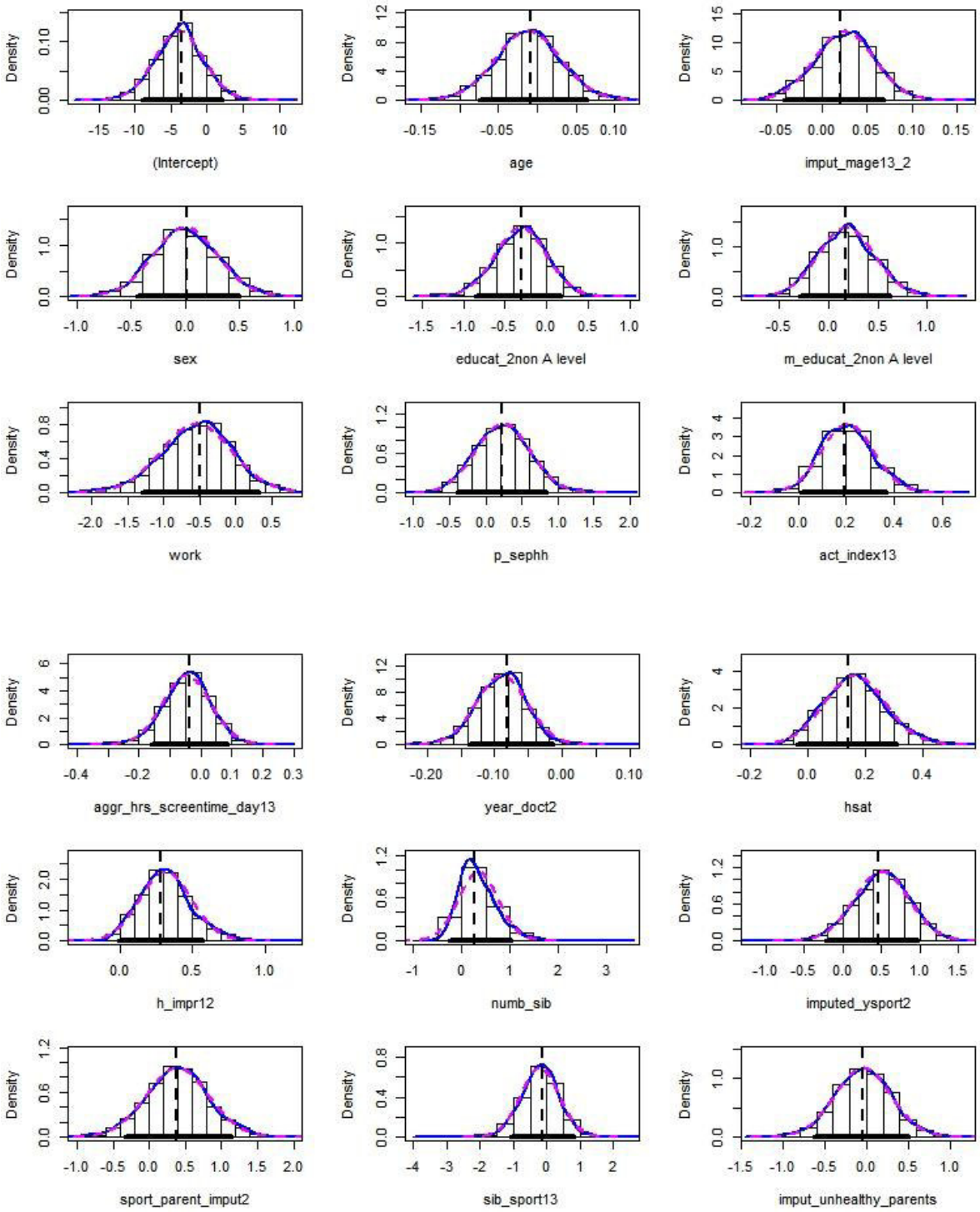
younger individuals. This result may highlight even further the greater importance of the parents for younger individuals.

- The different effect of household income on ones' sport participation becomes less clear, since only underweight individuals have a negative association with sport participation. However, there is still no statistical significance present.

Also, note that reason why the standard-error from the variable *Parent Sport when Young* for the underweighted individuals (model 4) is so high, is because there is only one observation in the control-group with a parent-pair that does no sport. This is also the reason why I did the matching-estimation in the main text *without* the parents' sports participation *when young*, since the amount of observation in the subsamples are already very low and R would have automatically tossed away too many individuals without full-cases.

F) Matching-Estimations: Robustness Check by Bootstrapping

Figure 8: Bootstrapping to Approximate the Coefficient's Sample-Distribution with 1,000 Draws for Normally Weighted Adults using Equation (21)



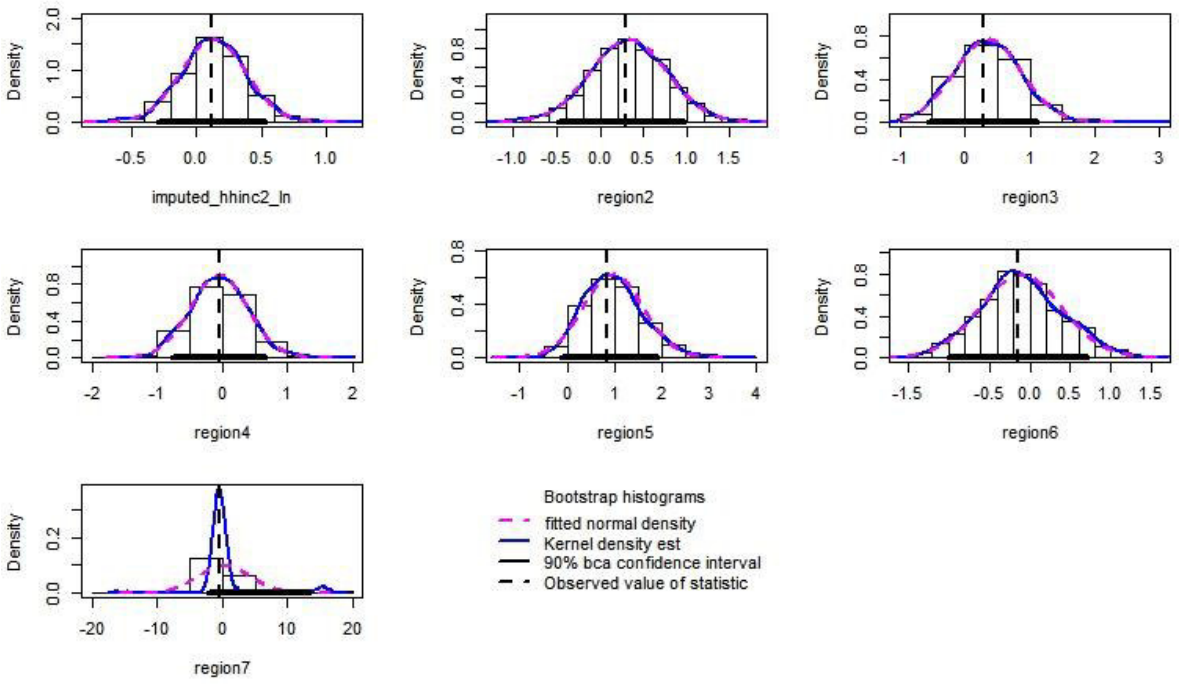
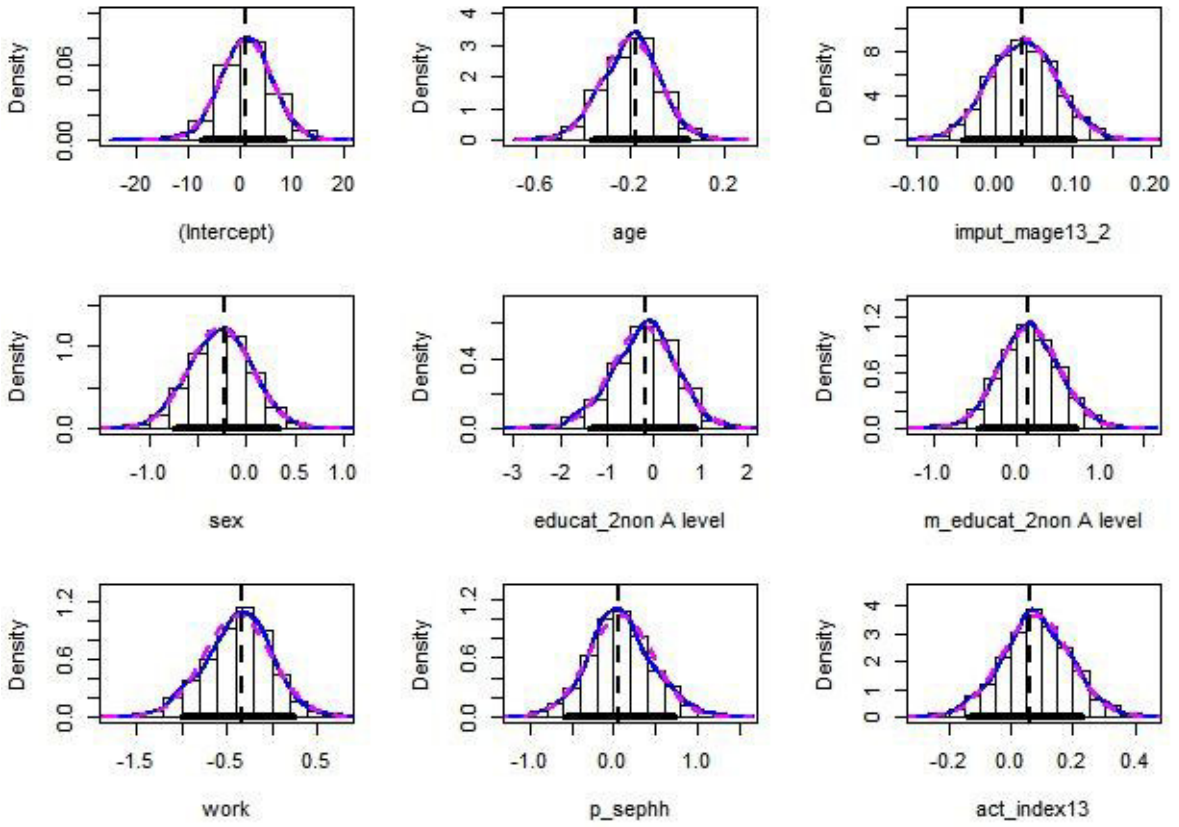


Figure 9: Bootstrapping to Approximate the Coefficient's Sample-Distribution with 1,000 Draws for Normally Weighted Young Individuals using Equation (21)



