Optimal Paternalism? A New View of the Taxation of Unhealthy Food

Zarko Kalamov* Marco Runkel†

March 7, 2017

Abstract

This paper analyzes optimal consumption choices and weight when consumers make errors due to self-control problems and naiveté. Contrary to the existing literature, we show that both types of errors do not determine whether an individual is overweight or underweight. Instead, they affect only the extent of over- or underweight and do not impact a healthy weight consumer. Consequently, a paternalistic tax on unhealthy food that is designed to correct self-control problems and naiveté cannot induce individuals to have a healthy weight. Moreover, there exists a tax that induces both rational and non-rational individuals to have a healthy weight in steady state. This tax does not depend on the degree of self-control problems or naiveté.

Keywords: obesity; paternalism; time inconsistency; fat tax;
JEL Code: D11;I12;H21;H51;

1 Introduction

High prevalence of obesity is a major health problem in many countries which has led economists to search for its causes and policies to address it. The seminal articles of O'Donoghue and Rabin (2003, 2006) view the overconsumption of unhealthy foods as the

*TU Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany. zarko.y.kalamov@tu-berlin.de.
†TU Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany. marco.runkel@tu-berlin.de.

result of individuals' self-control problems. They investigate "optimal paternalism", i.e. search for policies which correct the errors of the non-rational individuals at the lowest costs for the remaining consumers. Within this framework, a tax on unhealthy foods (a fat tax) is such a policy, as it imposes only a second-order cost to the fully rational individuals and a first-order benefit to the consumers without full self-control.

While the literature on optimal paternalism assumes that fully rational individuals do not overconsume, another strand of the economic literature on obesity shows that consumers may choose to become overweight even if they fully take into account its negative health consequences (Levy, 2002). Levy shows that rational obesity occurs either when the instantaneous utility of unhealthy food is high or its price is low.

This paper merges the literatures on optimal paternalism and rational obesity. We use a dynamic version of the optimal paternalism framework of O'Donoghue and Rabin (2006) in which a representative consumer may exhibit not only self-control problems but also naive expectations regarding her future consumption. The individual may be overweight, underweight or have a healthy weight. The negative health consequences of abnormal weight are increasing in weight if the individual is overweight, decreasing if she is underweight and achieve a minimum at the healthy weight.

We analyze the effects of self-control problems and naiveté on the steady state weight. We show that a higher degree of self-control problems raises the weight of an overweight consumer, lowers the weight of an underweight individual and does not affect an individual with a healthy weight. Thus, this form of non-rationality does not determine whether an individual is over-, healthy- or underweight. Instead, it can only influence the degree of the problem of abnormal weight. Furthermore, we show that naiveté also does not influence how the consumer's steady state weight relates to the healthy weight.

These results emerge because in a steady state with a healthy weight, a marginal increase in consumption does not create additional health problems, i.e. there are zero marginal health costs. Therefore, whether an individual can maintain such a weight in steady state does not depend on how she discounts the health consequences of unhealthy consumption (the degree of self-control problems) or how she expects to discount them in the future (the degree of naiveté). An optimal consumption path is compatible with a healthy steady state if an only if the marginal instantaneous utility of consumption is

also zero. Hence, this steady state can only be achieved when the price of unhealthy foods is such that the consumption level compatible with a healthy weight equals the instantaneous utility-maximizing consumption level.

Our second purpose is to analyze the fat tax, when it is set by a social planner who is paternalistic, i.e. maximizes the 'true' utility of the consumer. First, we consider a tax that is returned in lump-sum fashion to the representative individual. This tax equals optimally zero if the consumer is fully rational and nonzero otherwise. Moreover, it cannot be used to induce rational or non-rational individuals to achieve a healthy weight. Second, we consider a tax whose proceeds are fully returned to the consumer only in steady state but may not be fully returned during transition. The optimal level of this tax does achieve a healthy weight. Furthermore, its steady state value is independent of the degree of non-rationality of the consumer. Finally, we consider a social planner who cannot condition the tax rate on the consumers' self-control problems and naiveté. In this case the government can induce a healthy weight by levying a constant tax rate equal to the steady state tax from the previous case. The tax in this last scenario is non-paternalistic as it does not correct errors in consumption choices. Nevertheless, it incentivizes both rational and non-rational individuals to follow a healthy consumption path.

The non-paternalistic tax has several advantages relative to paternalistic taxation. First, it can induce a healthy steady state, which is not possible for taxes designed to correct non-rationalities. Second, it is easier to implement as it does not require information on the share of the population with self-control problems and naiveté and the severity of these problems. This is even more important, considering the sensitivity of the optimal paternalistic tax to small changes in the degree of non-rationality. O'Donoghue and Rabin (2006) calibrate the optimal paternalistic tax on potato chips in a heterogeneous population where half of the population is fully rational, while the other half exhibit self-control problems, represented by hyperbolic discounting. If the health costs of chips consumption are relatively high, the optimal tax rises from 5% to 28% when the rate of hyperbolic discounting of the second half of the population drops from 0.99 to 0.95. When this rate drops further to 0.9, the optimal tax becomes 63%. This high sensitivity may lead to errors in designing the optimal paternalistic tax.

This paper is related to two strands of literature. The first investigates the causes of obesity and the second examines the paternalistic taxation of unhealthy food.

The literature on the causes of obesity is very diverse. Obesity could be explained through rational addiction (Becker and Murphy, 1988) or technical change that lowers food prices and raises the cost of physical activity (Philipson and Posner, 1999; Lakdawalla et al., 2005). A third strand of literature focuses on self-control problems, that are formalized as quasi-hyperbolic discounting, as a reason for over- or under-consumption (see e.g. Laibson, 1997; O'Donoghue and Rabin, 2003, 2006). Yet another explanation is provided by the rational obesity literature. Levy (2002) shows that a rational individual may choose to be obese. The model of rational obesity can explain the occurrence of alternating diets and binges if habits in consumption are considered (Dragone, 2009). Moreover, (Dragone and Savorelli, 2012) show that the rational individual may also choose to be underweight if the utility of consumption is relatively low or the price of food high.

This article contributes to the literature by embedding the rational obesity model to the framework of quasi-hyperbolic discounting. The resulting model can better explain the role that non-rational behavior plays in the determination of consumption and weight.

The optimal paternalistic tax on unhealthy food is first studied by O'Donoghue and Rabin (2003, 2006). They derive large values for the optimal tax in a heterogeneous population where a small share of consumers exhibit non-rational behavior. This result emerges because the costs that the tax imposes on rational individuals is of second order. Moreover, they show that there exist Pareto improving positive tax rates when the tax revenues are returned as lump-sum transfers to consumers, as this policy results in redistribution from the individuals with high consumption to those with low consumption. Haavio and Kotakorpi (2011) show that individuals with self-control problems have incentives to vote for such taxes.

Our contribution to the literature consists in deriving a non-paternalistic tax that can induce both individuals with self-control problems and those without to be healthy weight. Moreover, we show that the paternalistic tax can only correct the problems arising from non-rational behavior but cannot be health-maximizing, i.e. the optimal paternalistic tax cannot induce an overweight individual to choose a consumption path compatible with a healthy steady state weight.

The rest of the article is organized as follows. In Section 2 we present the model and analyze the optimal consumption path and steady state weight. In Section 3 we analyze the optimal government policy. Section 4 concludes.

2 The Model

This model merges the literature on optimal paternalism (see e.g. O'Donoghue and Rabin, 2006) and the literature on rational obesity (Levy, 2002; Dragone and Savorelli, 2012). A representative individual consumes unhealthy food x_t and a bundle of other goods z_t in period t. They give rise to a quasi-linear instantaneous utility $u_t \equiv v(x_t)+z_t-c(w_t)$, where $v'(\cdot) > 0 > v''(\cdot)$, w_t denotes the weight of the individual in period t and $c(w_t)$ represents the negative health consequences of abnormal weight. There exists a healthy weight w^H, which minimizes the health problems, i.e. $c'(w_t) \gtreqless 0 \Leftrightarrow w_t \lesseqgtr w^H$. Furthermore, we require $c''(\cdot) > 0$, which guarantees that the consumption choices of the individual are well-behaved. Weight at time t depends on the consumption of junk-food in all previous periods according to the following equation of motion:

$$w_t = x_{t-1} + (1-d)w_{t-1}, \tag{1}$$

where $d \in]0,1[$ denotes the effect of burning calories on weight. We assume that the individual may have a present-bias, i.e. seek immediate gratification, which is inconsistent with her long term preferences. This present-bias is modelled by allowing for quasi-hyperbolic discounting in the lifetime utility of the agent in period t, U_t, as introduced by Laibson (1997):

$$U_t = u_t + \beta \sum_{s=t+1}^{T} \delta^{s-t} u_s, \tag{2}$$

where $\delta \in]0,1]$ denotes the degree of exponential discounting and $\beta \in]0,1]$ the rate of hyperbolic discounting. If $\beta = 1$, then there is no present-bias and the preferences are time-consistent. On the other hand, $\beta < 1$ denotes desire for immediate gratification and time-inconsistency, as the discount factor between any two consequtive future periods δ is larger than the discount rate between the current and next period $\beta\delta$. In the subsequent analysis, we will use interchangeably the terms 'self-control problems' and 'present-bias' in referring to the case $\beta < 1$.

The individual's income at time t is denoted by I. Both goods are produced at constant unitary marginal cost under perfect competition and, therefore, their prices equal one. However, the government may impose a tax τ_t on unhealthy food in period t. The proceeds are either returned to the individual as a lump-sum transfer ℓ_t or are used to subsidize the composite good at a rate σ_t. This subsidy may be interpreted as a reduction in the sales tax on other goods in such a way that the introduction of a tax on junk-food is revenue-neutral. Thus, the time t budget constraint is

$$(1+\tau_t)x_t + (1-\sigma_t)z_t = I + \ell_t. \tag{3}$$

Each period the individual chooses x_t and z_t so as to maximize the lifetime utility (2) under consideration of the equation of motion for weight (1) and the budget constraint (3). If the individual exhibits present-bias, then the optimal consumption path depends on whether and to what extent the individual expects her future selves to behave time-inconsistently, i.e. how sophisticated the agent is. We follow O'Donoghue and Rabin (2001) and assume that an agent with discount rate β expects her future selves to have a taste for immediate gratification $\hat{\beta} \in [\beta, 1]$. If $\hat{\beta} = \beta < 1$, then the individual is said to be sophisticated, i.e. she anticipates perfectly her future self-control problems. On the other hand, an individual is naive if she is characterized by $\beta < 1 \wedge \hat{\beta} = 1$, as this individual is not aware of the present-bias of her future selves. Partial naiveté is present when $\beta < \hat{\beta} < 1$.[1] In order to distinguish the different types of individuals in the remaining analysis, we will index consumption and weight using a superscript $i = s, n$, where s denotes a sophisticated individual and n a (fully or partially) naive individual.

Before we solve the individual problem of utility maximization, we can summarize how the model differs from the existing literature. The literature on optimal paternalistic taxes (see e.g. O'Donoghue and Rabin, 2003, 2006, and others) assumes as a simplification that the health problems in period t are monotonically increasing in the unhealthy consumption in period $t-1$. By modelling the negative health consequences as a function of weight and postulating a non-monotone relationship between weight and health centered

[1] This form of modelling the degree of sophistication of individuals with self-control problems has become standard in the literature. See e.g. Gruber and Köszegi (2001, 2004) for application to cigarette consumption, Diamond and Köszegi (2003) in the context of quasi-hyperbolic discounting and retirement, and others.

aroung the healthy weight level w^H, we allow for rational agents to be either overweight or underweight. This approach is consistent with the literature on rational obesity, as developed by Levy (2002); Dragone and Savorelli (2012) and others. However, this strand of literature does not discuss how a rational individual differs from a non-rational individual in becoming obese and what the consequences of rational obesity for the optimal taxation of unhealthy consumption are. This is the purpose of this article.

2.1 Optimal Consumption

The representative individual of type i maximizes the perceived lifetime utility at time t, given by Equation (2). The optimal consumption is derived from the solution of the Bellman equation

$$V^i(w_t^i) = \max_{x_t^i} \left\{ u(x_t^i, w_t^i) + \beta\delta V(w_{t+1}^i) \right\}, \tag{4}$$

where $V^i(w_t^i)$ is the value function, which gives the discounted lifetime utility of leaving a weight w_t^i in period t and consuming optimally afterwards. Using Equations (1) and (3), we can derive the following first-order condition:

$$v'(x_t^i) - \frac{1+\tau_t}{1-\sigma_t} + \beta\delta V^{i\prime}(w_{t+1}^i)\frac{\partial w_{t+1}^i}{\partial x_t^i} = v'(x_t^i) - \frac{1+\tau_t}{1-\sigma_t} + \beta\delta V^{i\prime}(w_{t+1}^i) = 0. \tag{5}$$

As a second step, we derive $V^{i\prime}(w_{t+1}^i)$, which from the perspective of the self in period t is determined by

$$V^i(w_{t+1}^i) = u_{t+1}(x_{t+1}^s(\hat{\beta}), w_{t+1}^i) + \delta V^i(w_{t+2}^i). \tag{6}$$

Two comments are necessary. First, Equation (6) is derived from the perspective of the self in period t and, therefore, the individual discounts exponentially at the rate δ between periods $t+1$ and $t+2$ in accordance with the lifetime utility (2). Second, the individual believes that her future selves in periods $t+1, t+2, \ldots$ will have self-control problems $\hat{\beta}$. Thus, she expects to be a sophisticated consumer with $\beta = \hat{\beta}$ from period $t+1$ onwards and to consume $x_{t+1}^s(\hat{\beta})$ in that period. We differentiate the above equation with respect to w_{t+1}^i and derive the following value for $V^{i\prime}(w_{t+1}^i)$:

$$V^{i\prime}(w_{t+1}^i) = \left[v'(x_{t+1}^s(\hat{\beta})) - \frac{1+\tau_{t+1}}{1-\sigma_{t+1}} \right] \frac{\partial x_{t+1}^s(\hat{\beta})}{\partial w_{t+1}^i} - c'(w_{t+1}^i) + \delta V^{i\prime}(w_{t+2}^i) \left[(1-d) + \frac{\partial x_{t+1}^s(\hat{\beta})}{\partial w_{t+1}^i} \right]. \tag{7}$$

The last step in deriving the optimal stream of consumption is to solve the maximization problem that the self in t expects to solve in $t+1$, which is given by

$$V^i(w^i_{t+1}) = \max_{x^s_{t+1}} \left\{ u(x^s_{t+1}, w^i_{t+1}) + \hat{\beta}\delta V^i(w^i_{t+2}) \right\}. \tag{8}$$

Note that the only difference between Equations (4) and (8) is that the expected self-control problem $\hat{\beta}$ may differ from the actual present-bias β. The expected first-order condition is given by

$$v'(x^s_{t+1}(\hat{\beta})) - \frac{1+\tau_{t+1}}{1-\sigma_{t+1}} + \hat{\beta}\delta V^{i\prime}(w^i_{t+2}) = 0. \tag{9}$$

Lastly, one can plug Equation (5) in (7), solve for $V^{i\prime}(w^i_{t+2})$ and plug the resulting expression in Equation (9) in order to derive the Euler equation of the individual. Denoting the relative price of unhealthy food in period t as $p_t \equiv (1+\tau_t)/(1-\sigma_t)$, we derive the following result:

$$v'(x^i_t) - p_t = \frac{\beta\delta}{\hat{\beta}} \left[\left(v'(x^s_{t+1}(\hat{\beta})) - p_{t+1} \right) \left((1-d) + (1-\hat{\beta}) \frac{\partial x^s_{t+1}(\hat{\beta})}{\partial w^i_{t+1}} \right) + \hat{\beta} c'(w^i_{t+1}) \right]. \tag{10}$$

Equation (10) looks complicated, but can be easily interpreted. Assume for the moment that both sides of (10) are positive. Note furthermore that along the optimal path, a small increase in consumption in period t, followed by a small reduction in period $t+1$, does not affect utility. The term on the left-hand side of (10) gives the marginal utility that a consumer derives of consuming one more unit of unhealthy food in period t, while the first term in brackets on the right-hand side gives the reduction in utility from consuming one unit less in the next period. The second term on the right-hand side displays the utility loss of this perturbation of the consumption path in terms of higher weight and, therefore, more health problems.

A steady state level of consumption and weight can be reached when the relative price is constant. Denote the steady state values of the variables as $\tilde{x}^i, \tilde{p}, \tilde{w}^i$. While the determinants of $\tilde{p}$ are analyzed in the next section, where we consider the optimal government policy, $\tilde{x}^i$ and $\tilde{w}^i$ are determined by Equations (1) and (10) and are given by

$$d\tilde{w}^i = \tilde{x}^i, \tag{11}$$

$$\left(v'(\tilde{x}^i) - \tilde{p} \right) = \left(v'(\tilde{x}^s(\hat{\beta})) - \tilde{p} \right) \frac{\beta\delta}{\hat{\beta}} \left[(1-d) + (1-\hat{\beta}) \frac{\partial \tilde{x}^s(\hat{\beta})}{\partial \tilde{w}^i} \right] + \beta\delta c'(\tilde{w}^i). \tag{12}$$

The literature on rational obesity examines whether an individual without self-control problems is optimally over- or underweight in steady state (see e.g. Dragone and Savorelli, 2012). In order to compare our results to that literature, we first need to derive a solution for the steady-state consumption $\tilde{x}^i$ and its dependance on weight.

2.2 Closed-form solution for the consumption path

We follow Gruber and Köszegi (2001, 2004) and derive a closed-form solution for junk-food consumption by assuming that $v(x)$ and $c(w)$ are quadratic:

$$v(x) = \gamma x - \frac{\varepsilon}{2}x^2, c(w) = \bar{c} + \frac{\omega}{2}(w - w^H)^2. \tag{13}$$

In Appendix A we show that the Euler equation (10) and the assumed functional forms in (13) lead to junk-food consumption being a linear function of weight, given by $x_t^i = \lambda_t^i w_t^i + \mu_t^i$. We derive the following results:

Proposition 1. *If the instantaneous utility is represented by quadratic functional forms, then optimal consumption satisfies* $x_t^i = \lambda_t^i w_t^i + \mu_t^i$. *If the consumer has an infinite time horizon and* $\beta \geq 1/2$, *then* λ_t^i *converges to the constant value* $\lambda^{*i} \in]-(1-d), 0[$, *which is independent of the price of unhealthy food. Additionally,* μ_t^i *converges to the value* $\mu_t^{*i}(\mathbf{p_t})$, *where* $\mathbf{p_t} = (p_t, p_{t+1}, \ldots)^T$.

Proof: See Appendix A.

Several comments are necessary. First, we can show that the assumption of quasi-linear preferences leads to λ^{*i} being independent of the price of unhealthy food and, hence, of the tax rate. This result simplifies the analysis as the introduction of a tax impacts consumption only through the term μ_t^{*i}. Second, $\lambda^{*i} < 0$ emerges because of the additive-separability of the instantaneous utility in current-period consumption and weight. If instead $u_{xw} > 0$, then the possibility of $\lambda^{*i} > 0$ emerges. However, the exact value of λ^{*i} is not essential to the remaining analysis.

2.3 Steady State

Proposition 1 can be used to analyze the steady state described by Equations (11)-(12). In order to simplify the subsequent analysis, it is useful to define a satiation level of

consumption x^F, as the junk-food intake, which maximizes instantaneous utility in steady state and the healthy consumption x^H, which is compatible with healthy long-term weight, i.e.

$$v'(x^F) - \tilde{p} = 0, \quad dw^H = x^H. \tag{14}$$

A consumer is said to be overconsuming if $\tilde{x}^i > x^F$ and underconsuming if $\tilde{x}^i < x^F$. Moreover, an individual is overweight if $\tilde{x}^i > x^H$ and underweight otherwise. Note furthermore that x^F and x^H are the same for all types of individuals, because x^F is determined by the instantaneous utility function and the relative price, while x^H is determined by the level of healthy weight and the equation of motion. We can derive the following results:

Proposition 2. *There exist three possible steady states for the consumer of type $i = s, n$. The individual is either (a) overweight and underconsuming if $x^H < \tilde{x}^i < x^F$, (b) underweight and overconsuming if $x^F < \tilde{x}^i < x^H$ or (c) healthy weight and consuming until satiation is achieved if $x^H = \tilde{x}^i = x^F$.*

The condition $x^F > x^H$ is necessary and sufficient for an overweight steady state. The conditions $x^F < x^H$ is necessary and sufficient for an underweight steady state. A necessary and sufficient condition for a healthy weight steady state is $x^F = x^H$.

Proof: See Appendix B.

The first part of Proposition 2 is analogous to the result of Dragone and Savorelli (2012), who consider only rational individuals. We generalize their result by proving that it continues to hold when consumers might exhibit self-control problems and naiveté. Furthermore, according to Proposition 2, the relation between the satiation level of consumption and the healthy consumption is sufficient for the determination of whether an individual is overweight or underweight. In the next Proposition we consider explicitly the implications of self-control problems and naiveté for the steady state weight:

Proposition 3. *The degree of self-control problems does not impact the decision of being under-/healthy- or overweight. An increase in the degree of self-control problems (a reduction in β) raises the weight of an overweight consumer, lowers the weight of an underweight individual and does not impact a healthy weight individual.*

The degree of naiveté does not impact the decision of being under-/healthy- or overweight. An increase in naiveté (higher $\hat{\beta}$) does not affect the steady state weight of a

healthy weight individual and has an ambiguous effect on the steady state weight, if the individual is over- or underweight.

Proof: See Appendix C.

Proposition 3 has the following interpretation. Self-control problems lead to under-evaluation of the future negative healthy consequences associated with abnormal weight. An over-(under-)weight individual underestimates the negative consequences of being over-(under-)weight and, thus, consumes more (less) than a consumer without present-bias. Therefore, hyperbolic discounting can only worsen the problem of abnormal weight but cannot determine whether an individual's weight is above or below the healthy level. While naiveté also does not determine whether $\tilde{w}^n \gtreqless w^H$, its impact on the steady state weight is ambiguous. On the one hand, lack of sophistication leads to wrong expectations of future consumption that lower the perceived effect of consumption today on weight and aggravate the problem of over- or underweight. On the other hand, if the individual expects to consume time-consistently in the future, consumption smoothing mitigates the problem of over- or underconsumption. It is unclear which effect dominates.

Note that the condition for a healthy steady state is a knife edge condition. While x^F is determined by the instantaneous utility and the relative price of unhealthy food, x^H is determined by how many calories a healthy weight individual burns per period, i.e. dw^H. However, public policy can influence x^F through changes in the relative price of unhealthy food and induce $\tilde{w}^i = w^H$.

2.4 The effect of price changes on consumption

Proposition 1 shows that prices affect consumption only through the parameter μ_t^{*i}. In Appendix A we derive the values of μ_t^{*i} for sophisticated and for naive consumers. Since the degree of naiveté impacts how prices affect consumption, we consider first the response of a sophisticated individual to a change in the tax on unhealthy food in period t. In Appendix A we show that

$$\mu_t^{*s} = \tilde{\mu}^{*s} - \frac{(p_t - \tilde{p}) - [\delta\left((1-d) + (1-\beta)\lambda^{*s}\right) - k^{*s}] \sum_{i=1}^{\infty} k^{*s\,i-1}(p_{t+i} - \tilde{p})}{\varepsilon - \delta\left(\varepsilon\lambda^{*s}\left((1-d) + (1-\beta)\lambda^{*s}\right) - \beta\omega\right)}, \tag{15}$$

where $k^{*s} \in]0,1[$ is a constant defined in Appendix A and $\tilde{\mu}^{*s}$ is the steady state value of μ_t^{*s} associated with the steady state price $\tilde{p}$ (see Equation (A.18) for a formal expression of $\tilde{\mu}^{*s}$). We see that consumption depends not only on the current price, but also on the prices in all subsequent periods. Therefore, the impact of a change of the tax rate in period t on consumption in period t depends on what the consumer expects to be its effect on future prices. We assume that the individual expects future prices to evolve according to the following general rule:

$$(p_{t+1} - \tilde{p}) = a(p_t - \tilde{p}),\ a \in]0,1[, \tag{16}$$

where $a \in]0,1[$ ensures that the tax rate and, hence, the price converge to their steady state values. We will later prove that the government's optimal policy follows a rule of the same form as in Equation (16). Nevertheless, we do not assume that the individual has perfect foresight regarding the value of a and may or may not expect a to equal the actual value chosen by the government. Thus, μ_t^{*s} can be rewritten as

$$\mu_t^{*s} = \tilde{\mu}^{*s} - \frac{1 - a\delta\left((1-d) + (1-\beta)\lambda^{*s}\right)}{(1 - k^{*s}a)\left[\varepsilon - \delta\left(\varepsilon\lambda^{*s}\left((1-d) + (1-\beta)\lambda^{*s}\right) - \beta\omega\right)\right]}(p_t - \tilde{p}). \tag{17}$$

Now, we can derive the impact of a change in the tax rate τ_t on the consumption of a sophisticated individual in period t, which is given by

$$\frac{dx_t^s}{d\tau_t} = \frac{d\mu_t^{*s}}{d\tau_t} = -\frac{1 - a\delta\left((1-d) + (1-\beta)\lambda^{*s}\right)}{(1 - k^{*s}a)\left[\varepsilon - \delta\left(\varepsilon\lambda^{*s}\left((1-d) + (1-\beta)\lambda^{*s}\right) - \beta\omega\right)\right]}\frac{dp_t}{d\tau_t} < 0. \tag{18}$$

If an individual is naive, then μ_t^{*n} is determined by Equation (A.21). Following the same steps, we can show that the naif's consumption in period t changes according to

$$\frac{dx_t^n}{d\tau_t} = \frac{d\mu_t^{*n}}{d\tau_t} = -\frac{1 - a\frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)}{\left[\varepsilon - \frac{\beta\delta}{\hat{\beta}}\left(\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right) - \hat{\beta}\omega\right)\right]}\frac{dp_t}{d\tau_t} + ak^{*n}\frac{d\mu_{t+1}^{*s}(\hat{\beta})}{dp_{t+1}}\frac{dp_t}{d\tau_t} < 0, \tag{19}$$

where $\lambda^{*s}(\hat{\beta})$ and $\mu_{t+1}^{*s}(\hat{\beta})$ are the values of λ and μ, respectively, that a sophisticated individual with self-control problems $\hat{\beta}$ would follow.