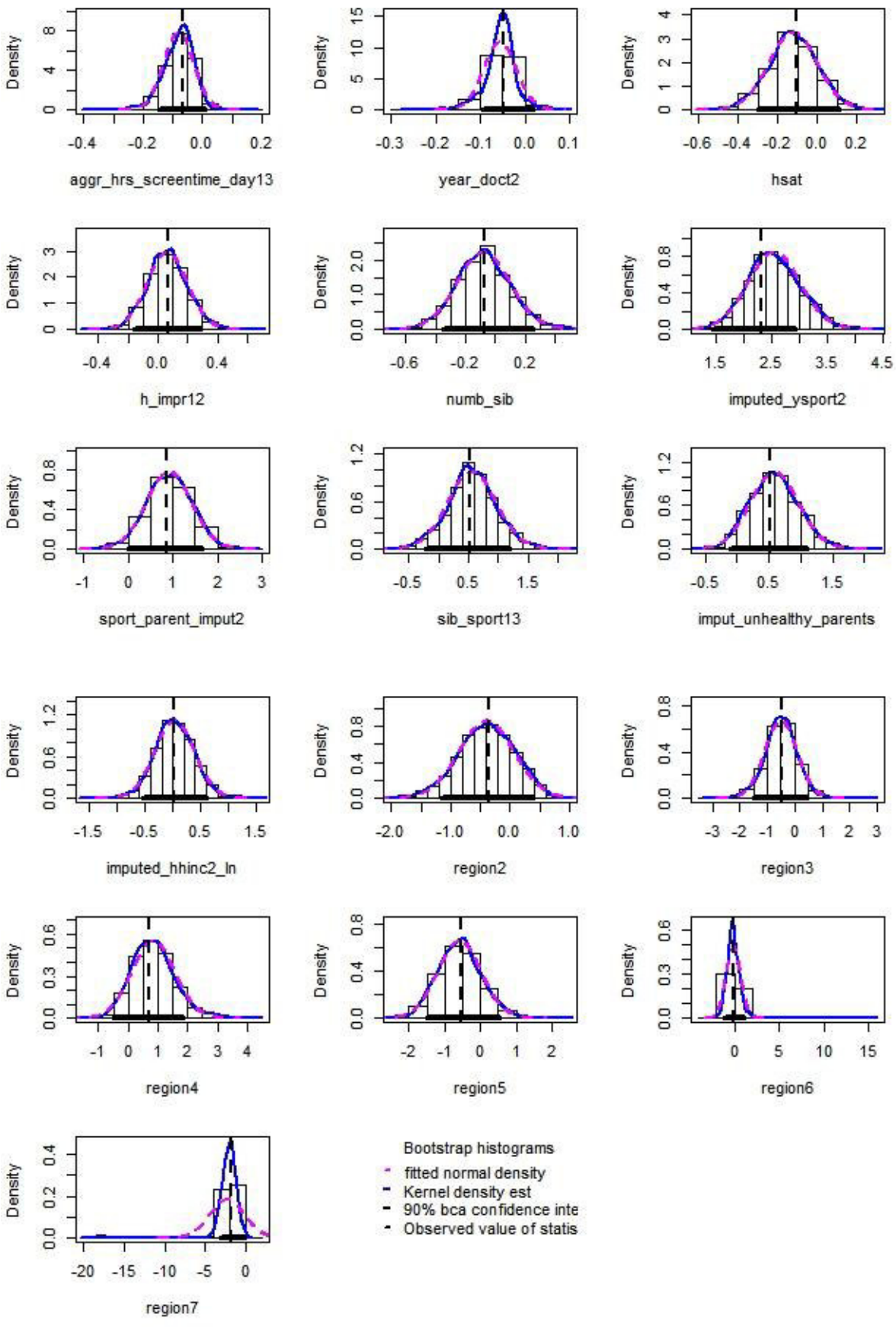
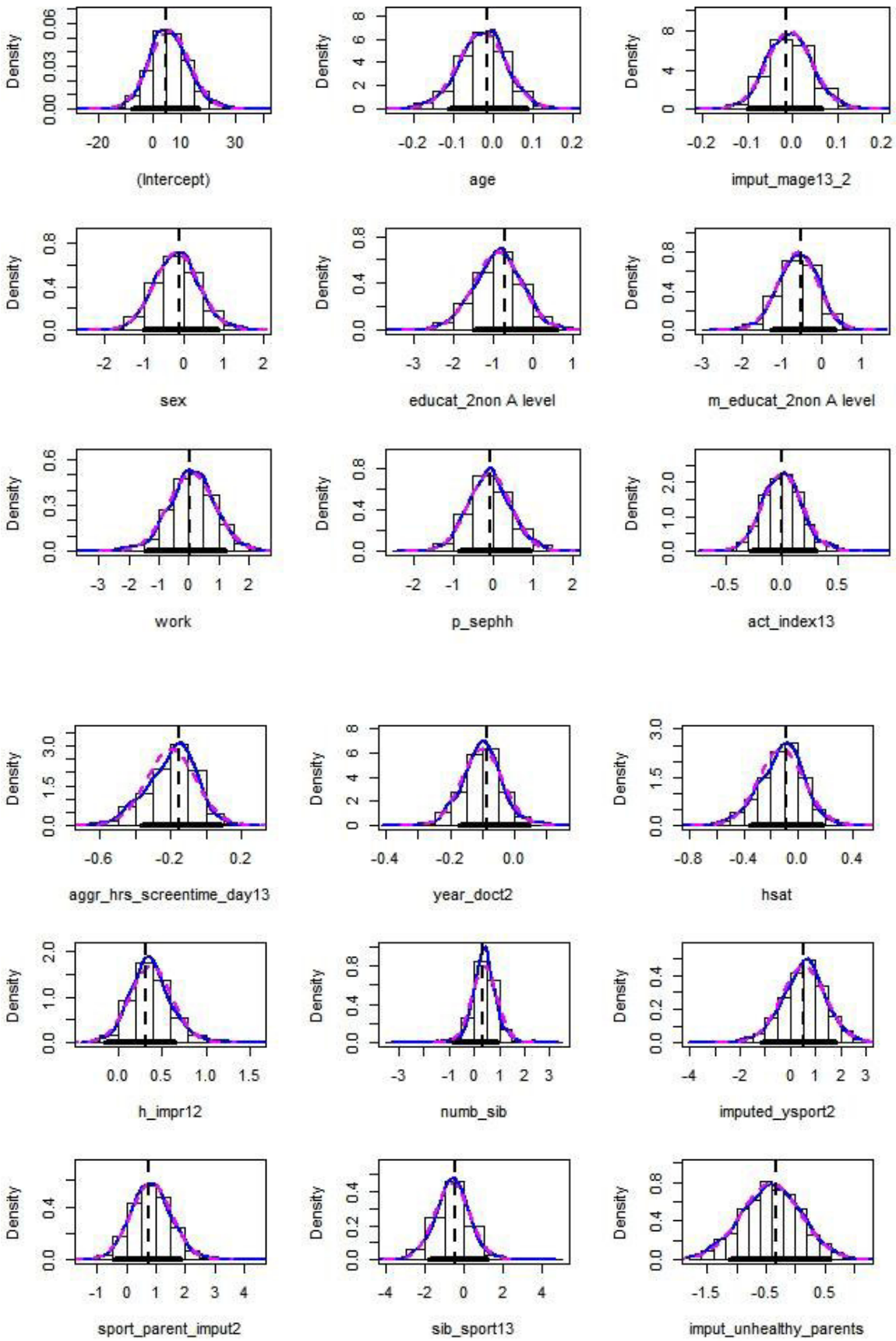

Density
aggr_hrs_screentime_day13
year_doct2
hsat
h_impr12
numb_sib
imputed_ysport2
sport_parent_imput2
sib_sport13
imput_unhealthy_parents
imputed_hhinc2_ln
region2
region3
region4
region5
region6
region7
Bootstrap histograms
fitted normal density
Kernel density est
90% bca confidence inte
Observed value of statis

Figure 10: Bootstrapping to Approximate the Coefficient's Sample-Distribution with 1,000 Draws for Overweight Individuals using Equation (21)