3 Goverment Policy

In analyzing government policy, we follow O'Donoghue and Rabin (2003, 2006) and Gruber and Köszegi (2001) and assume that the social planner maximizes the long-run utility

of a long-lived representative consumer, which is given by the exponentially discounted sum of instantaneous utilities

$$W_t^i = \sum_{t=0}^{\infty} \delta^t u_t^i \tag{20}$$

subject to

$$\begin{aligned} w^i(0) &= w_0^i, \\ w_{t+1}^i &= x_t^i + (1-d)w_t^i, \\ x_t^i &= \lambda^{*i} w_t^i + \mu_t^{*i}(p_t), \\ I + \ell_t &= (1+\tau_t)x_t^i + (1-\sigma_t)z_t^i. \end{aligned}$$

However, we depart from the existing literature in two aspects. First, we do not constrain the tax rate to be constant as in Gruber and Köszegi (2001). Second, we assume that the tax revenues are not necessarily paid back to the individual in the form of a lump-sum transfer.

In the following, we consider three cases. First, we examine the most widely analyzed case in the existing literature of a tax on unhealthy food, which is returned to the consumer in lump-sum fashion. We will see that such a tax can correct over-or underweight resulting from self-control problems and naiveté, but cannot induce an individual to become healthy weight, if she is not healthy weight in the absence of the tax. Second, we will consider a tax that is used to subsidize the composite good z and show that it can achieve a healthy steady state. Third, we will analyze how the social planner can induce individuals to be healthy weight when it does not have information on the values of β and $\hat{\beta}$.

3.1 Case I

Consider the transfer $\ell_t = \tau_t x_t^i$ and subsidy $\sigma_t = 0$. In this case the optimization problem of the social planner is to

$$\max_{\{\tau_t\}_{t=0}^{\infty}} W_t^i = \sum_{t=0}^{\infty} \delta^t \left[v(\lambda^{*i} w_t^i + \mu_t^{*i}(p_t)) - c(w_t^i) + I - (\lambda^{*i} w_t^i + \mu_t^{*i}(p_t)) \right] \tag{20.1}$$

subject to

$$w^i(0) = w_0^i, w_{t+1}^i = (1 - d + \lambda^{*i}) w_t^i + \mu_t^{*i}(p_t).$$

Denoting the value function of the social planner in period t as $V^{SP}(w_t^i)$, the optimal tax is derived from the solution of the following Bellman equation:

$$V^{SP}(w_t^i) = \max_{\tau_t} \left\{ v(\lambda^{*i} w_t^i + \mu_t^{*i}(p_t)) - c(w_t^i) + I - (\lambda^{*i} w_t^i + \mu_t^{*i}(p_t)) + \delta V^{SP}(w_{t+1}^i) \right\}. \tag{21}$$

The first-order condition is given by

$$\left[v'(x_t^i) - 1 + \delta V^{SP\prime}(w_{t+1}^i) \right] \frac{d\mu_t^{*i}}{dp_t} \frac{dp_t}{d\tau_t} = 0. \tag{22}$$

Moreover, the value function evolves according to

$$V^{SP}(w_{t+1}^i) = v(\lambda^{*i} w_{t+1}^i + \mu_{t+1}^{*i}(p_{t+1})) - c(w_{t+1}^i) + I - (\lambda^{*i} w_{t+1}^i + \mu_{t+1}^{*i}(p_{t+1})) + \delta V^{SP}(w_{t+2}^i). \tag{23}$$

Differentiating both sides of the above equation with respect to w_{t+1}^i, we get

$$\begin{aligned} V^{SP\prime}(w_{t+1}^i) = &\left[v'(x_{t+1}^i) - 1 + \delta V^{SP\prime}(w_{t+2}^i) \right] \frac{d\mu_{t+1}^{*i}}{dp_{t+1}} \frac{dp_{t+1}}{d\tau_{t+1}} \frac{d\tau_{t+1}}{dw_{t+1}^i} \\ &+ \left(v'(x_{t+1}^i) - 1 \right) \lambda^{*i} - c'(w_{t+1}^i) + \delta V^{SP\prime}(w_{t+2}^i)(1 - d + \lambda^{*i}). \end{aligned} \tag{24}$$

The Euler equation of the government can be derived by rewriting the first-order condition (22) for period $t+1$ and plugging the resulting expression and Equation (22) in (24) to get

$$\left[v'(x_t^i) - 1 \right] = \delta \left[(v'(x_{t+1}^i) - 1)(1 - d) + c'(w_{t+1}^i) \right]. \tag{25}$$

Lastly, we note that both Euler equations of the consumer and the government need to be satisfied. Plugging Equation (10) in Equation (25), we can derive the equation, which determines the optimal stream of tax rates:

$$\delta(1 - d)\tau_{t+1} - \tau_t = \delta(1 - \beta) \left[\left(v'(x_{t+1}^i) - p_{t+1} \right) \lambda^{*i} - c'(w_{t+1}^i) \right] + \Delta_{t+1}^i, \tag{26}$$

where Δ_{t+1}^i determines the error that a type i consumer makes regarding period $t+1$ and is defined as

$$\begin{aligned} \Delta_{t+1}^i \equiv &\frac{\beta\delta}{\hat{\beta}} \left[(v'(x_{t+1}^s(\hat{\beta})) - p_{t+1})((1 - d) + (1 - \hat{\beta})\lambda^{*s}(\hat{\beta})) \right] \\ &- \delta \left[(v'(x_{t+1}^i) - p_{t+1})((1 - d) + (1 - \beta)\lambda^{*i}) \right] \begin{cases} = 0, \text{ if } i = s, \\ \neq 0, \text{ if } i = n. \end{cases} \end{aligned}$$

Equation (26) shows that there are two reasons for taxation in this case: the first term on the right-hand side corrects for the self-control problem of the representative individual, while the second term corrects for the naiveté of the consumer. The optimal tax may be either positive or negative.

However, in the absence of self-control problems ($\beta = \hat{\beta} = 1$) the right-hand side of (26) is zero and, therefore, the only solution for the optimal tax is $\tau_t = 0, \forall t$. The reason is that the consumer is rational and maximizes the same lifetime utility as the social planner. Therefore, there is no need for an intervention.

Moreover, Equations (1) and (26) form a system of two linear first-order difference equations in w_t^i, τ_t. The steady state of the system is derived by setting $w_t^i = \tilde{w}^i, \tau_t = \tilde{\tau}, \forall t$ and is given by

$$\tilde{w}^i = \frac{\tilde{\mu}^{*i}}{d - \lambda^{*i}} = \frac{\tilde{x}^i}{d}, \quad \tilde{\tau}(1 - \delta(1-d)) = -\delta(1-\beta)\left[(v'(\tilde{x}^i) - (1+\tilde{\tau}))\lambda^{*i} - c'(\tilde{w}^i)\right] - \widetilde{\Delta}^i, \tag{27}$$

where

$$\widetilde{\Delta}^i \equiv \frac{\beta\delta}{\hat{\beta}}\left[v'(\tilde{x}^s(\hat{\beta})) - \tilde{p}\right]\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right) - \delta\left[v'(\tilde{x}^i) - \tilde{p}\right]\left((1-d) + (1-\beta)\lambda^{*i}\right).$$

Note furthermore that, according to Proposition 2, the right-hand side of Equation (27) equals zero for a healthy weight individual. This means that in this case the left-hand side also equals zero, i.e. $\tilde{\tau} = 0$. Thus, healthy weight can only be achieved if in the absence of taxation the individual finds it optimal to be healthy weight, but cannot be achieved by the use of taxation. Furthermore, according to Equation (27), there does not exist any non-zero tax that is compatible with a healthy steady state, because $\tilde{\tau} \neq 0$ requires the right-hand side of (27) to be non-zero in steady state, which is incompatible with $\tilde{w}^i = w^H$. This means that the government cannot use this tax in order to induce individuals with abnormal weight to choose $\tilde{w}^i = w^H$.

The system of difference equations (1) and (26) can be solved for τ_t and w_t^i as functions of w_0^i during the transition to steady state. The system is stable, if it has at least one eigenvalue which is in the interval $]-1, 1[$. In Appendix D we show that a sufficient condition for saddle-path stability is $\delta \in]\underline{\delta}, 1]$, where $\underline{\delta}$ is a lower bound for δ, defined in Equation (D.10). Denoting the eigenvalues of the system as $\nu_1 \in]0, 1[, \nu_2 > 1$, we show in

Appendix D that the optimal trajectories are given by

$$w_t^i = \tilde{w}^i + (w_0^i - \tilde{w}^i)\nu_1^t, \tag{28a}$$

$$\tau_t = \tilde{\tau} - \frac{a_{11} - \nu_1}{a_{12}}(w_0^i - \tilde{w}^i)\nu_1^t, \tag{28b}$$

where a_{11}, a_{12} are constants defined in Equations (D.5a) and (D.5b) Appendix D.

We summarize the solution to the optimal tax rate in the following Proposition:

Proposition 4. *When the tax revenues are returned to the consumer in lump-sum fashion, the optimal steady state tax is positive if the individual is overweight, negative if the individual is underweight and zero if (i) the individual has no self-control problems or (ii) the individual has a healthy weight in the absence of taxation. Moreover, there does not exist a non-zero steady state tax compatible with healthy weight.*

A sufficient condition for the system of difference equations in w_t^i and τ_t to be saddle-path stable is $\delta \in]\underline{\delta}, 1[$. Its solution is given by Equations (28a), (28b).

Proof: See Appendix D.

We note that in the first case of a zero tax in Proposition 2, the individual may be rationally overweight or underweight. Since this tax is designed to correct self-control problems and naiveté, it cannot correct for rationally abnormal weight. In the next subsection, we consider a tax which is designed to correct abnormal weight.

3.2 Case II

In this case we set $\ell_t = 0$. We also assume that the composite good is subsidized at a rate σ_t, which is proportional to τ_t. Note that if the system is balanced in each period, i.e. if the tax revenues $\tau_t x_t^i$ equal the expenditures $\sigma_t z_t^i$, then the budget constraint collapses to the budget constraint in Case I. In order to consider policy, which is different from Case I, we assume that the subsidy σ_t is defined in a constant relation to τ_t in such a way that the tax system is balanced in the steady state, but not necessarily balanced during the transition period. Thus, we assume that $\sigma_t \tilde{z}^i = \tau_t \tilde{x}^i$, or

$$\sigma_t = \tau_t \frac{\tilde{x}^i}{\tilde{z}^i}. \tag{29}$$

The system will generate a surplus (deficit) in period t, if $\tau_t(x_t^i/z_t^i - \tilde{x}^i/\tilde{z}^i) > (<)0$. We assume that any surplus is spent on goods and services not directly benefiting the representative consumer and if there is a deficit, it is financed by other taxes that are not levied on the consumer. The term $\tilde{x}^i/\tilde{z}^i$ can be interpreted as the social planner's desired unhealthy food consumption as a proportion of the total expenditures.

In this case, the problem of the social planner becomes to

$$\max_{\{\tau_t\}_{t=0}^{\infty}} W_t^i = \sum_{t=0}^{\infty} \delta^t \left[v(\lambda^{*i} w_t^i + \mu_t^{*i}(p_t)) - c(w_t^i) + \frac{I - (1+\tau_t)(\lambda^{*i} w_t^i + \mu_t^{*i}(p_t))}{1-\sigma_t} \right]. \quad (20.2)$$

The solution procedure is the same as in Case I and results in the following Euler equation of the government (analogue to Equation (25) in the previous subsection):

$$\left[v'(x_t^i) - p_t\right] = \delta\left[(v'(x_{t+1}^i) - p_{t+1})(1-d) + c'(w_{t+1}^i)\right] + \frac{x_t^i - z_t^i \tilde{x}^i/\tilde{z}^i}{\frac{d\mu_t^{*i}}{dp_t}\frac{dp_t}{d\tau_t}(1-\sigma_t)} - \delta(1-d+\lambda^{*i}) \frac{x_{t+1}^i - z_{t+1}^i \tilde{x}^i/\tilde{z}^i}{\frac{d\mu_{t+1}^{*i}}{dp_{t+1}}\frac{dp_{t+1}}{d\tau_{t+1}}(1-\sigma_{t+1})}. \quad (30)$$

In order to simplify the above equation, we note that $d\mu_t^{*i}/dp_t = d\mu_{t+1}^{*i}/dp_{t+1} = d\mu^{*i}/dp$ according to Equations (18) and (19). Moreover, from the definition of σ_t and the budget constraint of the individual, we can simplify the last two terms in Equation (30) in the following way:

$$\frac{x_t^i - z_t^i \tilde{x}^i/\tilde{z}^i}{\frac{d\mu_t^{*i}}{dp_t}\frac{dp_t}{d\tau_t}(1-\sigma_t)} = \frac{x_t^i - \frac{I-(1+\tau_t)x_t^i}{(1-\sigma_t)}\tilde{x}^i/\tilde{z}^i}{\frac{d\mu^{*i}}{dp}\left(1 + \frac{1+\tau_t}{1-\sigma_t}\frac{\tilde{x}^i}{\tilde{z}^i}\right)} = \frac{x_t^i\left(1+\frac{\tilde{x}^i}{\tilde{z}^i}\right) - \tilde{x}^i\left(1+\frac{\tilde{x}^i}{\tilde{z}^i}\right)}{\frac{d\mu^{*i}}{dp}\left(1+\frac{\tilde{x}^i}{\tilde{z}^i}\right)} = \frac{(x_t^i - \tilde{x}^i)}{\frac{d\mu^{*i}}{dp}},$$

where in the second equality we used $I = \tilde{x}^i + \tilde{z}^i$. Using the above result and the Euler equation of the consumer, which is given by (10), we can derive the following equation, determining the optimal path of the tax policy:

$$\frac{\left[(x_t^i - \tilde{x}^i) - \delta(1-d+\lambda^{*i})(x_{t+1}^i - \tilde{x}^i)\right]}{\frac{d\mu^{*i}}{dp}} = \delta(1-\beta)\left[(v'(x_{t+1}^i) - p_{t+1})\lambda^{*i} - c'(w_{t+1}^i)\right] + \Delta_{t+1}^i, \quad (31)$$

where Δ_{t+1}^i is defined analogously to the same term in Case I.

Now we can discuss the difference to the previous subsection. Note that Equations (1) and (31) define a system of linear first-order difference equations in w_t^i, p_t, while in the previous case the endogenous variables were w_t^i, τ_t. Therefore, it is more convenient to present the optimal tax policy in terms of the optimal choice of the relative price p_t and

not τ_t. The system of difference equations does not necessarily converge and saddle-path stability has to be assumed. If there is convergence, then the solution to the system is given by

$$w_t^i = \tilde{w}^i + (w_0^i - \tilde{w}^i)\overline{\nu}_1^t, \tag{32a}$$

$$p_t = \tilde{p} - \frac{\overline{a}_{11} - \overline{\nu}_1}{\overline{a}_{12}}(w_0^i - \tilde{w}^i)\overline{\nu}_1^t, \tag{32b}$$

where $\overline{a}_{11}, \overline{a}_{12}$ are constants defined in Equations (E.7a), (E.7b) and $\overline{\nu}_1$ is an eigenvalue defined in Equation (E.12) in Appendix E.

We can prove the following results:

Proposition 5. *The steady state relative price is given by* $\tilde{p} = v'(\tilde{x}^i) = \gamma - \varepsilon dw^H$, *where* $\tilde{x}^i = x^F = x^H = dw^H$, *and is independent of the degree of self-control problems* β *and the degree of naiveté* $\hat{\beta} - \beta$. *In the steady state, the consumer is healthy-weight* $\tilde{w}^i = w^H$ *and does not make errors in her expected consumption choices, i.e.* $\widetilde{\Delta}^i = 0, \forall i$.

If the system of difference equations in (w_t, p_t) *is characterized by a saddle-path, then its solution is given by Equations* (32a), (32b). *Otherwise, the system is unstable.*

Proof: See Appendix E.

Using Proposition 5, one can derive an explicit solution for the optimal steady state tax:

$$\tilde{\tau} = \frac{((\gamma - \varepsilon dw^H) - 1)}{(\gamma - \varepsilon dw^H)\frac{dw^H}{I - dw^H} + 1}. \tag{33}$$

Thus, the steady state tax depends only on (i) the instantaneous utility of the consumer, (ii) the consumer's income and (iii) the calorie expenditure when the consumer is healthy weight. The intuition behind this result is the following. When the individual has a healthy weight, then a marginal weight change does not affect her health. The only possiblity to maintain this weight is if the instantaneous utility is maximized in each period, i.e. if the individual consumes until satiation, because in this case a perturbation of the optimal consumption path in period t, which is undone in period $t+1$ also has a zero impact on utility. Therefore, the price that ensures $\tilde{x}^i = x^F$ also induces the individual to be healthy weight.

Nevertheless, according to Equations (32a) and (32b), the optimal tax τ_t does depend upon β and $\hat{\beta}$ during transition if the system converges. If the system is unstable, then

the only solution to the social planner's problem may be to set the price of unhealthy food equal to $\tilde{p}$ in each period. In the next subsection we show that this policy guarantees convergence to the same steady state weight $\tilde{w}^i = w^H$. A second reason for considering this policy is that it does not require information on the values of β and $\hat{\beta}$.

3.3 Case III: second-best policy

Suppose that the government cannot implement the policy from Case II either because the steady state is unstable or because the government does not know whether the representative individual makes errors and behaves time-inconsistently. Assume furthermore that it levies a constant tax rate on unhealthy food $\tau_t = \bar{\tau}, \forall t$. The tax revenues may either be returned each period to the consumer in the form of a lump-sum transfer or used to subsidize the bundle of other goods. Then we can prove the following result:

Proposition 6. *Suppose that the government levies a constant tax rate $\tau_t = \bar{\tau}$, such that the relative price of unhealthy food equals the steady state price from Case II, i.e. $p_t = \bar{p} = \gamma - \varepsilon d w^H$. Then the consumer unambiguously achieves a steady state of healthy weight $\tilde{w}^i = w^H, \forall i$ and healthy consumption $\tilde{x}^i = x^H, \forall i$ irrespective of her degree of self-control problems and naiveté.*

Proof: The proof consists of two parts. First, we prove that the steady state is indeed $\tilde{w}^i = w^H, \forall i$. Second, we show that the steady state is stable, i.e. the weight w_t^i converges to w^H.

Note that under a constant tax rate, μ_t^{*i} and $\mu_t^{*s}(\hat{\beta})$ become constant and equal their steady state values in each period t. Furthermore, we showed in Appendix A that $\tilde{\mu}^{*i} = \tilde{w}^i(d - \lambda^{*i})$ and $\tilde{\mu}^{*s}(\hat{\beta}) = \tilde{w}^s(\hat{\beta})(d - \lambda^{*s}(\hat{\beta}))$. Thus, consumption in period t is given by $x_t^i = \lambda^{*i} w_t^i + \tilde{\mu}^{*i} = d\tilde{w}^i + \lambda^{*i}(w_t^i - \tilde{w}^i)$ and expected consumption in period $t+1$ from the viewpoint of period t is $x_{t+1}^s(\hat{\beta}) = \lambda^{*s}(\hat{\beta}) w_{t+1}^i + \tilde{\mu}^{*s}(\hat{\beta}) = d\tilde{w}^s(\hat{\beta}) + \lambda^{*s}(\hat{\beta})(w_{t+1}^i - \tilde{w}^s(\hat{\beta}))$. We can use these expressions and $p_t = \bar{p} = \gamma - \varepsilon d w^H$ in the Euler equation of the consumer (10) and rewrite it in the following way:

$$\begin{aligned}\varepsilon\lambda^{*i}(w_t^i - \tilde{w}^i) = {} & \frac{\beta\delta}{\hat{\beta}}\varepsilon\lambda^{*s}(\hat{\beta})((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}))(w_{t+1}^i - \tilde{w}^s(\hat{\beta})) - \beta\delta\omega(w_{t+1}^i - w^H) \\ & - \varepsilon d(\tilde{w}^i - w^H) + \frac{\beta\delta}{\hat{\beta}}\varepsilon d(\tilde{w}^s(\hat{\beta}) - w^H)((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})). \qquad (34)\end{aligned}$$

Now we can derive the steady state. Consider first the case $i = s$. In a steady state $w_t^s = w_{t+1}^s = \tilde{w}^s$ and Equation (34) collapses to

$$(\tilde{w}^s - w^H)[\beta\delta\omega + \varepsilon d(1 - \delta((1-d) + (1-\beta)\lambda^{*s}))] = 0.$$

The solution of the above equation is $\tilde{w}^s = w^H$. Now we can consider the second case $i = n$. If we plug $\tilde{w}^s(\hat{\beta}) = w^H$ in (34), evaluate at steady state and simplify, we get

$$(\tilde{w}^n - w^H)\left(\varepsilon d + \beta\delta\omega - \frac{\beta\delta}{\hat{\beta}}\varepsilon\lambda^{*s}(\hat{\beta})((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}))\right) = 0.$$

The solution to the above equation is $\tilde{w}^n = w^H$. Hence, we have proven that the steady state is $\tilde{w}^i = w^H, i = s, n$ and what is left is to show that the weight converges to this value irrespective of the starting weight. Plugging $\tilde{w}^i = w^H$ in Equation (34) and simplifying, we get

$$(w_{t+1}^i - w^H) = \frac{-\varepsilon\lambda^{*i}}{\frac{\beta\delta}{\hat{\beta}}\left[-\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right) + \hat{\beta}\omega\right]}(w_t^i - w^H).$$

Using the definition of λ^{*i} from Equation (A.11), we can show that the term in front of $(w_t^i - w^H)$ is in the interval $]0, 1[$. Thus, the weight of the individual converges to $\tilde{w}^i = w^H$ and the consumption of junk-food converges to $\tilde{x}^i = x^H$. Q.E.D.

The policy considered in Proposition 6 is only a second-best instrument which can be used in the absence of information on the errors that consumers make. Even though it does not achieve the first-best, as it does not optimize the utility of the representative consumer during the transition to the steady state, it does achieve the health-maximizing steady state and the utility-maximizing steady state from Case II. Moreover, it is not paternalistic, since it does not seek to correct any errors that individuals might make. The reason why the health-maximizing steady state is reached by this policy is that individuals do not make errors in the steady state. Hence, no paternalistic policy is required.

4 Conclusion

This paper has merged the optimal paternalism and rational obesity literatures in order to study how rational and irrational individuals differ in their consumption choices

and weight. Our first contribution is in showing that irrationality in the form of self-control problems and naive expectations regarding future consumption cannot induce an individual to have a steady state weight below or above the healthy weight.

This result makes it possible for policy makers to construct a non-paternalistic tax that incentivizes both rational and irrational individuals to have a healthy weight. While this policy is only health-maximizing, but not utility-maximizing, it has two advantages relative to the paternalistic policy. First, it is easy to implement, as it does not require information on the type and degree of irrationality that consumers exhibit. Second, the paternalistic tax cannot induce obese individuals to have a healthy weight. While the paternalistic tax may be utility-maximizing, the externality that obesity causes on the rest of society through its medical treatment costs may make the health-maximizing policy more desirable to a social planner.

A Proof of Proposition 1

The derivation of the closed-form solution for junk-food consumption follows closely the analysis of Gruber and Köszegi (2001). First, we insert the assumed functional forms for $v(x)$ and $c(w)$ in Equation (10):

$$\gamma - \varepsilon x_t^i - p_t = \frac{\beta\delta}{\hat{\beta}} \left[\left(\gamma - \varepsilon x_{t+1}^s(\hat{\beta}) - p_{t+1} \right) \left((1-d) + (1-\hat{\beta}) \frac{\partial x_{t+1}^s(\hat{\beta})}{\partial w_{t+1}^i} \right) + \hat{\beta}\omega(w_{t+1}^i - w^H) \right]. \tag{A.1}$$

We solve the above equation by the method of undetermined coefficients. Assume that $x_t^i = \lambda_t^i w_t^i + \mu_t^i$, where λ_t^i and μ_t^i are constants to be determined. In this case $\partial x_t^i / \partial w_t^i = \lambda_t^i$. Moreover, Equation (1) changes to $w_{t+1}^i = (1-d)w_t^i + \lambda_t^i w_t^i + \mu_t^i$. Inserting the terms for $x_t^i, \partial x_t^i / \partial w_t^i$ and w_{t+1}^i in (A.1), we get

$$\gamma - \varepsilon(\lambda_t^i w_t^i + \mu_t^i) - p_t = \frac{\beta\delta}{\hat{\beta}} \left[\left(\gamma - \varepsilon[\lambda_{t+1}^s(\hat{\beta})((1-d+\lambda_t^i)w_t^i + \mu_t^i) + \mu_{t+1}^s(\hat{\beta})] - p_{t+1} \right) \left((1-d) + (1-\hat{\beta})\lambda_{t+1}^s(\hat{\beta}) \right) + \hat{\beta}\omega[(1-d+\lambda_t^i)w_t^i + \mu_t^i) - w^H] \right]. \tag{A.2}$$

In order to determine λ_t^i, we have to equate the terms in front of w_t^i on the left- and the right-hand side of (A.2). This leads to the following equation:

$$-\varepsilon\lambda_t^i = \frac{\beta\delta}{\hat{\beta}} \left[-\varepsilon\lambda_{t+1}^s(\hat{\beta})(1-d+\lambda_t^i) \left((1-d) + (1-\hat{\beta})\lambda_{t+1}^s(\hat{\beta}) \right) + \hat{\beta}\omega(1-d+\lambda_t^i) \right]. \tag{A.3}$$

The above expression can be simplified to give:

$$\lambda_t^i = -(1-d) + \frac{\varepsilon(1-d)}{\varepsilon - \frac{\beta\delta}{\hat{\beta}} \left[\varepsilon\lambda_{t+1}^s(\hat{\beta}) \left((1-d) + (1-\hat{\beta})\lambda_{t+1}^s(\hat{\beta}) \right) - \hat{\beta}\omega \right]}. \tag{A.4}$$

In the above equation $\lambda_{t+1}^s(\hat{\beta})$ is the constant proportion of weight that a sophisticated individual with $\beta = \hat{\beta}$ consumes, as the self in period t expects to be a sophisticate with present-bias $\hat{\beta}$ from time $t+1$ onwards. Thus, in order to show that λ_t^i is constant, we first need to show that the $\lambda_{t+1}^s(\hat{\beta})$ converges to a constant. Thus, we derive first the solution of λ_t^s when $\beta = \hat{\beta}$:

$$\lambda_t^s = -(1-d) + \frac{\varepsilon(1-d)}{\varepsilon - \delta \left[\varepsilon\lambda_{t+1}^s \left((1-d) + (1-\beta)\lambda_{t+1}^s \right) - \beta\omega \right]}. \tag{A.5}$$