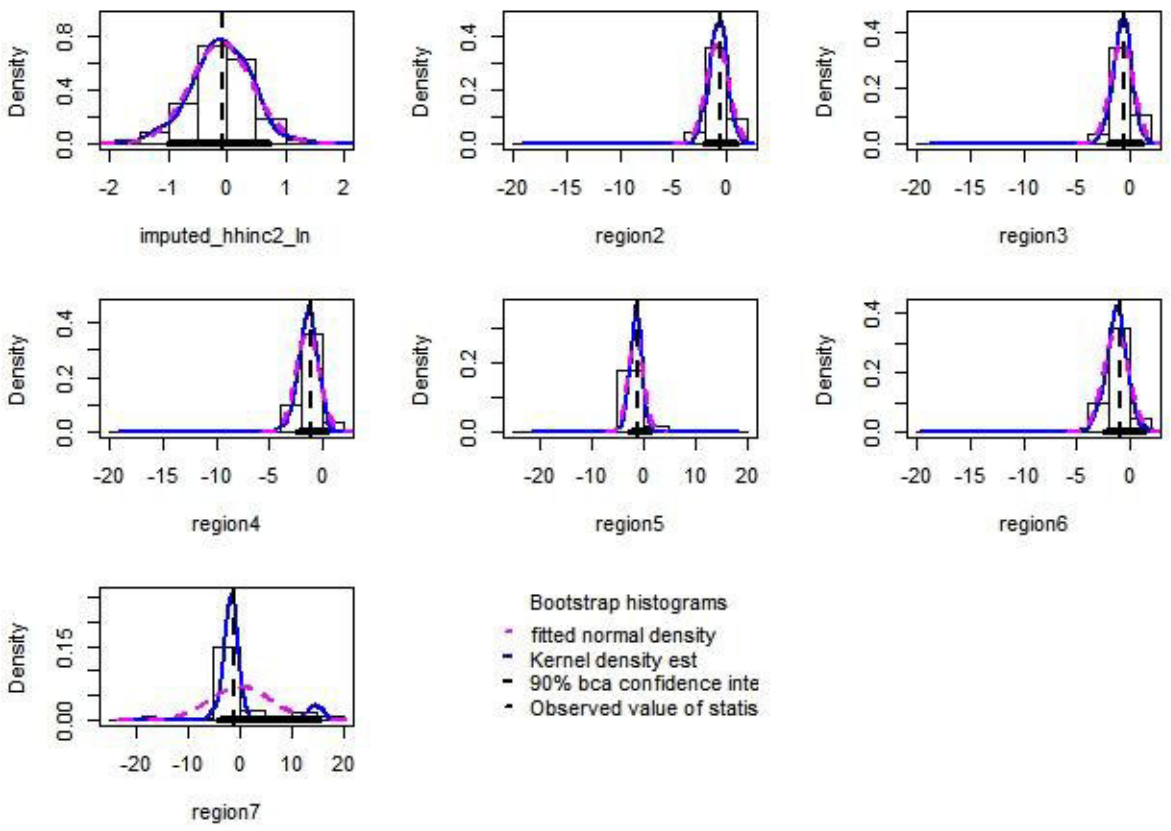
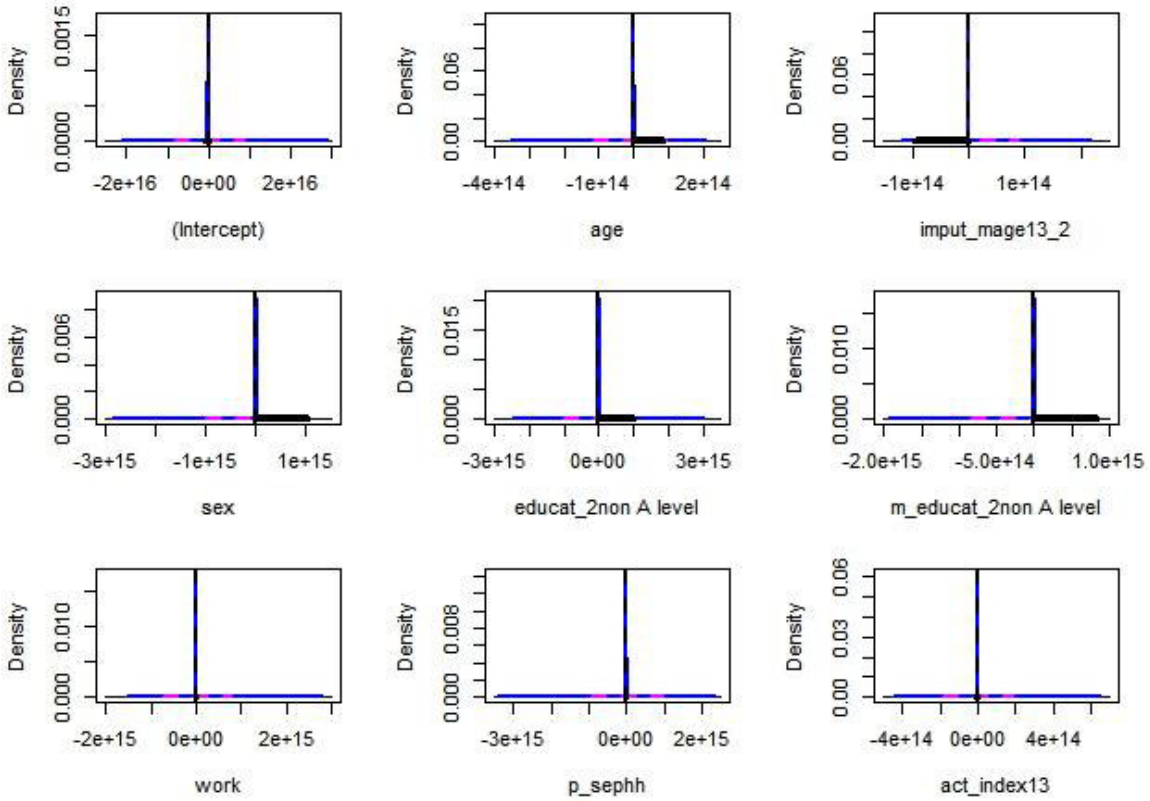
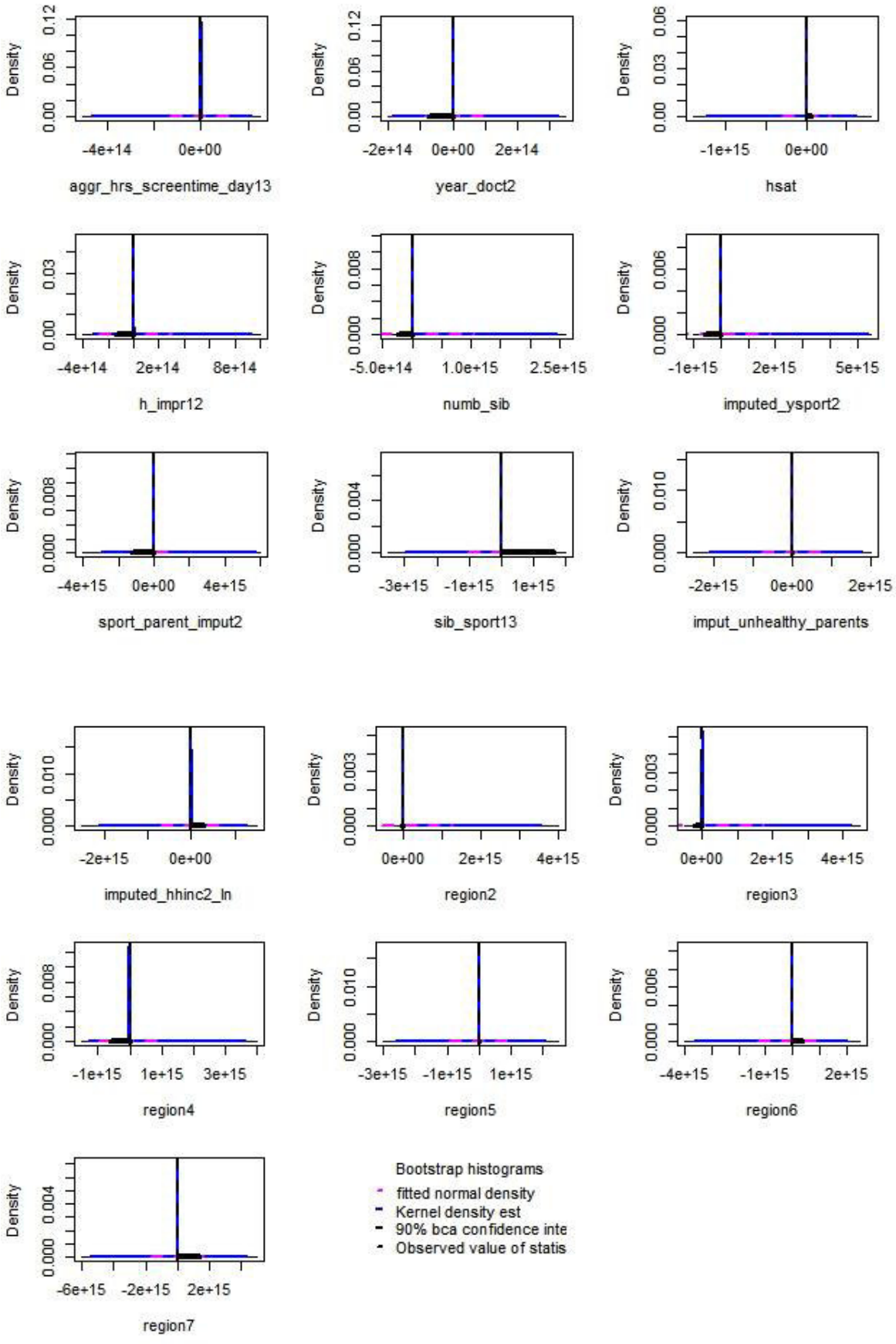


Figure 11: Bootstrapping to Approximate the Coefficient's Sample-Distribution with 1,000 Draws for Underweight Individuals using Equation (21)




Density
aggr_hrs_screentime_day13
year_doct2
hsat
h_impr12
numb_sib
imputed_ysport2
sport_parent_imput2
sib_sport13
imput_unhealthy_parents
imputed_hhinc2_ln
region2
region3
region4
region5
region6
region7
Bootstrap histograms
fitted normal density
Kernel density est
90% bca confidence inte
Observed value of statis

G) Matching-Estimations: Alternative Specifications and Estimators of ATT and ATE

As described in section 4, I tried different functional forms in order to get better matching-estimates. I used the following alternative logit-regressions for the propensity scores:

1. Including squared- and interaction-terms:

$$sport_{iw} = \beta_0 + \beta_1 ysport_{iw} + \beta_2 p\_sport_{iw} + \beta_3 p\_ysport_{iw} + \beta_4 sib\_sport_{iw} + \beta_5 age_{iw} + \beta_6 age_{iw}^2 + \beta_7 sex_{iw} + \beta_8 educ_{iw} + \beta_9 m\_educ_{iw} + \beta_{10} f\_educ_{iw} + \beta_{11} leisure\_index10_{iw} + \beta_{12} doct\_cons_{iw} + \beta_{13} hsat_{iw} + \beta_{14}(hsat_{iw} * age_{iw}) + \beta_{15} h\_impr\_lagged_{iw} + \beta_{16} BMI\_growth_{iw} + \beta_{17} p\_unhealthy_{iw} + \beta_{18} hrs\_worked_{iw} + \beta_{19} hh\_inc_{iw} + \beta_{20} work_{iw} + \beta_{21} region_{iw} \quad (25)$$

2. Including squared-terms and first-differences:

$$sport_{iw} = \beta_0 + \beta_1 ysport_{iw} + \beta_2 p\_sport_{iw} + \beta_3 p\_ysport_{iw} + \beta_4 sib\_sport_{iw} + \beta_5 age_{iw} + \beta_6 age_{iw}^2 + \beta_7 sex_{iw} + \beta_8 educ_{iw} + \beta_9 m\_educ_{iw} + \beta_{10} f\_educ_{iw} + \beta_{11} leisure\_index_{iw} + \beta_{12} doct\_cons_{iw} + \beta_{13} \Delta hsat_{iw} + \beta_{15} h\_impr\_lagged_{iw} + \beta_{16} BMI\_growth_{iw} + \beta_{17} p\_unhealthy_{iw} + \beta_{18} hrs\_worked_{iw} + \beta_{19} hh\_inc_{iw} + \beta_{20} work_{iw} + \beta_{21} region_{iw} \quad (26)$$

3. Including squared- and interaction-term, as well as one additional covariate of an alternative leisure activity:

$$sport_{iw} = \beta_0 + \beta_1 ysport_{iw} + \beta_2 p\_sport_{iw} + \beta_3 p\_ysport_{iw} + \beta_4 sib\_sport_{iw} + \beta_5 age_{iw} + \beta_6 age_{iw}^2 + \beta_7 sex_{iw} + \beta_8 educ_{iw} + \beta_9 m\_educ_{iw} + \beta_{10} f\_educ_{iw} + \beta_{11} leisure\_index10_{iw} + \beta_{12} doct\_cons_{iw} + \beta_{13} hsat_{iw} + \beta_{14}(hsat_{iw} * age_{iw}) + \beta_{15} h\_impr\_lagged_{iw} + \beta_{16} BMI\_growth_{iw} + \beta_{17} p\_unhealthy_{iw} + \beta_{18} hrs\_worked_{iw} + \beta_{19} hh\_inc_{iw} + \beta_{20} work_{iw} + \beta_{21} region_{iw} \quad (27)$$

Where $sport_{iw}$ denotes if individual i with the, in Table 2, defined weight-categories $w \in \{underweight, normal\ weight\ \&\ young, normal\ weight\ \&\ adult, overweight\ or\ obsese\}$ practices sport. The exact specification of all abbreviated independent variables in the equations (25), (26) and (27) can be found in the Tables 7, 8 or 9 in Appendix B. Note that I used different functional forms depending on the weight-category *w* an individual belonged. This gives the advantage of adapting the covariates on *w* instead of using the same model for all *w*'s, thus giving more modelling flexibility.

Like in section 4, I use the regression $Health - Improvement\ in\ 2014_i = \beta_0 + \beta_1 * Weekly\ Sport\ in\ 2013_i$ to get the results in Table 14 below. Using again the simple one to one matching, I get the following matching-estimates:

Table 14: Alternative Matching-Estimators

Alternative Matching-Estimators	estimated ATT	p-value ATT	estimated ATE	p-value ATE	Equation of Reference
Normal Adult	0.17241	0.41921	0.052381	0.79494	eq. (26)
Normal Young	0.70769	0.048313	0.59091	0.045547	eq. (25)
Overweighted	0.16	0.62752	0.11504	0.68331	eq. (27)
Underweighted	0.23333	0.75754	0.068182	0.91601	eq. (27)

H) Matching-Estimations: Region of Common Support

Figure 12: Region of Common Support for Normally Weighted Adults using Equation (21) and MatchIt-Package