The above equation represents backward recursion in λ_t^s. In order to prove that it converges to a constant value λ^{*s}, we define the right-hand side of (A.5) as $f_s(\lambda_{t+1}^s)$. Note that the second term of $f_s(\cdot)$ is the reciprocal of a quadratic equation in λ_{t+1}^s with a negative coefficient in front of λ_{t+1}^{s2}. Thus, if it is positive for two λ_{t+1}^s values, it is also positive for all values in between. We will show that it is positive for $\lambda_{t+1}^s = -(1-d)$ and $\lambda_{t+1}^s = 0$ and that λ_t^s is bounded by these values, i.e. (i) $f_s(-(1-d)) > -(1-d)$ and (ii) $f_s(0) < 0$. First, we have:

$$f_s(-(1-d)) = -(1-d)\left[1 - \frac{\varepsilon}{\varepsilon + \delta\left[\varepsilon(1-d)^2\beta + \beta\omega\right]}\right] > -(1-d). \tag{A.6}$$

Second:

$$f_s(0) = -(1-d)\left[1 - \frac{\varepsilon}{\varepsilon + \delta\beta\omega}\right] < 0. \tag{A.7}$$

Equations (A.6) and (A.7) prove that λ_t^s is bounded in the interval $]-(1-d), 0[$. In order to prove convergence, it is sufficient to show that $f_s(\lambda_{t+1}^s)$ is continuous and monotonically increasing in λ_{t+1}^s on the interval $]-(1-d), 0[$. The derivative of f_s with respect to λ_{t+1}^s is given by

$$f_s'(\lambda_{t+1}^s) = \frac{\delta\varepsilon^2(1-d)\left[2(1-\beta)\lambda_{t+1}^s + (1-d)\right]}{\left[\varepsilon - \delta\left(\varepsilon\lambda_{t+1}^s\left((1-d) + (1-\beta)\lambda_{t+1}^s\right) - \beta\omega\right)\right]^2}. \tag{A.8}$$

Next, note that $f_s(\cdot)$ is strictly convex in the interval $\lambda_{t+1}^s \in]-(1-d), 0[$, i.e. $f_s''(\cdot) > 0$ in the interval of interest. Thus, $f_s'(\cdot) > 0$ for all values of λ_{t+1}^s if $f_s'(\lambda_{t+1}^s = -(1-d)) \geq 0$, i.e. if

$$f_s'(-(1-d)) = \frac{\delta\varepsilon^2(1-d)^2\left[2\beta - 1\right]}{\left[\varepsilon - \delta\left(\varepsilon\lambda_{t+1}^s\left((1-d) + (1-\beta)\lambda_{t+1}^s\right) - \beta\omega\right)\right]^2} \geq 0. \tag{A.9}$$

The above inequality holds if $\beta \geq 1/2$. Therefore, if there are infinitely many periods till the end of the time horizon, λ_t^s converges to the unique solution of

$$\lambda^{*s} = -(1-d)\left[1 - \frac{\varepsilon}{\varepsilon - \delta\left(\varepsilon\lambda^{*s}\left((1-d) + (1-\beta)\lambda^{*s}\right) - \beta\omega\right)}\right] \in]-(1-d), 0[. \tag{A.10}$$

On the other hand, a naive consumer, who makes a decision in period t, expects to behave as a sophisticated consumer with present-bias $\hat{\beta}$ from $t+1$ onwards. Thus, she expects to consume according to $\lambda^{*s}(\hat{\beta})$ from period $t+1$ onwards, and chooses the consumption

in period t according to

$$\lambda^{*n} = -(1-d)\left[1 - \frac{\varepsilon}{\varepsilon - \frac{\beta\delta}{\hat{\beta}}\left(\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right) - \hat{\beta}\omega\right)}\right] \in]-(1-d), 0[. \tag{A.11}$$

Equations (A.10) and (A.11) show that λ^{*i} is independent of the price level.

In order to determine the parameter μ_t^i, we equate the terms on the left-and right-hand sides of Equation (A.2), which are not multiplicative of w_t^i:

$$\begin{aligned}\gamma - \varepsilon\mu_t^i - p_t = \frac{\beta\delta}{\hat{\beta}}\Big[&\left(\gamma - \varepsilon[\lambda_{t+1}^s(\hat{\beta})\mu_t^i + \mu_{t+1}^s(\hat{\beta})] - p_{t+1}\right)\left((1-d)+(1-\hat{\beta})\lambda_{t+1}^s(\hat{\beta})\right) \\ &+\hat{\beta}\omega[\mu_t^i - w^H]\Big].\end{aligned} \tag{A.12}$$

The above equation can be solved for μ_t^i:

$$\begin{aligned}\mu_t^i \;=\; &\frac{\gamma - \frac{\beta\delta}{\hat{\beta}}\left[\gamma\left((1-d)+(1-\hat{\beta})\lambda_{t+1}^s(\hat{\beta})\right) - \hat{\beta}\omega w^H\right]}{\varepsilon - \frac{\beta\delta}{\hat{\beta}}\left[\varepsilon\lambda_{t+1}^s(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda_{t+1}^s(\hat{\beta})\right) - \hat{\beta}\omega\right]} \\ &- \frac{p_t - p_{t+1}\frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda_{t+1}^s(\hat{\beta})\right)}{\varepsilon - \frac{\beta\delta}{\hat{\beta}}\left[\varepsilon\lambda_{t+1}^s(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda_{t+1}^s(\hat{\beta})\right) - \hat{\beta}\omega\right]} \\ &+ \frac{\varepsilon\frac{\beta\delta}{\hat{\beta}}\left[(1-d)+(1-\hat{\beta})\lambda_{t+1}^s(\hat{\beta})\right]}{\varepsilon - \frac{\beta\delta}{\hat{\beta}}\left[\varepsilon\lambda_{t+1}^s(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda_{t+1}^s(\hat{\beta})\right) - \hat{\beta}\omega\right]}\mu_{t+1}^s(\hat{\beta}).\end{aligned} \tag{A.13}$$

Note that the value of μ_t^i converges if the term in front of $\mu_{t+1}^s(\hat{\beta})$ is between zero and one, because $\lambda_{t+1}^s(\hat{\beta})$ converges to $\lambda^{*s}(\hat{\beta})$ and the price level also converges to a given $\tilde{p}$. We denote the limit value of the term of interest as k^{*i} and using Equation (A.11) show that it is always in the interval $]0, 1[$:

$$\begin{aligned}k^{*i} \;&\equiv\; \frac{\varepsilon\frac{\beta\delta}{\hat{\beta}}\left[(1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right]}{\varepsilon - \frac{\beta\delta}{\hat{\beta}}\left[\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right) - \hat{\beta}\omega\right]} \\ &=\; \frac{\frac{\beta\delta}{\hat{\beta}}\left[(1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right](\lambda^{*n}+(1-d))}{1-d} \in]0, 1[.\end{aligned} \tag{A.14}$$

In order to solve for $\mu_t^{*i}(\hat{\beta})$, we solve first for a sophisticated consumer's μ_t^{*s}, which is given by the solution of (A.13) when $\beta = \hat{\beta}$, i.e. the solution of

$$\mu_t^s \;=\; \frac{\gamma - \delta\left[\gamma\left((1-d)+(1-\beta)\lambda^{*s}\right) - \beta\omega w^H\right]}{\varepsilon - \delta\left[\varepsilon\lambda^{*s}\left((1-d)+(1-\beta)\lambda^{*s}\right) - \beta\omega\right]} - \frac{p_t - p_{t+1}\delta\left((1-d)+(1-\beta)\lambda^{*s}\right)}{\varepsilon - \delta\left[\varepsilon\lambda^{*s}\left((1-d)+(1-\beta)\lambda^{*s}\right) - \beta\omega\right]}$$

$$+\frac{\varepsilon\delta\left[(1-d)+(1-\beta)\lambda^{*s})\right]}{\varepsilon-\delta\left[\varepsilon\lambda^{*s}\left((1-d)+(1-\beta)\lambda^{*s}\right)-\beta\omega\right]}\mu_{t+1}^{s}. \tag{A.15}$$

Denoting the term in front of μ_{t+1}^{s} as k^{*s} and noting that Equation (A.15) is a simple geometric progression, we get

$$\begin{aligned}\mu_t^{*s} = &\frac{\gamma-\delta\left[\gamma\left((1-d)+(1-\beta)\lambda^{*s}\right)-\beta\omega w^H\right]}{(1-k^{*s})\left[\varepsilon-\delta\left(\varepsilon\lambda^{*s}\left((1-d)+(1-\beta)\lambda^{*s}\right)-\beta\omega\right)\right]}\\ &-\frac{p_t-\left[\delta\left((1-d)+(1-\beta)\lambda^{*s}\right)-k^{*s}\right]\sum_{i=1}^{\infty}k^{*s\,i-1}p_{t+i}}{\varepsilon-\delta\left(\varepsilon\lambda^{*s}\left((1-d)+(1-\beta)\lambda^{*s}\right)-\beta\omega\right)}.\end{aligned} \tag{A.16}$$

Furthermore, if the tax rate and, hence, the relative price reach their steady state values, μ_t^{*s} also reaches a steady state, defined by

$$\tilde{\mu}^{*s}=\frac{(\gamma-\tilde{p})\left[1-\delta\left((1-d)+(1-\beta)\lambda^{*s}\right)\right]+\delta\beta\omega w^H}{(1-k^{*s})\left[\varepsilon-\delta\left(\varepsilon\lambda^{*s}\left((1-d)+(1-\beta)\lambda^{*s}\right)-\beta\omega\right)\right]}. \tag{A.17}$$

Note that $\tilde{\mu}^{*s}$ can be simplified significantly by solving Equation (A.2) for a sophisticated individual in a steady state, which gives

$$(\gamma-\tilde{p})\left[1-\delta\left((1-d)+(1-\beta)\lambda^{*s}\right)\right]+\delta\beta\omega w^H=\varepsilon(\lambda^{*s}\tilde{w}^s+\tilde{\mu}^{*s})\left[1-\delta\left((1-d)+(1-\beta)\lambda^{*s}\right)\right]+\beta\delta\omega\tilde{w}^s.$$

Plugging the above Equation in (A.17) and simplifying, we get

$$\tilde{\mu}^{*s}=\tilde{w}^s(d-\lambda^{*s}). \tag{A.18}$$

Lastly, we can rewrite the solution of μ_t^{*s} in a more convenient form by expressing it as a function of $p_t-\tilde{p}$. Using Equations (A.16) and (A.17), we get

$$\mu_t^{*s}=\tilde{\mu}^{*s}-\frac{(p_t-\tilde{p})-\left[\delta\left((1-d)+(1-\beta)\lambda^{*s}\right)-k^{*s}\right]\sum_{i=1}^{\infty}k^{*si-1}(p_{t+i}-\tilde{p})}{\varepsilon-\delta\left(\varepsilon\lambda^{*s}\left((1-d)+(1-\beta)\lambda^{*s}\right)-\beta\omega\right)}. \tag{A.19}$$

The solution for naive consumers can be easily derived from Equations (A.13) and (A.16):

$$\begin{aligned}\mu_t^{*n} = &\frac{\gamma-\frac{\beta\delta}{\hat{\beta}}\left[\gamma\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)-\hat{\beta}\omega w^H\right]}{\varepsilon-\frac{\beta\delta}{\hat{\beta}}\left[\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)-\hat{\beta}\omega\right]}\\ &-\frac{p_t-p_{t+1}\frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)}{\varepsilon-\frac{\beta\delta}{\hat{\beta}}\left[\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)-\hat{\beta}\omega\right]}+k^{*n}\mu_{t+1}^{*s}(\hat{\beta}).\end{aligned} \tag{A.20}$$

Analogous to μ_t^{*s}, we can express μ_t^{*n} more conveniently as a function of $p_t - \tilde{p}$. Following the same steps as in the derivation of Equation (A.19), we get

$$\mu_t^{*n} = \tilde{\mu}^{*n} - \frac{(p_t - \tilde{p}) - \frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)(p_{t+1} - \tilde{p})}{\varepsilon - \frac{\beta\delta}{\hat{\beta}}\left[\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right) - \hat{\beta}\omega\right]} + k^{*n}\left(\mu_{t+1}^{*s}(\hat{\beta}) - \tilde{\mu}^{*s}(\hat{\beta})\right), \tag{A.21}$$

where the steady state values $\tilde{\mu}^{*n}$ and $\tilde{\mu}^{*s}(\hat{\beta})$ are defined in the following way:

$$\tilde{\mu}^{*n} = \frac{(\gamma - \tilde{p})\left[1 - \frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right] + \beta\delta\omega w^H}{\varepsilon - \frac{\beta\delta}{\hat{\beta}}\left[\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right) - \hat{\beta}\omega\right]} + k^{*n}\tilde{\mu}^{*s}(\hat{\beta}) = \tilde{w}^n(d - \lambda^{*n}), \tag{A.22}$$

$$\tilde{\mu}^{*s}(\hat{\beta}) = \tilde{w}^s(\hat{\beta})(d - \lambda^{*s}(\hat{\beta})). \tag{A.23}$$

Note that in deriving the second equality in Equation (A.22) we replaced the numerator in the first term in Equation (A.22) by its steady state value as derived by evaluating Equation (A.2) in a steady state. On the other hand, Equation (A.23) is derived in the same way as Equation (A.18) with the difference that it applies to a sophisticated individual with present-bias $\hat{\beta}$. Q.E.D.

B Proof of Proposition 2

First, we prove the first part of Proposition 2, which states that there are three possible steady states: (a) $x^H < \tilde{x}^i < x^F$, (b) $x^F < \tilde{x}^i < x^H$, (c) $x^H = \tilde{x}^i = x^F$. Using Equations (A.18), (A.22) and (A.23) from Appendix A, we can rewrite the steady state, described by Equations (11) and (12), in the following way:

$$\begin{aligned}\beta\delta\omega(\tilde{w}^i - w^H) = \left(\gamma - \varepsilon d\tilde{w}^i - \tilde{p}\right)\left[1 - \frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right] \\ + (\tilde{w}^s(\hat{\beta}) - \tilde{w}^i)(d - \lambda^{*s}(\hat{\beta}))\varepsilon\frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right).\end{aligned} \tag{B.1}$$

We will consider separately the two types of individuals. Assume first that the individual is sophisticated. In this case Equation (B.1) simplifies to

$$\beta\delta\omega(\tilde{w}^s - w^H) = (\gamma - \varepsilon d\tilde{w}^s - \tilde{p})\left[1 - \delta\left((1-d) + (1-\beta)\lambda^{*s}\right)\right] \tag{B.2}$$

Noting that the term in brackets on the right-hand side of (B.2) is positive and using the definitions of x^H and x^F from Equation (14), the proof for a sophisticated individual follows immediately.

We turn now to the naive individual. In this case the steady state is determined by

$$\begin{aligned}\beta\delta\omega(\tilde{w}^n - w^H) = (\gamma - \varepsilon d\tilde{w}^n - \tilde{p})\left[1 - \frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right] \\ + (\tilde{w}^s(\hat{\beta}) - \tilde{w}^n)(d - \lambda^{*s}(\hat{\beta}))\varepsilon\frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right),\end{aligned} \tag{B.3}$$

where $\tilde{w}^s(\hat{\beta})$ is determined by Equation (B.2) when $\beta = \hat{\beta}$. One can use Equations (B.2) and (B.3) in order to prove the following intermediate result:

$$\tilde{w}^n \gtreqless \tilde{w}^s(\hat{\beta}) \quad \Leftrightarrow \quad \tilde{w}^s(\hat{\beta}) \gtreqless w^H. \tag{B.4}$$

Assume first that $\tilde{w}^s(\hat{\beta}) = w^H$. Inserting this in Equation (B.2) gives $\gamma - \tilde{p} = \varepsilon d\tilde{w}^s(\hat{\beta}) = \varepsilon d w^H$. Inserting this equality in (B.3) and simplifying, we get

$$(\tilde{w}^n - w^H)(\beta\delta\omega + \varepsilon d - \varepsilon\lambda^{*s}(\hat{\beta})\beta\delta(1 - d + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}))/\hat{\beta}) = 0.$$

This equation can only be satisfied if $\tilde{w}^n = w^H$.

There are two remaining cases in Equation (B.4): $\tilde{w}^s(\hat{\beta}) > (<) w^H$. We use proof by contradiction to show that (B.4) holds in these cases. To do so, evaluate Equation (B.2) for $\beta = \hat{\beta}$, insert it in Equation (B.3) and rewrite the resulting expression in the following way:

$$\begin{aligned}\beta\delta\omega(\tilde{w}^n - w^H) - \hat{\beta}\delta\omega(\tilde{w}^s(\hat{\beta}) - w^H) = (\gamma - \varepsilon d\tilde{w}^n - \tilde{p})\left[1 - \frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right] \\ - \left(\gamma - \varepsilon d\tilde{w}^s(\hat{\beta}) - \tilde{p}\right)\left[1 - \delta\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right] \\ + (\tilde{w}^s(\hat{\beta}) - \tilde{w}^n)(d - \lambda^{*s}(\hat{\beta}))\varepsilon\frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right].\end{aligned} \tag{B.5}$$

Consider now the first case: $\tilde{w}^s(\hat{\beta}) > w^H$ and assume that $\tilde{w}^n \leq \tilde{w}^s(\hat{\beta})$. It is trivial to show that in this case the assumption $\tilde{w}^n \leq \tilde{w}^s(\hat{\beta})$ makes the left-hand side of Equation (B.5) negative and its right-hand side positive. This is a contradiction and we conclude that if $\tilde{w}^s(\hat{\beta}) > w^H$, then it must hold true that $\tilde{w}^n > \tilde{w}^s(\hat{\beta})$.

In the second case $\tilde{w}^s(\hat{\beta}) < w^H$. Assume now that $\tilde{w}^n \geq \tilde{w}^s(\hat{\beta})$. It is again trivial to show that this assumption makes the left-hand side of (B.5) positive and its right-hand side negative. This is a contradiction and we conclude that if $\tilde{w}^s(\hat{\beta}) < w^H$, then it must hold true that $\tilde{w}^n < \tilde{w}^s(\hat{\beta})$. Thus, Equation (B.4) is always true.

Note now that Equations (B.3) and (B.4) together determine the following relations:

$$\operatorname{sgn}\{\tilde{w}^n - w^H\} = \operatorname{sgn}\{\gamma - \varepsilon d\tilde{w}^n - \tilde{p}\} = \operatorname{sgn}\{\tilde{w}^s(\hat{\beta}) - w^H\}. \tag{B.6}$$

The first equality in Equation (B.6) proves the results from the first part of Proposition 2 in the case of a naive consumer.

Consider now the second part of Proposition 2. Suppose that $x^F > x^H$. This implies $\gamma - \varepsilon\tilde{x}^i - \tilde{p} > 0$ for $\tilde{x}^i \leq x^H$. Assume that the individual achieves a healthy or underweight steady state, i.e. $\tilde{w}^i \leq w^H$ and $\tilde{x}^i = d\tilde{w}^i \leq x^H$. In this case Equations (B.4) implies $\tilde{w}^i \leq \tilde{w}^s(\hat{\beta})$, where strict inequality applies for $i = n$ and equality for $i = s$. Therefore, we have

$$\begin{aligned}(\gamma - \varepsilon\tilde{x}^i - \tilde{p})\left[1 - \frac{\beta\delta}{\hat{\beta}}((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}))\right] > 0 \geq \beta\delta\omega(\tilde{w}^i - w^H) \\ - (\tilde{w}^s(\hat{\beta}) - \tilde{w}^i)(d - \lambda^{*s}(\hat{\beta}))\varepsilon\frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right).\end{aligned} \tag{B.7}$$

Therefore, Equation (B.1) is not satisfied and consumption must increase. Thus, in steady state $\tilde{x}^i > x^H$ and $\tilde{w}^i > w^H$. This has proven that $x^F > x^H$ is a sufficient condition for an overweight steady state. In order to see that this condition is also necessary, assume that the opposite holds, i.e. $x^F \leq x^H$. In this case we use proof by contradiction to show that a steady state of $\tilde{w}^i > w^H$ is not possible. Suppose that $\tilde{w}^i > w^H$ is the steady state. Then $\gamma - \varepsilon\tilde{x}^i - \tilde{p} < \gamma - \varepsilon x^H - \tilde{p} \leq \gamma - \varepsilon x^F - \tilde{p} = 0$. Moreover, according to Equation (B.4) we have $\tilde{w}^i \geq \tilde{w}^s(\hat{\beta})$, where strict inequality applies for $i = n$ and equality for $i = s$. Thus, we can show

$$\begin{aligned}(\gamma - \varepsilon\tilde{x}^i - \tilde{p})\left[1 - \frac{\beta\delta}{\hat{\beta}}((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}))\right] < 0 < \beta\delta\omega(\tilde{w}^i - w^H) \\ - (\tilde{w}^s(\hat{\beta}) - \tilde{w}^i)(d - \lambda^{*s}(\hat{\beta}))\varepsilon\frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right).\end{aligned} \tag{B.8}$$

Therefore, Equation (B.1) is violated. This means that $x^F > x^H$ is also a necessary condition.

Next we show that $x^F < x^H$ is a sufficient condition for an underweight steady state. If this condition holds, then $\gamma - \varepsilon\tilde{x}^i - \tilde{p} < 0$ for $\tilde{x}^i \geq x^H$. Suppose that the individual achieves a healthy or overweight steady state, i.e. $\tilde{w}^i \geq w^H$. Using these conditions and (B.4), we can show that

$$(\gamma - \varepsilon\tilde{x}^i - \tilde{p}) \left[1 - \frac{\beta\delta}{\hat{\beta}}((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}))\right] < 0 \leq \beta\delta\omega(\tilde{w}^i - w^H)$$
$$- (\tilde{w}^s(\hat{\beta}) - \tilde{w}^i)(d - \lambda^{*s}(\hat{\beta}))\varepsilon\frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right). \tag{B.9}$$

The above inequality violates Equation (B.1) and we conclude that $x^F < x^H$ is sufficient for an underweight steady state. To show that it is also necessary, assume that it does not hold, i.e. $x^F \geq x^H$. Assume furthermore that $\tilde{w}^i < w^H$. Under these conditions $\gamma - \varepsilon\tilde{x}^i - \tilde{p} > \gamma - \varepsilon x^H - \tilde{p} \geq \gamma - \varepsilon x^F - \tilde{p} = 0$. Using Equation (B.4), we can show that

$$(\gamma - \varepsilon\tilde{x}^i - \tilde{p}) \left[1 - \frac{\beta\delta}{\hat{\beta}}((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}))\right] > 0 > \beta\delta\omega(\tilde{w}^i - w^H)$$
$$- (\tilde{w}^s(\hat{\beta}) - \tilde{w}^i)(d - \lambda^{*s}(\hat{\beta}))\varepsilon\frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right). \tag{B.10}$$

The above inequality violates Equation (B.1) and, therefore, an underweight steady state is not possible under $x^F \geq x^H$, i.e. $x^F < x^H$ is also necessary for this steady state.

Consider lastly the case $x^F = x^H$. Plugging this condition in Equation (B.2), we conclude that $\tilde{w}^s = w^H, \forall\beta$. Using this result and Equation (B.6), we get $\tilde{w}^n = w^H$. In order to show that it is also necessary, assume $x^F \neq x^H$ and $\tilde{w}^i = w^H$. From Equation (B.6), we know that the latter condition implies $\tilde{w}^s(\hat{\beta}) = \tilde{w}^i$. Thus, Equation (B.1) can be rewritten as

$$(\gamma - \varepsilon\tilde{x}^i - \tilde{p}) \left[1 - \frac{\beta\delta}{\hat{\beta}}((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}))\right] = 0. \tag{B.11}$$

The above equality is satisfied if and only if $\tilde{x}^i = x^F$. However, $\tilde{w}^i = w^H$ implies $\tilde{x}^i = x^H$. This contradicts the assumption $x^F \neq x^H$ and we conclude that $x^F = x^H$ is both necessary and sufficient condition for a healthy steady state. Q.E.D.

C Proof of Proposition 3

In order to derive the effects of β and $\hat{\beta}$ on the steady state weight, note first that x^F and x^H do not depend on these variables. Second, $\hat{\beta}$ does not impact a sophisticated

individual. Thus, we only need to derive the impact of β on the steady state weight of a sophisticated consumer, which is solely determined by its effect on $\tilde{w}^s$ and through it on $\tilde{x}^s$. Therefore, we differentiate Equation (B.2) with respect to β and $\tilde{w}^s$ and simplify:

$$\frac{d\tilde{w}^s}{d\beta} = -\frac{\delta\omega(\tilde{w}^s - w^H)\left[1 - \frac{\beta\delta\left(\lambda^{*s} - (1-\beta)\frac{\partial\lambda^{*s}}{\partial\beta}\right)}{(1-\delta((1-d)+(1-\beta)\lambda^{*s}))}\right]}{\varepsilon d\left(1-\delta\left((1-d)+(1-\beta)\lambda^{*s}\right)\right)+\beta\delta\omega} \begin{cases} < 0, & \text{if} \quad \tilde{w}^s > w^H \\ > 0, & \text{if} \quad \tilde{w}^s < w^H. \\ = 0, & \text{if} \quad \tilde{w}^s = w^H. \end{cases} \tag{C.1}$$

Note that in determining the sign of Equation (C.1), we need to prove that the term in brackets in the numerator of (C.1) is positive. This condition can be simplified in the following way:

$$1 - \frac{\beta\delta\left(\lambda^{*s} - (1-\beta)\frac{\partial\lambda^{*s}}{\partial\beta}\right)}{(1-\delta((1-d)+(1-\beta)\lambda^{*s}))} > 0 \quad \Leftrightarrow \quad 1-\delta(1-d+\lambda^{*s})+\delta\beta(1-\beta)\frac{\partial\lambda^{*s}}{\partial\beta} > 0, \tag{C.2}$$

where $\partial\lambda^{*s}/\partial\beta$ can be derived by totally differentiating Equation (A.10) with respect to λ^{*s} and β and is given by

$$\frac{\partial\lambda^{*s}}{\partial\beta} = -\frac{(1-d+\lambda)\delta(\varepsilon\lambda^{*s\,2}+\omega)}{\varepsilon\left[1-\delta(1-d)^2-\delta\lambda^{*s}(2(1-d)(2-\beta)+3\lambda^{*s}(1-\beta))\right]+\delta\beta\omega} < 0. \tag{C.3}$$

Inserting (C.3) in (C.2) and rearranging, we can show after some tedious calculations that it always holds:

$$\begin{aligned} 1-\delta(1-d+\lambda^{*s})+\delta\beta(1-\beta)\frac{\partial\lambda^{*s}}{\partial\beta} &= \frac{\varepsilon(1-\delta(1-d+\lambda^{*s}))(1-\delta((1-d)^2-\lambda^{*s}d))}{\varepsilon\left[1-\delta(1-d)^2-\delta\lambda^{*s}(2(1-d)(2-\beta)+3\lambda^{*s}(1-\beta))\right]+\delta\beta\omega} \\ -\varepsilon\delta\lambda^{*s}\frac{(1-d+\lambda^{*s})(1-\beta)(1-\delta(1-d+\lambda^{*s}))+\frac{1-\delta(1-d+\lambda^{*s})(\delta+(d-\lambda^{*s})(1-\beta-\delta))}{\delta(1-d+\lambda^{*s})}}{\varepsilon\left[1-\delta(1-d)^2-\delta\lambda^{*s}(2(1-d)(2-\beta)+3\lambda^{*s}(1-\beta))\right]+\delta\beta\omega} &> 0. \end{aligned} \tag{C.4}$$

Thus, the numerator of Equation (C.1) has the same sign as $\tilde{w}^s - w^H$.