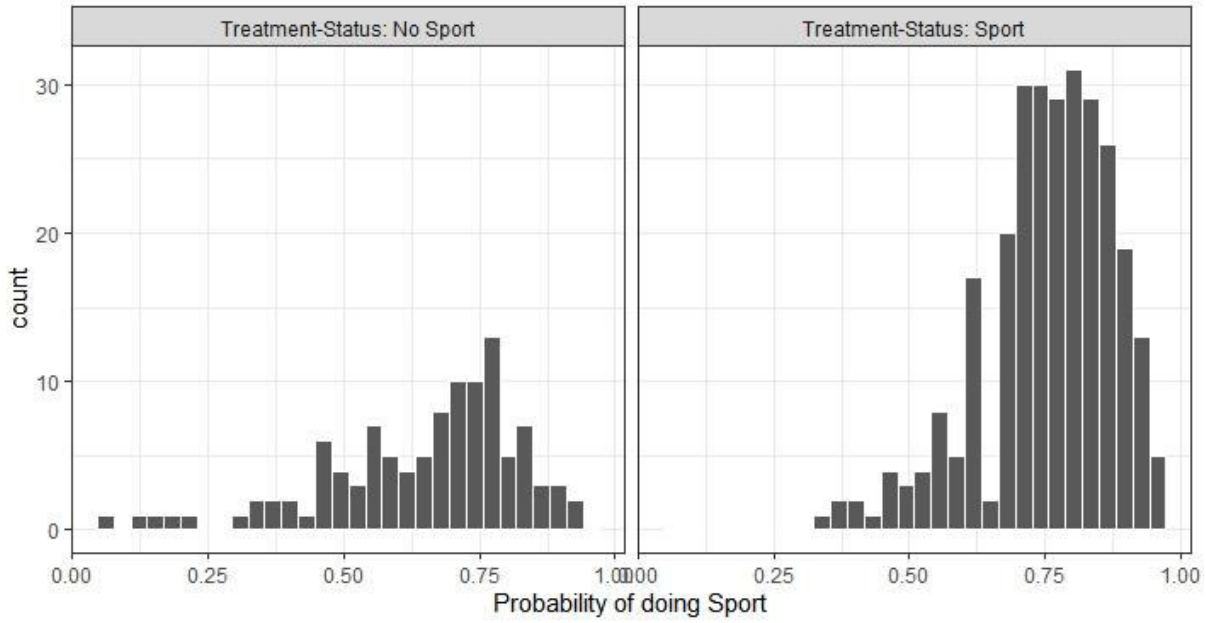


Figure 13: Region of Common Support for Overweight Individuals using Equation (21) and MatchIt-Package

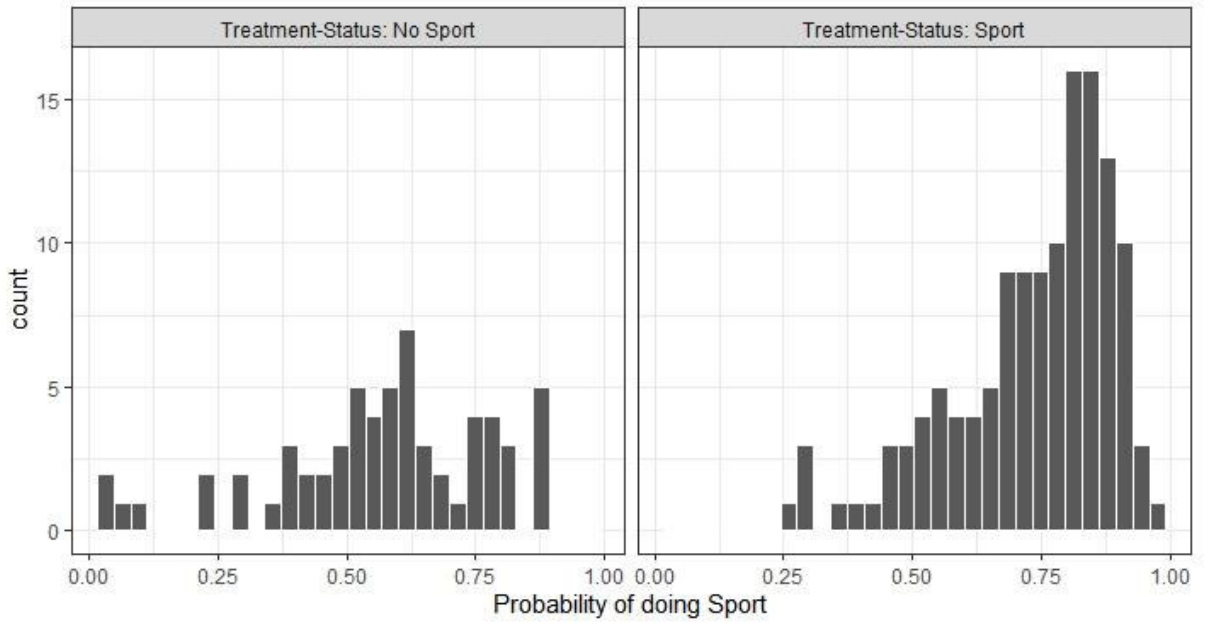


Figure 14: Region of Common Support for Underweight Individuals using equation (21) and MatchIt-Package

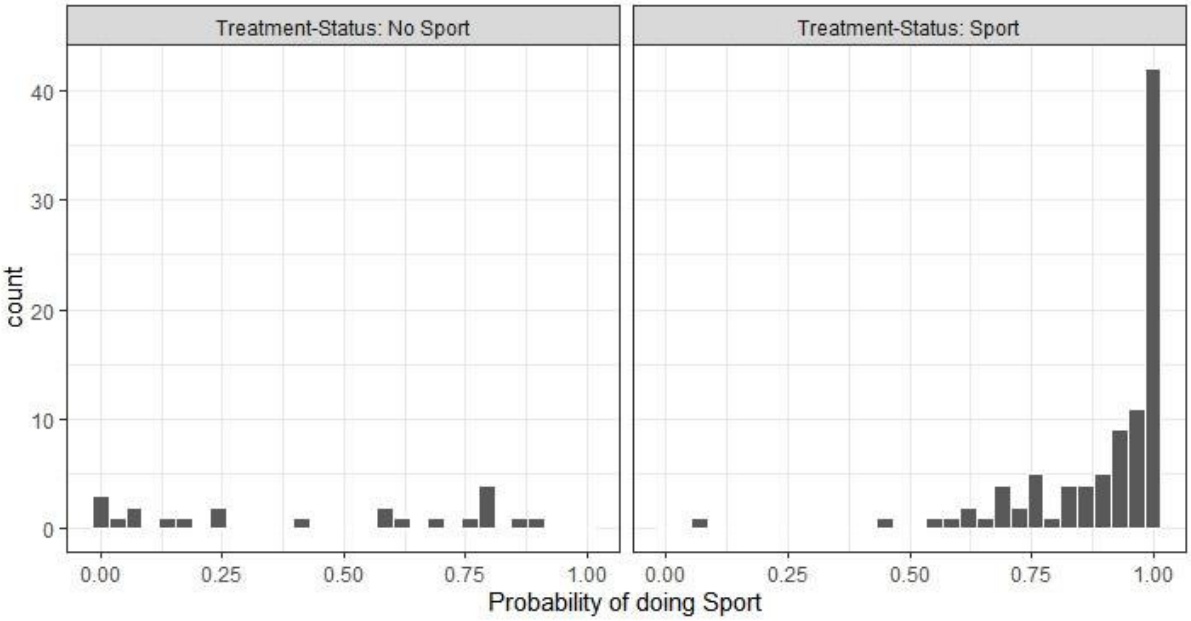
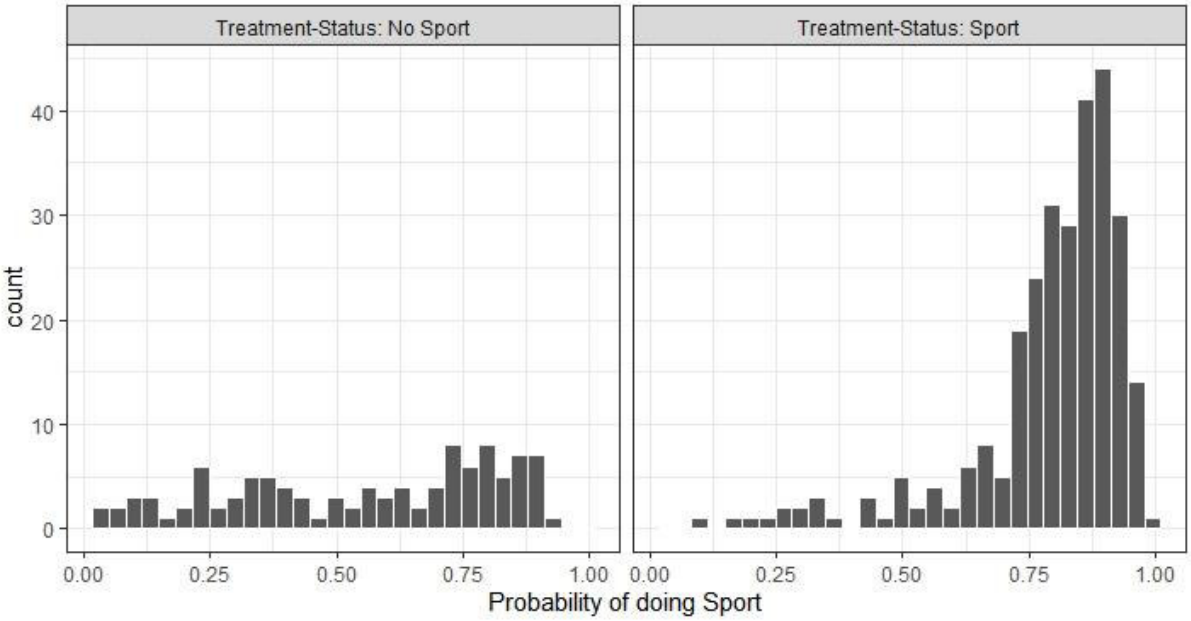


Figure 15: Region of Common Support for Normally Weighted Young Individuals using equation (21) and MatchIt-Package



I) Matching Estimations: Visualization of Covariate Balance

Figure 16: Covariate Balance after Matching for Normally Weighted Adults using Equation (21) and MatchIt-Package

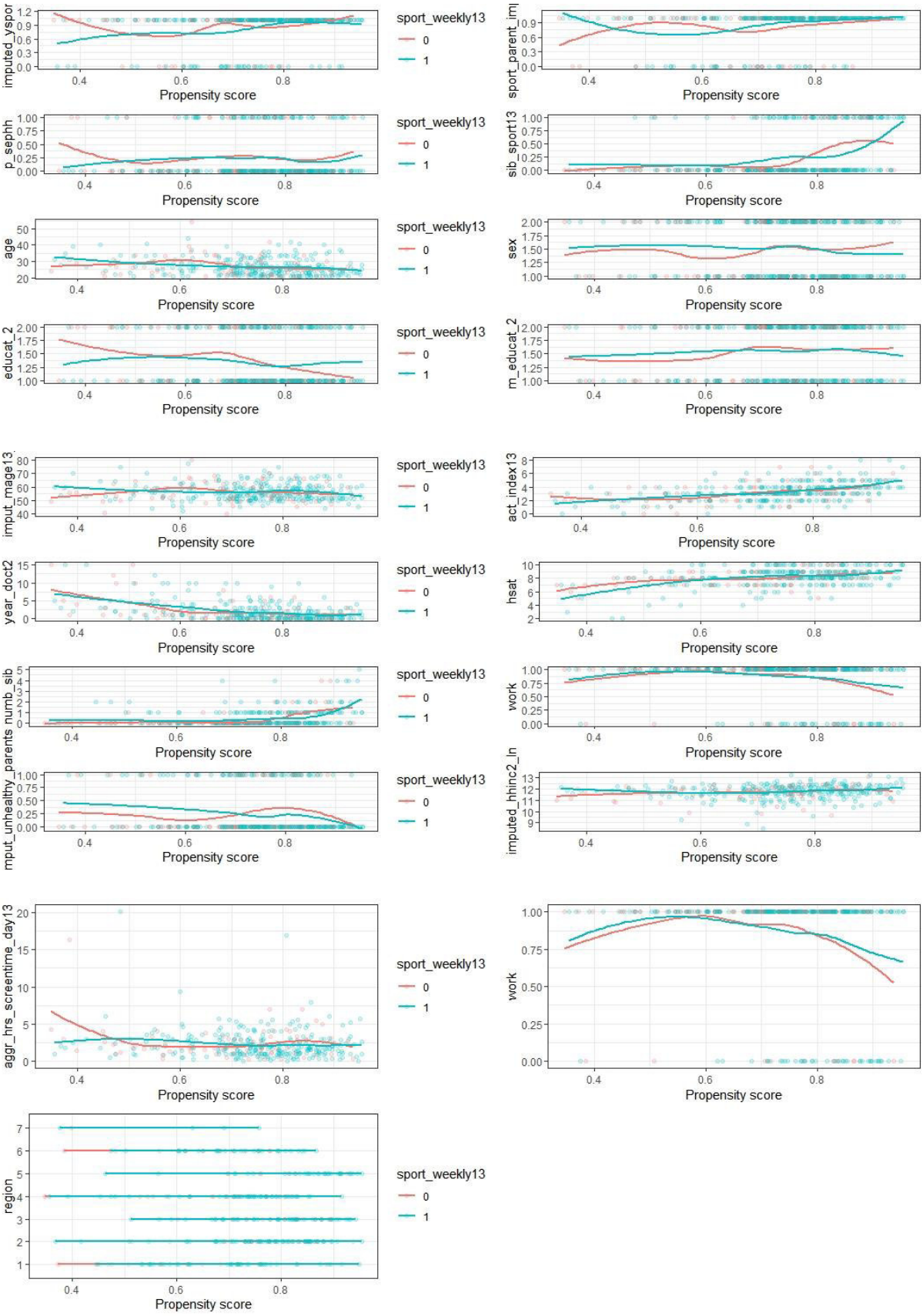


Figure 17: Covariate Balance after Matching for Overweight Individuals using Equation (21) and MatchIt-Package

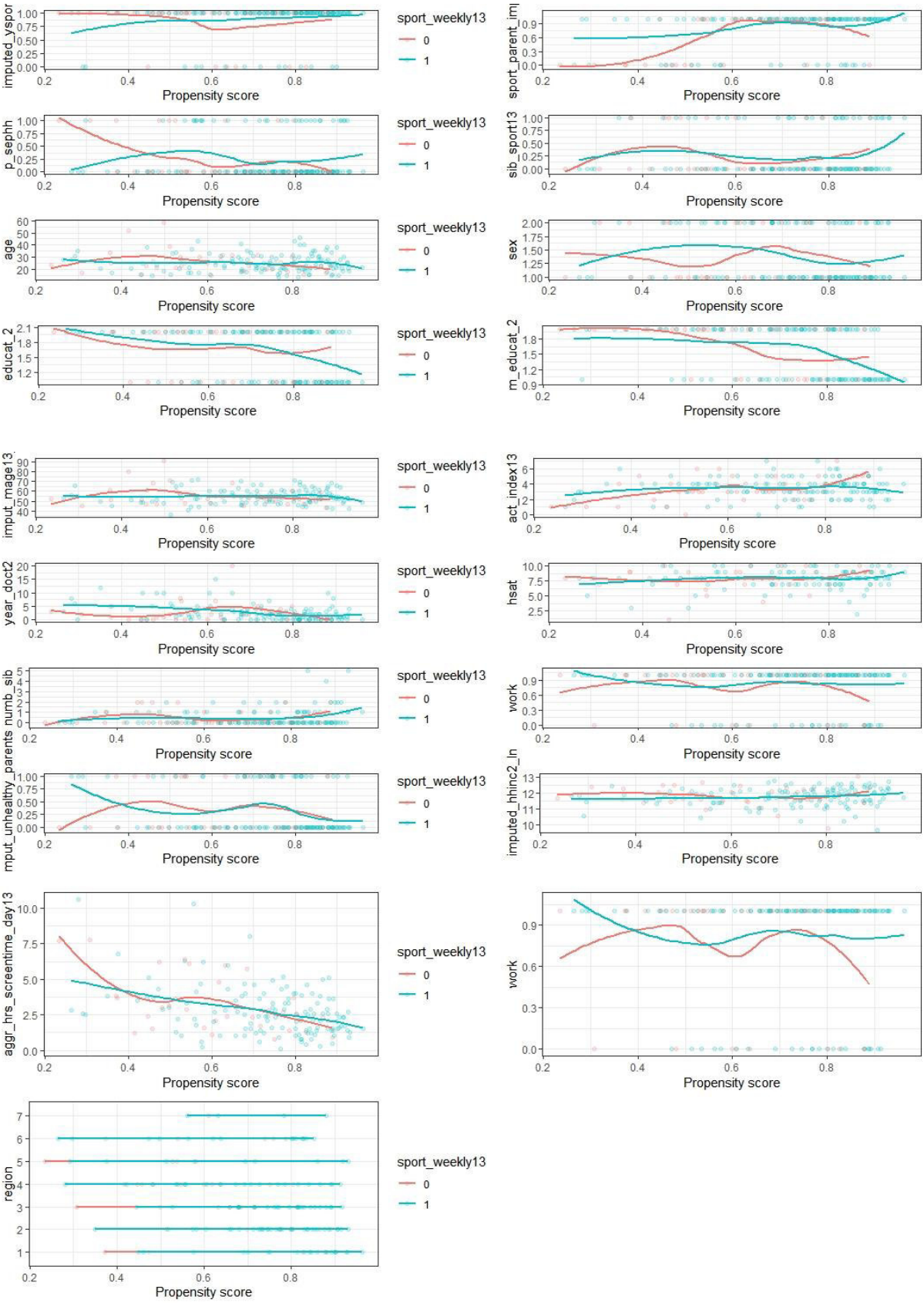


Figure 18: Covariate Balance after Matching for Underweight Individuals using Equation (21) and MatchIt-Package

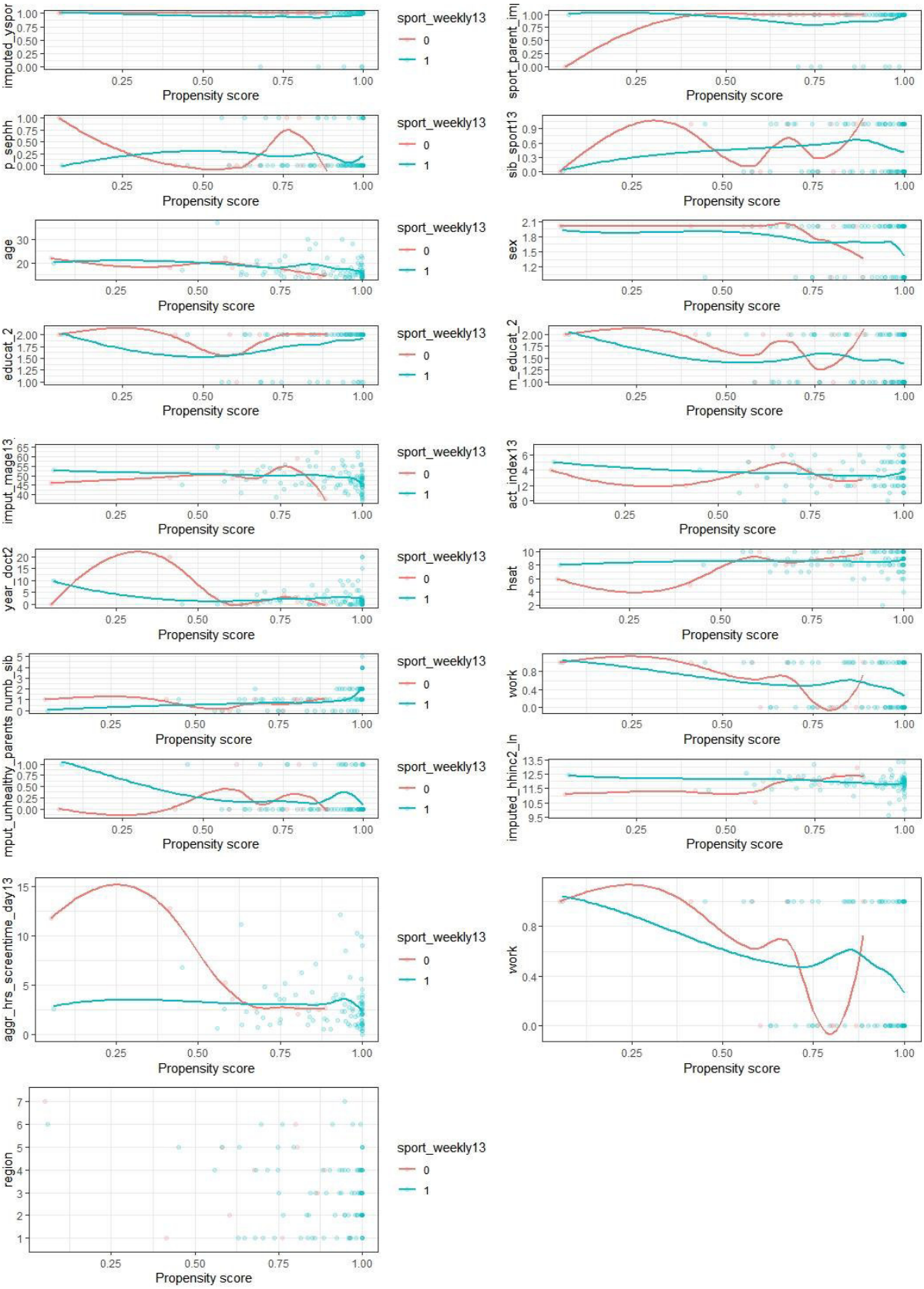


Figure 19: Covariate Balance after Matching for Normally Weighted Young Individuals using Equation (21) and MatchIt-Package