Now we totally differentiate Equation (B.3) with respect to β and $\tilde{w}^n$ and simplify:

$$\frac{d\tilde{w}^n}{d\beta} = \frac{-\delta\omega(\tilde{w}^n - w^H) - \left[(\gamma - \varepsilon d\tilde{w}^n - \tilde{p}) - (d - \lambda^{*s}(\hat{\beta}))\varepsilon(\tilde{w}^s(\hat{\beta}) - \tilde{w}^n)\right]\frac{\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)}{\beta\delta\omega + \varepsilon d - \varepsilon\lambda^{*s}(\hat{\beta})\frac{\beta\delta}{\hat{\beta}}(1-d+(1-\hat{\beta})\lambda^{*s}(\hat{\beta}))}. \tag{C.5}$$

Equations (B.4), (B.6) and (C.5) together prove that the sign of the effect of β on the weight of the naive individual is the same as in the case of a sophisticated individual.

Therefore, (C.1) and (C.5) together prove the first part of Proposition 3: a reduction in β cannot determine whether the individual is overweight, underweight or healthy weight, because it does not impact the healthy weight consumer and raises (lowers) the weight of an overweight (underweight) individual.

Next we show that naiveté does not impact the individual's decision to be over-/healthy- or underweight. This follows immediately from Equations (B.4) and (C.1):

$$\text{sgn}\{\tilde{w}^n - \tilde{w}^s(\hat{\beta})\} = \text{sgn}\{\tilde{w}^s(\hat{\beta}) - w^H\} = \text{sgn}\{\tilde{w}^s(\beta) - w^H\}, \quad \forall \beta, \hat{\beta}. \tag{C.6}$$

The first equality in (C.6) follows directly from Equation (B.4) and the second from (C.1). Thus, a naive individual is overweight if a sophisticated consumer with the same β is overweight. However, the above result does not state whether the difference $\tilde{w}^n - w^H$ is greater or smaller than $\tilde{w}^s(\beta) - w^H$.

Therefore, next we consider the impact of naivete on steady state weight.

In order to derive the effect of naivete on the steady state weight, we totally differentiate Equation (B.3) with respect to weight and $\hat{\beta}$, which gives

$$\begin{aligned}\frac{d\tilde{w}^n}{d\hat{\beta}} = & \frac{\varepsilon\frac{\beta\delta}{\hat{\beta}}\Big((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\Big)\Big[(d-\lambda^{*s}(\hat{\beta}))\frac{d\tilde{w}^s(\hat{\beta})}{d\hat{\beta}} - (\tilde{w}^s(\hat{\beta})-\tilde{w}^n)\frac{\partial\lambda^{*s}(\hat{\beta})}{\partial\hat{\beta}}\Big]}{\beta\delta\omega+\varepsilon d-\varepsilon\lambda^{*s}(\hat{\beta})\frac{\beta\delta}{\hat{\beta}}(1-d+(1-\hat{\beta})\lambda^{*s}(\hat{\beta}))} \\ & + \frac{\Big[(\gamma-\varepsilon d\tilde{w}^n-\tilde{p})-(d-\lambda^{*s}(\hat{\beta}))\varepsilon(\tilde{w}^s(\hat{\beta})-\tilde{w}^n)\Big]\frac{\beta\delta}{\hat{\beta}^2}\Big(1-d+\lambda^{*s}(\hat{\beta})-\hat{\beta}(1-\hat{\beta})\frac{\partial\lambda^{*s}(\hat{\beta})}{\partial\hat{\beta}}\Big)}{\beta\delta\omega+\varepsilon d-\varepsilon\lambda^{*s}(\hat{\beta})\frac{\beta\delta}{\hat{\beta}}(1-d+(1-\hat{\beta})\lambda^{*s}(\hat{\beta}))}.\end{aligned} \tag{C.7}$$

Using Equations (B.4) and (B.6), we can immediately show that the right-hand side of (C.7) is zero, if $\tilde{w}^n = w^H$. On the other hand, if $\tilde{w}^n > w^H$, then the first row is negative and the second row is positive. The opposite holds for $\tilde{w}^n < w^H$. It is not possible to derive an explicit result for the sign of (C.7) in the latter two cases. Therefore, we conclude that naivete does not affect the weight of a healthy-weight individual and has an ambiguous impact on the weight of a non-healthy weight individual. Q.E.D.

D Proof of Proposition 4

First we derive the sign of the steady state tax when the individual is overweight or underweight. If the consumer is sophisticated, then $\tilde{\Delta}^s = 0$ and $\tilde{\tau} > (<0) \Leftrightarrow \tilde{w}^s > (<$

$)w^H$ follows immediately from Proposition 2 and Equation (27). In the case of a naive consumer, plug Equation (B.3) in (27) in order to derive the following expression for the steady state tax:

$$\tilde{\tau}(1-\delta(1-d)) = \delta\omega(\tilde{w}^n - w^H) - (\gamma - \varepsilon d\tilde{w}^n - \tilde{p})(1-\delta(1-d)). \tag{D.1}$$

Suppose $\tilde{w}^n - w^H > 0$. Then, the following relations hold:

$$\begin{aligned}\delta\omega(\tilde{w}^n - w^H) > \delta\omega(\tilde{w}^s(\hat{\beta}) - w^H) &= \frac{(\gamma - \varepsilon d\tilde{w}^s(\hat{\beta}) - \tilde{p})(1-\delta((1-d)) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}))}{\hat{\beta}} \\ &> (\gamma - \varepsilon d\tilde{w}^n - \tilde{p})(1-\delta(1-d)) > 0. \end{aligned} \tag{D.2}$$

The first inequality in Equation (D.2) stems from (B.4), the next equality comes from (B.2) and the third inequality is also a consequence of (B.4) and $\lambda^{*s}(\hat{\beta}) < 0$. Together (D.1) and (D.2) determine $\tilde{\tau} > 0$ if $\tilde{w}^n > w^H$.

Suppose $\tilde{w}^n - w^H < 0$. Then, the following relations hold:

$$\begin{aligned}\delta\omega(\tilde{w}^n - w^H) < \delta\omega(\tilde{w}^s(\hat{\beta}) - w^H) &= \frac{(\gamma - \varepsilon d\tilde{w}^s(\hat{\beta}) - \tilde{p})(1-\delta((1-d)) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}))}{\hat{\beta}} \\ &< (\gamma - \varepsilon d\tilde{w}^n - \tilde{p})(1-\delta(1-d)) < 0. \end{aligned} \tag{D.3}$$

The inequalities in (D.3) come from (B.4) and the equality from (B.2). Together (D.1) and (D.3) determine $\tilde{\tau} < 0$ if $\tilde{w}^n < w^H$. Thus, we have proven that the tax rate is positive, if the individual is overweight, and negative, if the individual is underweight.

The next result in Propositoin 5 states that the steady state tax is zero if (i) the individual has no self-control problems or (ii) the individual has a healthy weight in the absence of taxation. Moreover, there does not exist a non-zero steady state tax compatible with healthy weight. These results follow directly from evaluating Equation (27) first at $\beta = \hat{\beta} = 1$ and secondly at $\tilde{w}^i = w^H$.

Lastly, we derive the optimal trajectory for τ_t during the transition to steady state. The equation of motion for weight (1) and Equation (26) determine the following system of two linear first-order difference equations:

$$\begin{pmatrix} w^i_{t+1} \\ \tau_{t+1} \end{pmatrix} = \underbrace{\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}}_{=J} \begin{pmatrix} w^i_t \\ \tau_t \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \tag{D.4}$$

where

$$a_{11} \equiv 1 - d + \lambda^{*i} > 0, \tag{D.5a}$$

$$a_{12} \equiv \frac{d\mu_t^{*i}}{dp_t} = \frac{d\mu^{*i}}{dp} < 0, \tag{D.5b}$$

$$a_{21} \equiv \frac{(1-d+\lambda^{*i})\left[\delta(1-d)\varepsilon\lambda^{*i} - \delta(1-\beta)\omega - \frac{\beta\delta}{\hat{\beta}}\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right]}{-\delta(1-d)\varepsilon\frac{d\mu^{*i}}{dp} + \frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)} < 0, \tag{D.5c}$$

$$a_{22} \equiv \frac{1+\frac{d\mu^{*i}}{dp}\left[\delta(1-d)\varepsilon\lambda^{*i} - \delta(1-\beta)\omega - \frac{\beta\delta}{\hat{\beta}}\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right]}{-\delta(1-d)\varepsilon\frac{d\mu^{*i}}{dp} + \frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)} > 0. \tag{D.5d}$$

The terms b_1, b_2 are constants. The system of difference equations (D.4) has two eigenvalues denoted by $\nu_i, i = 1, 2$ and is saddle-path stable if one eigenvalue is in the interval $]0, 1[$ while the other is greater than one. In order to determine the stability properties of this system, we need to derive and determine the signs of the trace and determinant of the matrix J:

$$\begin{aligned} Tr(J) &= a_{11} + a_{22} \\ &= \frac{\left\{\begin{array}{c} 1 + (1-d+\lambda^{*i})\frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right) \\ -\frac{d\mu^{*i}}{dp}\left[\delta(1-d)^2\varepsilon + \delta(1-\beta)\omega + \frac{\beta\delta}{\hat{\beta}}\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right] \end{array}\right\}}{-\delta(1-d)\varepsilon\frac{d\mu^{*i}}{dp} + \frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)} > 0, \end{aligned} \tag{D.6}$$

$$\begin{aligned} Det(J) &= a_{11}a_{22} - a_{12}a_{21} \\ &= \frac{1-d+\lambda^{*i}}{-\delta(1-d)\varepsilon\frac{d\mu^{*i}}{dp} + \frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)} > 0. \end{aligned} \tag{D.7}$$

In determining the signs of the above expressions, we used Equation (18) in order to show that the last term in the denominator is positive:

$$\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right) = \frac{-\delta\left[\varepsilon\lambda^{*s}(\hat{\beta})((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})) - \beta\omega\right]}{-\delta\left[\varepsilon\lambda^{*s}(\hat{\beta})((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})) - \beta\omega\right] + \varepsilon[1 - a\delta((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta}))]} \in]0, \tag{D.8}$$

Since $\nu_1 + \nu_2 = Tr(J) > 0$ and $\nu_1\nu_2 = Det(J) > 0$, we conclude that both eigenvalues are positive. However, a sufficient condition for a saddle path is $Det(J) \in]0, 1[$. On order to prove whether this condition holds, use Equations (19) and (A.14) to derive the following expression for the effect of the price of unhealthy food on its consumption:

$$\frac{d\mu^{*n}}{dp} = -\frac{1-d+\lambda^{*n}}{(1-d)\varepsilon}\left[1 - a\frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1 + \varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)\right]. \tag{D.9}$$

Note that if $\beta = \hat{\beta}$, then Equation (D.9) coincides with (18), which determines $d\mu^{*s}/dp$. Therefore, Equation (D.9) can be rewritten with a superscript i instead of n as it can be solved for both sophisticated and naive individuals. Then, the determinant can be rewritten as

$$Det(J) = \frac{1-d+\lambda^{*i}}{\delta(1-d+\lambda^{*i}) + \frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1 + \varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)(1 - a\delta(1-d+\lambda^{*i}))}. \tag{D.10}$$

If δ is equal to or close to one, then the determinant is unambiguously smaller than one. However, there exists a small enough δ, which makes the above term greater than one. Define the value of δ which makes $Det(J) = 1$ as $\underline{\delta}$. Then, a sufficient condition for the system to be saddle-path is $\delta \in]\underline{\delta}, 1[$.

Denoting the eigenvalue that is less than one as ν_1, we can express it as

$$\nu_1 = \frac{Tr(J) - \sqrt{Tr(J)^2 - 4Det(J)}}{2}. \tag{D.11}$$

The system of difference equations (D.4) can then be solved and the solution is given by

$$w_t^i = \tilde{w}^i + (w_0^i - \tilde{w}^i)\nu_1^t, \tag{D.12a}$$

$$\tau_t = \tilde{\tau} - \frac{a_{11} - \nu_1}{a_{12}}(w_0^i - \tilde{w}^i)\nu_1^t. \tag{D.12b}$$

Q.E.D.

E Proof of Proposition 5

In order to derive the steady state price level $\tilde{p}$, evaluate Equation (31) in the steady state. Its left-hand side equals zero and the right-hand side can be extended to

$$0 = \delta(1-\beta)\omega(\tilde{w}^i - w^H) + (\gamma - \varepsilon d\tilde{w}^i - \tilde{p})\left[\delta(1-d) - \frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right]$$

$$+\varepsilon(d-\lambda^{*s}(\hat{\beta}))(\tilde{w}^s(\hat{\beta})-\tilde{w}^i)\frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right). \tag{E.1}$$

Assume first that the individual is sophisticated. In this case the second row of Equation (E.1) is identically equal to zero, while Equation (B.2) from Appendix B shows that the two terms on the first row are of the same sign. Thus, if $\gamma-\varepsilon d\tilde{w}^s-\tilde{p}>(<)0$, the right-hand side of (E.1) does not equal zero. The unique steady state price level which satisfies (E.1) in the case of a sophisticated consumer is, thus, $\tilde{p}=\gamma-\varepsilon d\tilde{w}^s=\gamma-\varepsilon dw^H$.