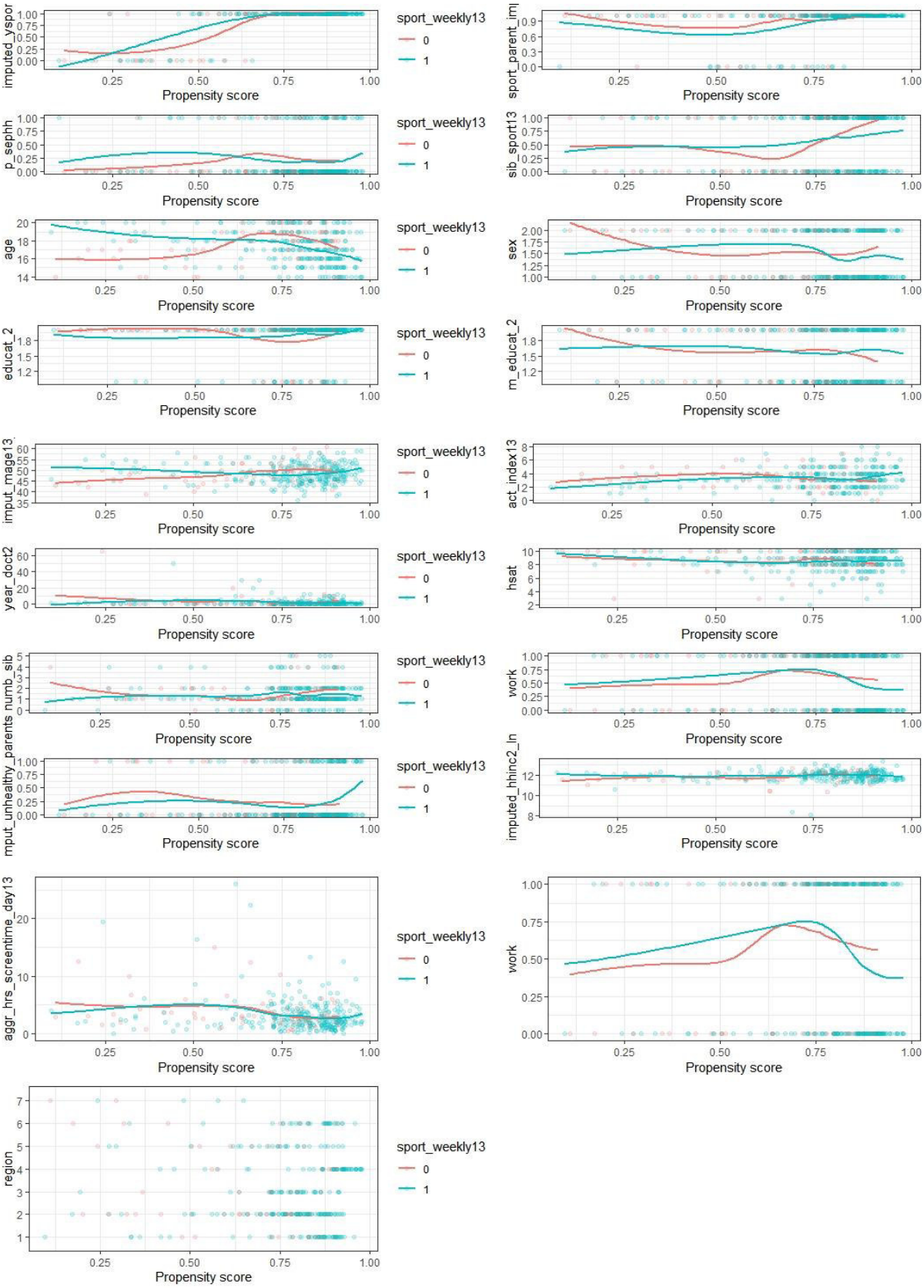


Table 15: Standardized Differences after Matching based on Equation (21) and Sub-Samples based on Table 2

Absolute Standardized Difference	Normal Adult	Overweight	Underweight	Normal Young
Age	0.089324428	0.123235114	0.169957856	0.24281812
Mother Age	0.046786316	0.07767916	0.039035262	0.05292891
Gender	0.027627633	0.010920923	0.445086087	0.112865293
Personal Education	0.097856409	0.14676509	0.17359264	0.069496102
Education of Mother	0.036280616	0.270283552	0.51513117	0.005083085
Work	0.051277322	0.096259753	0.187554423	0.080838647
Parents Separated	0.046986855	0.013153856	0.272248793	0.010012889
Leisure Activity Index	0.240910221	0.064915995	0.172687228	0.021245948
Number of hours on a Screen per Day	0.095172191	0.292562559	0.618631416	0.181682123
Number of Doctor Consultations: last 12 Months	0.110555192	0.053777512	0.051699813	0.138947959
Current Health-Satisfaction	0.169504532	0.048387196	0.149828986	0.043843974
Health-Improvement 2013: Last 12 Months	0.183385981	0.153184962	0.068979775	0.038351083
Number of Siblings	0.159607245	0.028432997	0.775629918	0.01027335
Sport as a Child	0.003307725	0.172963595	0.333413011	0.456976561
Parents doing Sport	0.175656752	0.465017221	0.049825569	0.062718183
Siblings doing Sport	0.140024685	0.004362365	0.041330986	0.122155664
Unhealthy Parents last Year	0.068954915	0.150624272	0.020817689	0.051935886
Hh-Income	0.09627015	0.071265533	0.059815476	0.154545494
Region 1 Dummy	0.058613777	0.142838119	0.030644777	0.074106213

Region 2 Dummy	0.067805018	0.243713989	0.382064744	0.091731446
Region 3 Dummy	0.034734145	0.284081448	0.228801768	0.010285322
Region 4 Dummy	0.072863636	0.100544961	0.005154393	0.18814831
Region 5 Dummy	0.140000199	0.147175792	0.320837839	0.045879137
Region 6 Dummy	0.102853175	0.096716102	0.379938836	0.081582656
Region 7 Dummy	0.009228207	0.016242208	0.127722227	0.103439223
Average Absolute Standardized Difference	0.093023493	0.131004171	0.224817227	0.098075663

Table 16: Number of Covariates with Remaining Imbalance after Matching based on Equation (21) and Sub-Samples based on Table 2

Covariate Count with Non-Negligible Imbalance *after* Maching: above 0.1	Normal Adult	Overweight	Underweight	Normal Young
Yes	9	13	16	9
No	16	12	9	16
Total	25	25	25	25

K) Multinomial Logit Model: Different Specifications

Table 17: Multinomial Logit Model with Subsample of Individuals aged 20 to 30

	Dependent variable: lifecycle choice-sequence for weekly Sport					
	Pooled MNL			MNL with only indiv. aged 20 to 30		
	Sport only as an Adult	Sport only as a Child	Never Sport	Sport only as an Adult	Sport only as a Child	Never Sport
	(1)	(2)	(3)	(4)	(5)	(6)
Parents doing Sport	1.000	-0.318	-0.495	0.096	-0.082	-0.461
	(0.626)	(0.291)	(0.349)	(0.674)	(0.457)	(0.726)
Parents Sport when Young	0.641	0.131	-0.373	-0.022	-0.427	-1.141
	(0.567)	(0.399)	(0.462)	(0.823)	(0.645)	(0.873)
Mother: Secondary Educ	-0.617	-0.207	-0.448	-0.358	-0.606	13.977***
	(0.599)	(0.426)	(0.477)	(0.770)	(0.629)	(1.315)
Mother: Tertiary Educ	-0.925	-0.288	-0.181	-0.714	-0.682	14.061***
	(0.649)	(0.463)	(0.519)	(0.827)	(0.684)	(1.316)
Father: Secondary Educ	1.043	-0.398	0.147	14.553***	-0.785	0.470
	(1.075)	(0.494)	(0.654)	(0.900)	(0.836)	(1.436)
Father: Tertiary Educ	0.953	-0.547	-0.594	14.259***	-0.957	-1.623
	(1.076)	(0.502)	(0.677)	(0.935)	(0.846)	(1.500)
2 People living in Hh	-0.609	-0.068	0.021	-0.100	0.196	-1.001
	(0.549)	(0.436)	(0.838)	(0.792)	(0.561)	(1.141)
3 People living in Hh	-0.322	0.533	0.973	0.499	0.511	0.046
	(0.592)	(0.483)	(0.852)	(0.890)	(0.651)	(1.184)
4 People living in Hh	-0.407	0.318	0.496	0.625	0.089	-0.224
	(0.581)	(0.484)	(0.862)	(0.912)	(0.678)	(1.234)
5 People living in Hh	-0.111	0.574	0.609	1.608*	0.910	1.448
	(0.618)	(0.519)	(0.886)	(0.948)	(0.752)	(1.275)
6 or more People living in Hh	-0.657	0.794	0.650	0.946	1.136	0.060
	(0.820)	(0.594)	(0.959)	(1.142)	(0.897)	(1.708)
Age	0.079***	-0.053*	-0.072*	0.075	-0.149**	0.028
	(0.030)	(0.028)	(0.043)	(0.081)	(0.070)	(0.113)
Gender	0.554**	0.471**	0.561**	0.447	0.110	0.813
	(0.271)	(0.197)	(0.254)	(0.376)	(0.316)	(0.514)
Number of Doctor Consultations: Last 12 Months	-0.047	0.089***	0.003	-0.070	0.160***	-0.201
	(0.044)	(0.018)	(0.037)	(0.071)	(0.043)	(0.149)
Being Underweighted	-0.209	-0.563	-0.747*	0.147	3.096***	2.462*
	(0.475)	(0.369)	(0.435)	(0.892)	(0.820)	(1.408)

Being Overweighted	0.136	-1.414***	-1.684***	0.273	-0.825*	-1.437
	(0.363)	(0.326)	(0.496)	(0.475)	(0.483)	(0.895)
Being Obese or Severly Obese	-1.205	-1.587***	-1.144	-14.115***	-0.550	-16.034***
	(1.050)	(0.605)	(0.750)	(0.00000)	(0.854)	(0.00000)
Number of hours worked per week	0.0002	0.016***	0.012	0.00001	-0.002	0.024
	(0.007)	(0.006)	(0.007)	(0.010)	(0.008)	(0.015)
Main Income-Contributor in the Hh	-0.293	0.073	-0.318	-0.019	0.382	-1.095
	(0.464)	(0.353)	(0.601)	(0.680)	(0.494)	(1.015)
Personal Education: Secondary Educ	0.245	-0.760***	-1.465***	-0.615	-0.639	-0.330
	(0.367)	(0.282)	(0.383)	(0.552)	(0.468)	(0.795)
Personal Education: Tertiary Educ	0.045	-1.357***	-1.484***	-0.334	-0.707	-0.151
	(0.517)	(0.392)	(0.556)	(0.676)	(0.558)	(0.905)
Number of Hours on a Screen per Day	-0.096	0.045	0.099**	-0.109	0.032	0.088
	(0.077)	(0.040)	(0.040)	(0.102)	(0.069)	(0.114)
Feeling Unheathy: last Year	0.115	0.160	0.610*	-0.268	0.933*	1.179
	(0.399)	(0.298)	(0.331)	(0.689)	(0.488)	(0.807)
Health-Expectation due to Sport	-0.082	-1.513***	-1.183***	0.013	-2.286***	-1.470***
	(0.122)	(0.127)	(0.161)	(0.145)	(0.240)	(0.364)
Constant	-5.871***	3.488***	2.953*	-18.073***	8.165***	-13.806***
	(1.766)	(1.151)	(1.587)	(1.799)	(2.350)	(2.566)
Akaike Inf. Crit.	1,834.112	1,834.112	1,834.112	850.493	850.493	850.493
Observations		1068			482	

Note: *p<0.1; **p<0.05; ***p<0.01

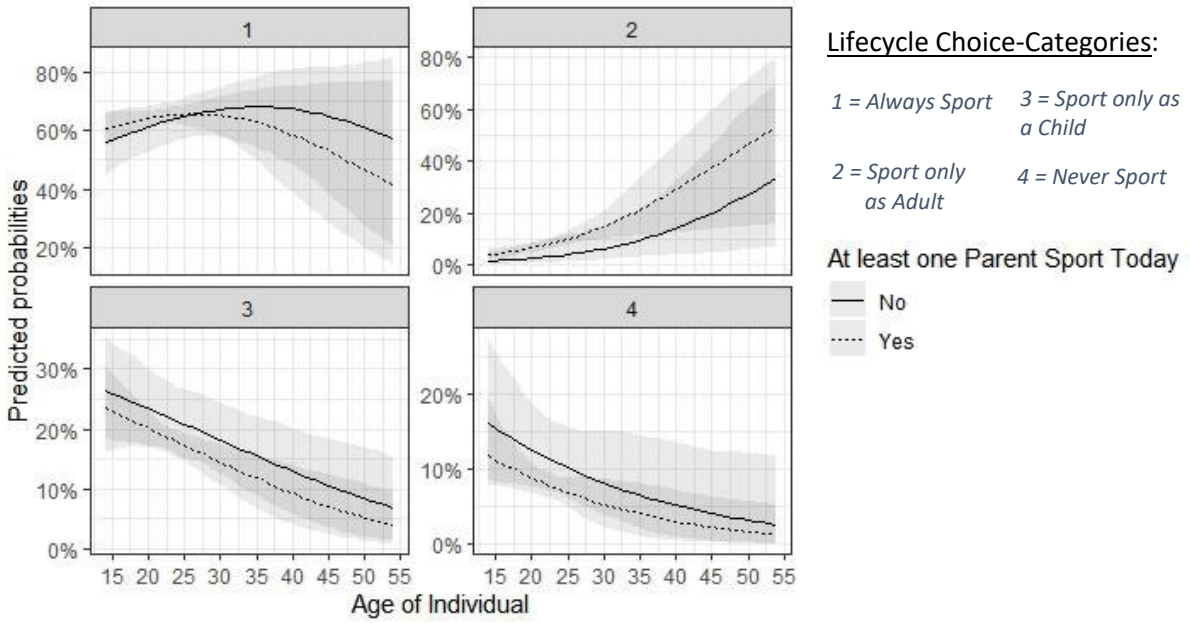
Table 18: Multinomial Logit Model with Lifecycle Sport Choice of Parents

	Dependent variable: lifecycle choice-sequence for weekly Sport					
	Pooled MNL			MNL with lifecycle choice of parents		
	Sport only as an Adult	Sport only as a Child	Never Sport	Sport only as an Adult	Sport only as a Child	Never Sport
	(1)	(2)	(3)	(4)	(5)	(6)
Parents doing Sport	1.000	-0.318	-0.495			
	(0.626)	(0.291)	(0.349)			
Parents Sport when Young	0.641	0.131	-0.373			
	(0.567)	(0.399)	(0.462)			
At least one Parent always Sport				0.815*	-0.059	-0.426
				(0.433)	(0.252)	(0.300)
Mother: Secondary Educ	-0.617	-0.207	-0.448	-0.626	-0.202	-0.443
	(0.599)	(0.426)	(0.477)	(0.597)	(0.428)	(0.477)
Mother: Tertiary Educ	-0.925	-0.288	-0.181	-0.946	-0.266	-0.169
	(0.649)	(0.463)	(0.519)	(0.648)	(0.464)	(0.519)
Father: Secondary Educ	1.043	-0.398	0.147	1.069	-0.454	0.092
	(1.075)	(0.494)	(0.654)	(1.072)	(0.488)	(0.650)
Father: Tertiary Educ	0.953	-0.547	-0.594	0.983	-0.598	-0.651
	(1.076)	(0.502)	(0.677)	(1.073)	(0.496)	(0.673)
2 People living in Hh	-0.609	-0.068	0.021	-0.639	-0.027	0.036
	(0.549)	(0.436)	(0.838)	(0.547)	(0.433)	(0.835)
3 People living in Hh	-0.322	0.533	0.973	-0.350	0.583	0.985
	(0.592)	(0.483)	(0.852)	(0.592)	(0.480)	(0.850)
4 People living in Hh	-0.407	0.318	0.496	-0.429	0.365	0.510
	(0.581)	(0.484)	(0.862)	(0.581)	(0.481)	(0.859)
5 People living in Hh	-0.111	0.574	0.609	-0.130	0.631	0.626
	(0.618)	(0.519)	(0.886)	(0.617)	(0.516)	(0.883)
6 or more People living in Hh	-0.657	0.794	0.650	-0.679	0.816	0.656
	(0.820)	(0.594)	(0.959)	(0.820)	(0.593)	(0.958)
Age	0.079***	-0.053*	-0.072*	0.077***	-0.054*	-0.071
	(0.030)	(0.028)	(0.043)	(0.029)	(0.028)	(0.043)
Gender	0.554**	0.471**	0.561**	0.554**	0.477**	0.564**
	(0.271)	(0.197)	(0.254)	(0.271)	(0.197)	(0.254)
Number of Doctor Consultations: Last 12 Months	-0.047	0.089***	0.003	-0.046	0.089***	0.002
	(0.044)	(0.018)	(0.037)	(0.044)	(0.018)	(0.038)
Being Underweighted	-0.209	-0.563	-0.747*	-0.210	-0.542	-0.735*
	(0.475)	(0.369)	(0.435)	(0.475)	(0.368)	(0.433)

Being Overweighted	0.136	-1.414***	-1.684***	0.133	-1.407***	-1.678***
	(0.363)	(0.326)	(0.496)	(0.362)	(0.326)	(0.495)
Being Obese or Severly Obese	-1.205	-1.587***	-1.144	-1.256	-1.567**	-1.118
	(1.050)	(0.605)	(0.750)	(1.049)	(0.611)	(0.752)
Number of hours worked per week	0.0002	0.016***	0.012	0.0002	0.016***	0.012
	(0.007)	(0.006)	(0.007)	(0.007)	(0.005)	(0.007)
Main Income-Contributor in the Hh	-0.293	0.073	-0.318	-0.294	0.080	-0.316
	(0.464)	(0.353)	(0.601)	(0.465)	(0.353)	(0.598)
Personal Education: Secondary Educ	0.245	-0.760***	-1.465***	0.260	-0.742***	-1.476***
	(0.367)	(0.282)	(0.383)	(0.366)	(0.282)	(0.382)
Personal Education: Tertiary Educ	0.045	-1.357***	-1.484***	0.083	-1.347***	-1.501***
	(0.517)	(0.392)	(0.556)	(0.513)	(0.393)	(0.557)
Number of Hours on a Screen per Day	-0.096	0.045	0.099**	-0.099	0.049	0.100**
	(0.077)	(0.040)	(0.040)	(0.077)	(0.040)	(0.040)
Feeling Unheathy: last Year	0.115	0.160	0.610*	0.126	0.155	0.607*
	(0.399)	(0.298)	(0.331)	(0.398)	(0.298)	(0.331)
Health-Expectation due to Sport	-0.082	-1.513***	-1.183***	-0.085	-1.515***	-1.183***
	(0.122)	(0.127)	(0.161)	(0.122)	(0.127)	(0.161)
Constant	-5.871***	3.488***	2.953*	-4.978***	3.370***	2.519*
	(1.766)	(1.151)	(1.587)	(1.607)	(1.087)	(1.530)
Akaike Inf. Crit.	1,834.112	1,834.112	1,834.112	1,830.401	1,830.401	1,830.401
Observations		1068			1068	
Note:						*p<0.1; **p<0.05; ***p<0.01

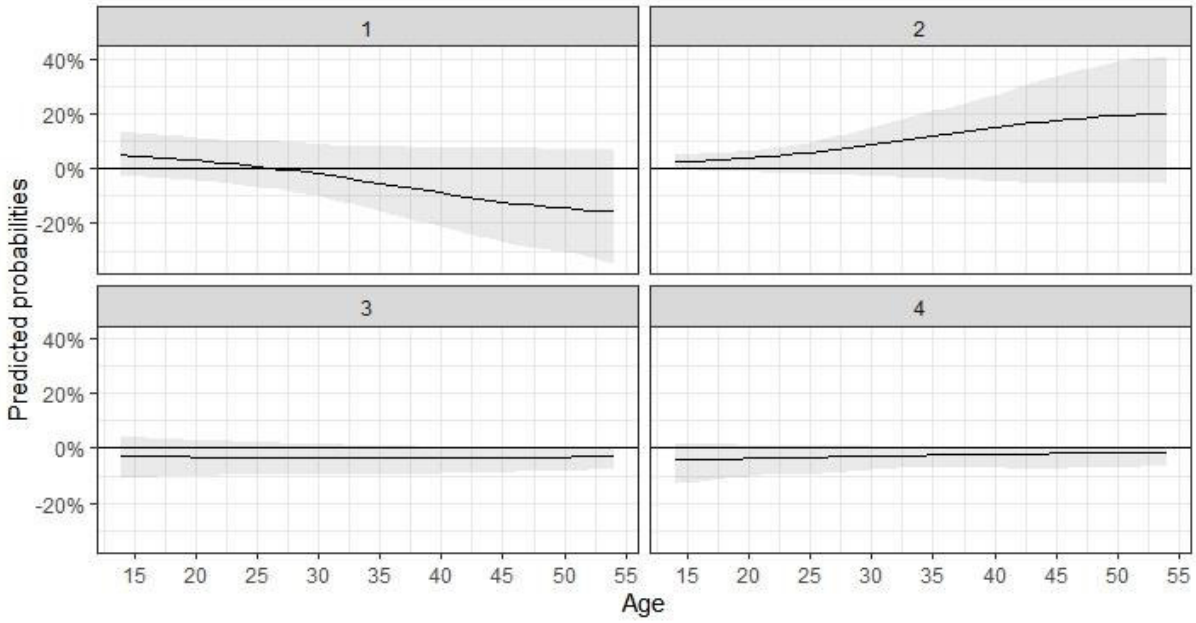
L) Multinomial Logit Model: Dynamics of Choice Probabilities and First Differences

Figure 20: Predicted Lifecycle Sport-Choice-Probabilities for Individuals with at least one Parent doing Sport Today VS. No Parent doing Sport Today across Ages based on Model in Table 6



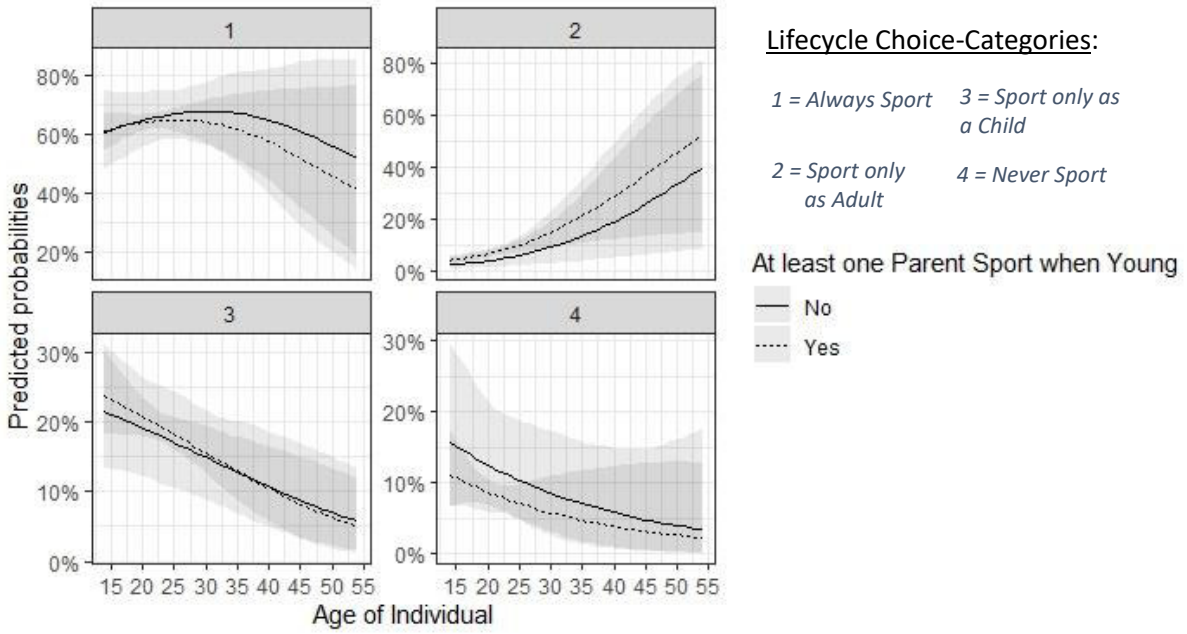
Note: The gray areas around the black dotted or straight lines are the 95%-confidence intervals.