On the other hand, if the consumer is naive, we evaluate Equation (30) in steady state in the case $i=n$:

$$(1-\delta(1-d))(\gamma-\varepsilon d\tilde{w}^n-\tilde{p})=\delta\omega(\tilde{w}^n-w^H). \tag{E.2}$$

We solve Equation (E.2) for $(\gamma-\varepsilon d\tilde{w}^n-\tilde{p})$, plug the resulting expression in Equation (E.1) and rewrite $(\tilde{w}^s(\hat{\beta})-\tilde{w}^n)=(\tilde{w}^s(\hat{\beta})-w^H)-(\tilde{w}^n-w^H)$ in order to derive after some rearrangement the following expression

$$\varepsilon(d-\lambda^{*s}(\hat{\beta}))\frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)(\tilde{w}^s(\hat{\beta})-w^H)=(\tilde{w}^n-w^H)\cdot$$
$$\cdot\left[\delta\omega\left(\beta-\frac{1-\frac{\beta\delta\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)}{\hat{\beta}}}{1-\delta(1-d)}\right)+\varepsilon(d-\lambda^{*s}(\hat{\beta}))\frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right]. \tag{E.3}$$

Next we use Equations (B.2) and (E.2) in order to show the following relation:

$$\begin{aligned}(\tilde{w}^s(\hat{\beta})-w^H)&=\frac{(\gamma-\varepsilon d\tilde{w}^s(\hat{\beta})-\tilde{p})\left[1-\delta\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right]}{\hat{\beta}\delta\omega}\\&=\frac{(\gamma-\varepsilon d\tilde{w}^n-\tilde{p}-\varepsilon d(\tilde{w}^s(\hat{\beta})-\tilde{w}^n))\left[1-\delta\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right]}{\hat{\beta}\delta\omega}\\&=\left[\frac{\delta\omega(\tilde{w}^n-w^H)}{1-\delta(1-d)}-\varepsilon d(\tilde{w}^s(\hat{\beta})-\tilde{w}^n)\right]\frac{\left[1-\delta\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right]}{\hat{\beta}\delta\omega}.\end{aligned} \tag{E.4}$$

Using Equation (E.4) to substitute for $(\tilde{w}^s(\hat{\beta})-w^H)$ in (E.3), we get after some rearrangement

$$\varepsilon^2(d-\lambda^{*s}(\hat{\beta}))\frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\frac{\left[1-\delta\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right]d}{\hat{\beta}\delta\omega}(\tilde{w}^n-\tilde{w}^s(\hat{\beta}))$$

$$= (\tilde{w}^n - w^H) \left[\delta\omega \left(\beta - \frac{1 - \frac{\beta\delta\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)}{\hat{\beta}}}{1-\delta(1-d)} \right) + \varepsilon(d - \lambda^{*s}(\hat{\beta})) \frac{\beta\delta}{\hat{\beta}} \left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}) \right) \cdot \right.$$
$$\left. \cdot \left(1 - \frac{\left[1 - \delta \left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}) \right) \right]}{\hat{\beta}(1-\delta(1-d))} \right) \right]. \tag{E.5}$$

Equations (B.4) and (B.6) show that if $\tilde{w}^n - w^H > 0$, then the right-hand side of (E.5) is negative, while its left-hand side is positive, i.e. this case cannot be a solution. If $\tilde{w}^n - w^H < 0$, then the right-hand side of (E.5) is positive, while its left-hand side is negative, i.e. this case also cannot be a solution. Lastly, $\tilde{w}^n = w^H$ makes both sides of (E.5) equal to zero and is the unique solution of the steady state. Using this last equality and Equation (E.2), we derive $\tilde{p} = \gamma - \varepsilon d w^H$. Hence, this is the unique price level, which solves the social planner's optimization problem in steady state.

In order to prove that $\widetilde{\Delta}^i$ equals zero, note that $\tilde{w}^s(\hat{\beta}) = \tilde{w}^n = w^H$. Thus, $\tilde{x}^s(\hat{\beta}) = \lambda^{*s}(\hat{\beta}) + \tilde{w}^s(\hat{\beta})(d - \lambda^{*s}(\hat{\beta})) = dw^H = \tilde{x}^n$. Hence, $(v'(\tilde{x}^s(\hat{\beta})) - \tilde{p}) = (v'(\tilde{x}^n) - \tilde{p}) = 0$ and from the definition of $\widetilde{\Delta}^n$, we immediately see that it also equals zero.

The optimal trajectory to the steady state is determine by Equations (1) and (31). They define the following system of first-order linear difference equations:

$$\begin{pmatrix} w^i_{t+1} \\ p_{t+1} \end{pmatrix} = \underbrace{\begin{pmatrix} \overline{a}_{11} & \overline{a}_{12} \\ \overline{a}_{21} & \overline{a}_{22} \end{pmatrix}}_{=\overline{J}} \begin{pmatrix} w^i_t \\ p_t \end{pmatrix} + \begin{pmatrix} \overline{b}_1 \\ \overline{b}_2 \end{pmatrix}, \tag{E.6}$$

where

$$\overline{a}_{11} \equiv 1 - d + \lambda^{*i} > 0, \tag{E.7a}$$

$$\overline{a}_{12} \equiv \frac{d\mu^{*i}_t}{dp_t} = \frac{d\mu^{*i}}{dp} < 0, \tag{E.7b}$$

$$\overline{a}_{21} \equiv \frac{\left\{ \begin{array}{c} (1-d+\lambda^{*i}) \left[-\delta(1-d)\varepsilon\lambda^{*i} + \delta(1-\beta)\omega + \frac{\beta\delta}{\hat{\beta}}\varepsilon\lambda^{*s}(\hat{\beta}) \left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}) \right) \right] \\ + \frac{\lambda^{*i}[1-\delta(1-d+\lambda^{*i})^2]}{d\mu^{*i}/dp} \end{array} \right\}}{\delta(1-d)\left(1 + \varepsilon\frac{d\mu^{*i}}{dp}\right) - \frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta}) \right)\left(1 + \varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right) + \delta(1-d+\lambda^{*i})} > 0, \tag{E.7c}$$

$$\overline{a}_{22} \equiv \frac{1-\delta\lambda^{*i}(1-d+\lambda^{*i}) - \frac{d\mu^{*i}}{dp}\left[\delta(1-d)\varepsilon\lambda^{*i} - \delta(1-\beta)\omega - \frac{\beta\delta}{\hat{\beta}}\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right]}{\delta(1-d)\left(1+\varepsilon\frac{d\mu^{*i}}{dp}\right) - \frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right) + \delta(1-d+\lambda^{*i})}. \tag{E.7d}$$

The terms $\overline{b}_1, \overline{b}_2$ are constants. The system of difference equations (E.6) has two eigenvalues denoted by $\overline{\nu}_i, i = 1, 2$. The trace and determinant of the matrix $\overline{J}$ are given by:

$$\begin{aligned} Tr(\overline{J}) &= \overline{a}_{11} + \overline{a}_{22} \\ &= \frac{\left\{\begin{array}{c} 1+(1-d+\lambda^{*i})\left[2\delta(1-d) + \frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)\right] \\ +\frac{d\mu^{*i}}{dp}\left[\delta(1-d)^2\varepsilon + \delta(1-\beta)\omega + \frac{\beta\delta}{\hat{\beta}}\varepsilon\lambda^{*s}(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right] \end{array}\right\}}{\delta(1-d)\left(1+\varepsilon\frac{d\mu^{*i}}{dp}\right) - \frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right) + \delta(1-d+\lambda^{*i})}, \end{aligned} \tag{E.8}$$

$$\begin{aligned} Det(\overline{J}) &= \overline{a}_{11}\overline{a}_{22} - \overline{a}_{12}\overline{a}_{21} \\ &= \frac{1-d}{\delta(1-d)\left(1+\varepsilon\frac{d\mu^{*i}}{dp}\right) - \frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right) + \delta(1-d+\lambda^{*i})} > 0. \end{aligned} \tag{E.9}$$

Using Equation (D.9), we can show that the trace is greater than zero:

$$Tr(\overline{J}) = \frac{\left\{\begin{array}{c} 1+\frac{d\mu^{*i}}{dp}\delta(1-\beta)\omega + \delta(1-d+\lambda^{*i})\left[(1-d) - \frac{\beta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)\right] \\ +(1-d+\lambda^{*i})\frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)a\delta\cdot \\ \cdot\left[(1-d)+\frac{\lambda^{*s}(\hat{\beta})\beta}{(1-d)\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right] \end{array}\right\}}{\delta(1-d) - \frac{\beta\delta}{\hat{\beta}}\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1+\varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)(1-a\delta(1-d+\lambda^{*i}))} \tag{E.10}$$

Note that the denominator is unambiguously positive and all the terms in the numerator are positive expect for the term involving $d\mu^{*i}/dp$. However, it can be shown that the sum of the first two terms is positive for $\beta \geq 1/2$:

$$\begin{aligned} 1+\frac{d\mu^{*i}}{dp}\delta(1-\beta)\omega &= 1+\frac{d\mu^{*i}}{dp}\frac{1-\beta}{\beta}\delta\beta\omega \\ &= 1+\frac{d\mu^{*i}}{dp}\frac{1-\beta}{\beta}\left[\frac{(1-d)\varepsilon}{1-d+\lambda^{*i}} - \varepsilon\left(1-\frac{\beta\delta}{\hat{\beta}}\lambda^{*s}(\hat{\beta})\left((1-d)+(1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right)\right] \end{aligned}$$

$$= 1 - \frac{1-\beta}{\beta}\left[1 - a\frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1 + \varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)\right]$$
$$- \frac{d\mu^{*i}}{dp}\frac{1-\beta}{\beta}\varepsilon\left(1 - \frac{\beta\delta}{\hat{\beta}}\lambda^{*s}(\hat{\beta})\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\right) > 0, \quad \text{if} \quad \beta \geq 1/2.$$

In deriving the second row of the above equation, we used Equation (A.11) and in deriving the third row we used (D.9).

Moreover, the determinant is greater than one:

$$Det(\overline{J}) = \frac{1-d}{\delta(1-d) - \frac{\beta\delta}{\hat{\beta}}\left((1-d) + (1-\hat{\beta})\lambda^{*s}(\hat{\beta})\right)\left(1 + \varepsilon\frac{d\mu^{*s}(\hat{\beta})}{dp}\right)(1 - a\delta(1-d+\lambda^{*i}))} > 1. \tag{E.11}$$

Equations (E.10) and (E.11) show that the sum of the two eigenvalues is positive and their product is greater than one. Therefore, both are positive and at least one of them is greater than one. The system is a saddle-path if the other eigenvalue is smaller than one and unstable otherwise. In the first case, the eigenvalue which is less than one can be denoted as $\overline{\nu}_1$ and is determined by

$$\overline{\nu}_1 = \frac{Tr(\overline{J}) - \sqrt{Tr(\overline{J})^2 - 4Det(\overline{J})}}{2}. \tag{E.12}$$

If the system of difference equations (E.6) is a saddle-path, then it can be easily solved and its solution is given by Equations (32a), (32b). Q.E.D.

References

Becker, G. S. and Murphy, K. M. (1988). A Theory of Rational Addiction. *Journal of Political Economy*, 96(4):675–700.

Diamond, P. and Köszegi, B. (2003). Quasi-hyperbolic discounting and retirement. *Journal of Public Economics*, 87(9-10):1839–1872.

Dragone, D. (2009). A rational eating model of binges, diets and obesity. *Journal of Health Economics*, 28(4):799–804.

Dragone, D. and Savorelli, L. (2012). Thinness and obesity: A model of food consumption, health concerns, and social pressure. *Journal of Health Economics*, 31(1):243–256.

Gruber, J. and Köszegi, B. (2001). Is Addiction "Rational"? Theory and Evidence. *The Quarterly Journal of Economics*, 116(4):1261–1303.

Gruber, J. and Köszegi, B. (2004). Tax incidence when individuals are time-inconsistent: the case of cigarette excise taxes. *Journal of Public Economics*, 88(9-10):1959–1987.

Haavio, M. and Kotakorpi, K. (2011). The political economy of sin taxes. *European Economic Review*, 55(4):575–594.

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

Lakdawalla, D., Philipson, T., and Bhattacharya, J. (2005). Welfare-Enhancing Technological Change and the Growth of Obesity. *American Economic Review*, 95(2):253–257.

Levy, A. (2002). Rational eating: can it lead to overweightness or underweightness? *Journal of Health Economics*, 21(5):887–899.

O'Donoghue, T. and Rabin, M. (2001). Choice and Procrastination. *The Quarterly Journal of Economics*, 116(1):121–160.

O'Donoghue, T. and Rabin, M. (2003). Studying Optimal Paternalism, Illustrated by a Model of Sin Taxes. *American Economic Review*, 93(2):186–191.

O'Donoghue, T. and Rabin, M. (2006). Optimal sin taxes. *Journal of Public Economics*, 90(10-11):1825–1849.

Philipson, T. J. and Posner, R. A. (1999). The Long-Run Growth in Obesity as a Function of Technological Change. NBER Working Papers 7423, National Bureau of Economic Research, Inc.