Figure 21: Differences in Lifecycle Sport-Choice-Probabilities between Individuals with at least one Parent doing Sport Today VS. No Parent doing Sport Today across Ages based on Figure 20



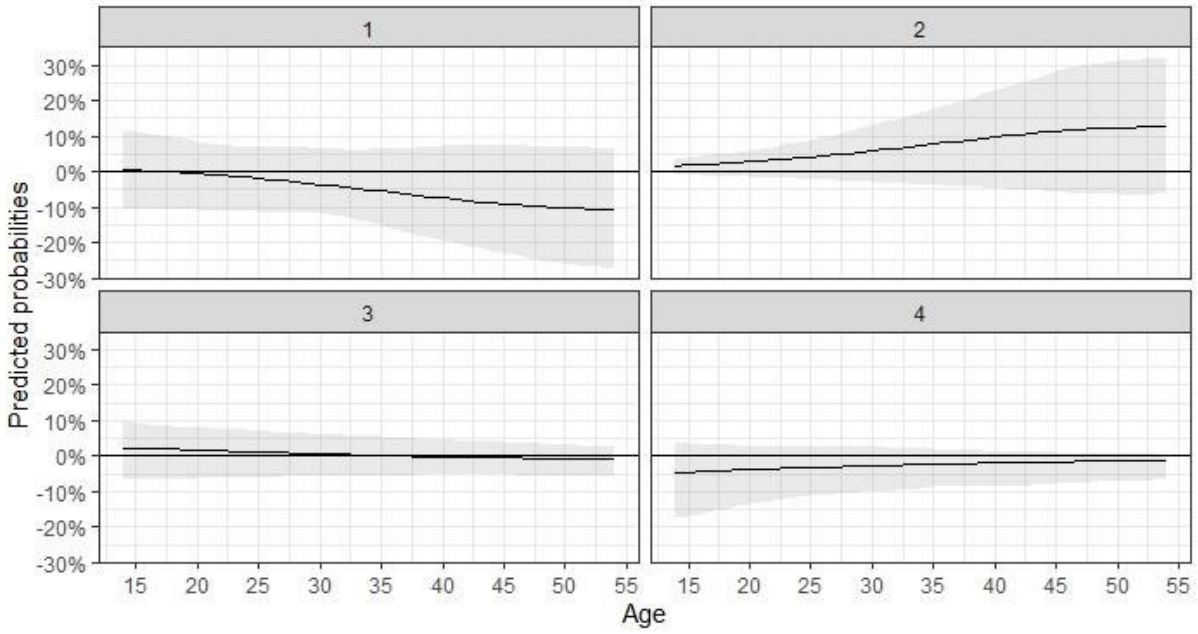
Note: The gray areas around the black lines are the 95%-confidence intervals. The Lifecycle choice-categories are defined as follows: 1) Always Sport; 2) Sport only as Adult; 3) Sport only as a Child; 4) Never Sport.

Figure 22: Predicted Lifecycle Sport-Choice-Probabilities for Individuals with at least one Parent doing Sport when Young VS. No Parent doing Sport during their Youth across Ages based on Model in Table 6



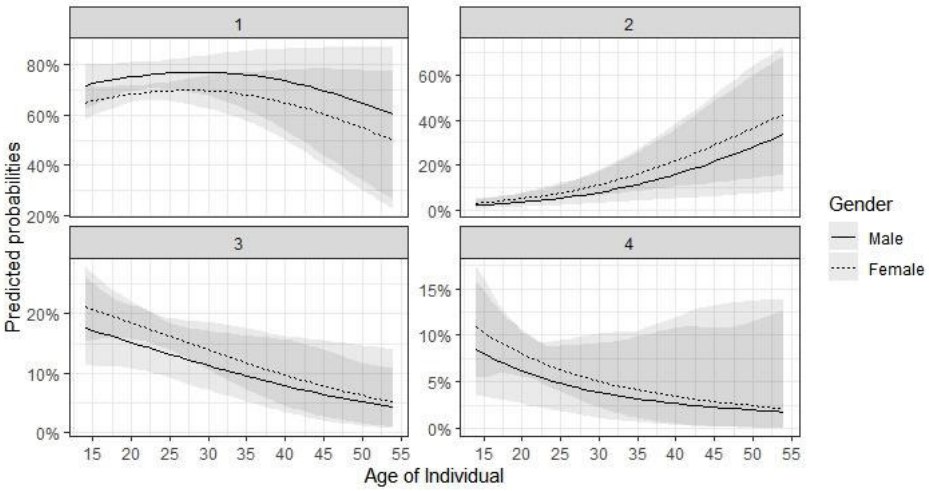
Note: The gray areas around the black lines are the 95%-confidence intervals

Figure 23: Differences in Lifecycle Sport-Choice-Probabilities between Individuals with at least one Parent doing Sport when Young VS. No Parent doing Sport during their Youth across Ages based on Figure 22



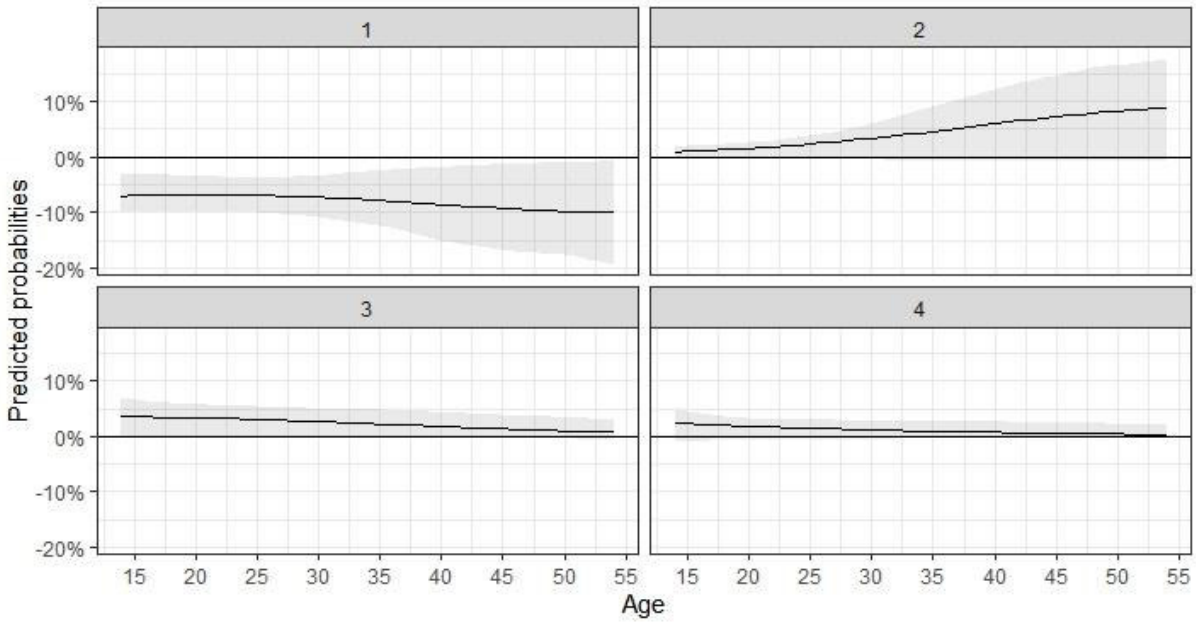
Note: The gray areas around the black lines are the 95%-confidence intervals. The Lifecycle choice-categories are defined as follows: 1) Always Sport; 2) Sport only as Adult; 3) Sport only as a Child; 4) Never Sport.

Figure 24: Predicted Lifecycle Sport-Choice-Probabilities for Women VS. Men during their Life across Ages based on Model in Table 6



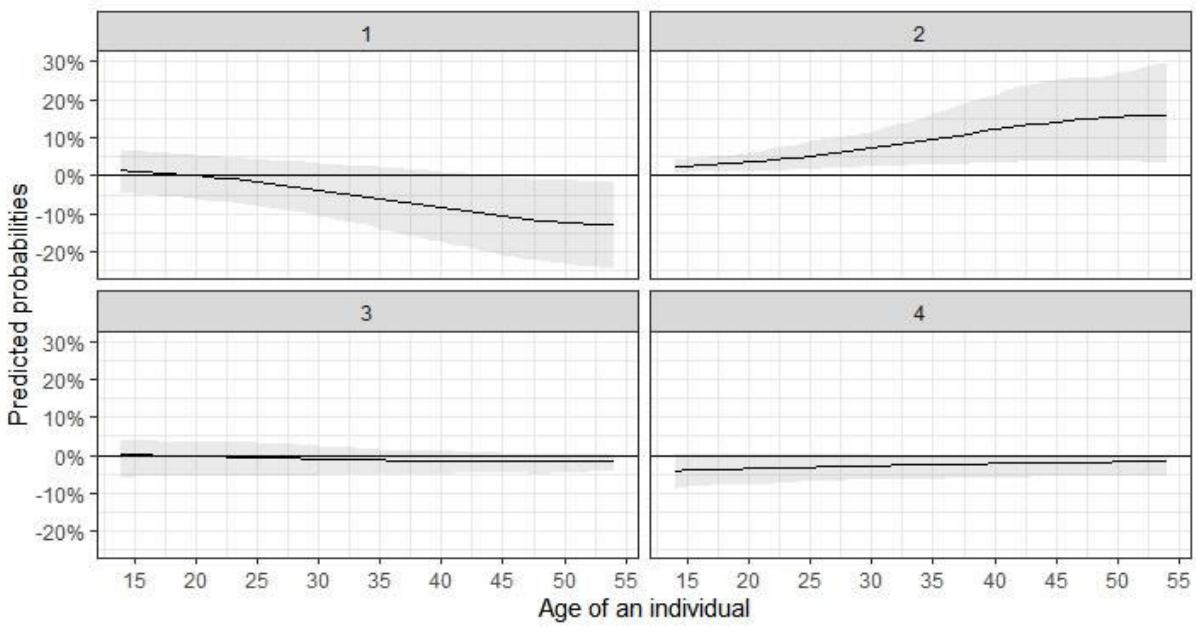
Note: The gray areas around the black dotted or straight lines are the 95%-confidence intervals. The Lifecycle choice-categories are defined as follows: 1) Always Sport; 2) Sport only as Adult; 3) Sport only as a Child; 4) Never Sport.

Figure 25: Differences in Lifecycle Sport-Choice-Probabilities for Women VS. Men during their Life across Ages based on Figure 24



Note: The gray areas around the black lines are the 95% confidence intervals. The Lifecycle choice-categories are defined as follows: 1) Always Sport; 2) Sport only as Adult; 3) Sport only as a Child; 4) Never Sport.

Figure 26: Differences in Lifecycle Sport-Choice-Probabilities between Individuals with at least one Parent always doing Sport VS. No Parent doing Sport during their Life across Ages based on Figure 3



Note: The gray areas around the black lines are the 90% confidence intervals. The Lifecycle choice-categories are defined as follows: 1) Always Sport; 2) Sport only as Adult; 3) Sport only as a Child; 4) Never Sport.

M) Multinomial Logit Model: Score Test for Heteroscedasticity

Table 19: Score-Test for Heteroscedastic Error-Term

Score-Test Results	Estimates
Chi-Squared	12.908
p-value	0.00484

Note: The Nullhypothesis is having homoscedastic errors, which is rejected on the 1% significance-level.

Statutory Declaration / Affidavit

I hereby declare that the thesis with the title

An Empirical Analysis of the Formation of Sports Preferences in Switzerland; with a Focus on Inter- and Intragenerational Factors

has been composed by myself autonomously and that no means other than those declared were used. In every single case, I have marked parts that were taken out of published or unpublished work, either verbatim or in a paraphrased manner, as such through quotation. This thesis has not yet been handed in or published in the same or similar form.

Zurich, 13.10.2020