Discussion Paper Series – CRC TR 224

Discussion Paper No. 450
Project C 01

# The Welfare Economics of Reference Dependence

Arthur Seibold [1]
Daniel Reck [2]

August 2023

[1] University of Mannheim, Email: seibold@uni-mannheim.de
[2] University of Maryland, Email: dreck@umd.edu

Support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 is gratefully acknowledged.

# The Welfare Economics of Reference Dependence*

Daniel Reck
University of Maryland

Arthur Seibold
University of Mannheim

June 2023

**Abstract**

Empirical evidence suggests that individuals often evaluate options relative to a reference point, especially seeking to avoid losses. We undertake the first welfare analysis under reference-dependent preferences. We characterize the welfare impact of changes in reference points and prices, decomposing these into direct and behavioral effects. The sign of direct and behavioral effects depends on the form of reference-dependent payoffs; which of these effects matter for welfare depends on whether reference dependence reflects a bias or a normative preference. We derive sufficient statistics formulas quantifying the social welfare effects of changes in reference points and prices in terms of estimable reduced-form parameters and normative judgments. We illustrate these findings with an empirical application to reference dependence exhibited in German workers' retirement decisions. We find positive social welfare effects of increasing the Normal Retirement Age, but ambiguous effects of financial incentives to postpone retirement.

JEL codes: D91, D60, H55, J26
Keywords: reference-dependent preferences, loss aversion, welfare, pension reform

*dreck@umd.edu and seibold@uni-mannheim.de. *We thank Zarek Brot-Goldberg, Jack Fisher, Jacob Goldin, Nathan Hendren, Xavier Jaravel, Dmitri Koustas, Camille Landais, Emel Filiz-Ozbay, Alex Rees-Jones, Emilie Sartre, Ludvig Sinander, Johannes Spinnewijn, Charlie Sprenger, Neil Thakral, Dmitry Taubinsky, Teju Velayudhan, and numerous seminar and conference participants for helpful comments and discussions. Felix Knau, Canishk Naik and Baptiste Roux provided excellent research assistance. Daniel Reck gratefully acknowledges financial support from the Suntory and Toyota International Centres for Economics and Related Disciplines (STICERD) at the London School of Economics. Arthur Seibold gratefully acknowledges financial support from the Daimler & Benz Foundation and from the German Research Foundation (DFG) through CRC TR 224 (Project C01).*

# 1 Introduction

Reference-dependent preferences are a cornerstone of behavioral economics.[1] In a vast array of settings, decision-makers appear to evaluate options relative to a reference point, and they evaluate losses relative to the reference point more strongly than equivalent gains - *loss aversion*. Early evidence of such behavior came from classic laboratory experiments by Kahneman and Tversky (1979). Since then, the experiments have been replicated and extended in many ways, in parallel with a rich theoretical literature seeking to model reference dependence (see O'Donoghue and Sprenger, 2018, for a review). Empirical evidence of reference dependence has been found in experiments around the world (Ruggeri et al., 2020), and a wide range of field settings including the daily labor supply of taxi drivers (Camerer et al., 1997; Crawford and Meng, 2011; Thakral and Tô, 2021) and bicycle messengers (Fehr and Goette, 2007), job search (DellaVigna et al., 2017), behavioral responses to taxation (Homonoff, 2018; Rees-Jones, 2018), housing transactions (Andersen et al., 2022), and the timing of retirement (Seibold, 2021).

As policymakers take notice of mounting evidence of the importance of reference-dependent preferences, difficult questions loom large. How should we evaluate welfare in the presence of reference dependence? What are the policy implications of all the evidence that reference dependence matters? The vast literature on reference dependence has so far refrained from conducting welfare analysis, mainly because of two fundamental challenges. The first challenge is that it is unclear whether reference dependence represents a bias on the part of decision-makers or non-standard but normative preferences. This challenge is widely recognized in prior literature but remains unresolved (see the discussion in O'Donoghue and Sprenger, 2018). A second challenge is that many different formulations of reference-dependent payoffs have been proposed. Particular functional forms are often adopted with the aim of tractability or in order to rationalize empirical patterns, but the varying implications of payoff formulations for welfare have received little attention.

This paper undertakes the first analysis of the welfare economics of reference dependence. We address the first challenge by parameterizing a normative judgment about whether reference dependence reflects a bias or a normative preference, and then mapping this judgment to welfare (Goldin and Reck, 2022). We address the second challenge by analyzing welfare under minimal assumptions that encompass virtually all formulations of reference-dependent payoffs consistent with the empirical evidence base or considered in prior literature. Following this approach, we characterize the welfare impact of changes in the reference point and of changes in prices (or taxes), in terms of empirically identifiable parameters and normative judgments. We illustrate our findings in the retirement setting of Seibold (2021), where statutory retirement ages set by public policy influence individuals' reference points and implicit prices are given by financial retirement incentives.

Our theoretical analysis begins with a general characterization of welfare under minimal assumptions about reference-dependent payoffs. We consider a deterministic setting in which an individual receives reference-dependent utility from gains or losses of their consumption relative to a reference point, as in Tversky and Kahneman (1991). We follow the literature and assume that reference-dependent payoffs are kinked at the reference point, which captures the key empirical patterns typically associated with reference dependence. This property is usually interpreted as *loss aversion*, but there could be other motives rationalizing such a kink in preferences. This general framework allows us to characterize the first-order welfare effects of policies in terms of direct and behavioral effects. We show that which of these effects matter for

[1] For instance, DellaVigna (2018, p. 699) describes the theory of reference-dependent preferences as "perhaps the most influential model in behavioral economics."

welfare depends on the normative judgment about reference dependence.

We first analyze the welfare effects of policies that influence the reference point. Examples of such policies include governments setting a "Normal Retirement Age" (Seibold, 2021), or income tax withholding rules creating a reference point when filing a tax return (Rees-Jones, 2018). We find that if reference dependence reflects a normative preference, only a direct welfare effect arises as a result of changing the reference point. Intuitively, the individual's reference-dependent payoff is modified directly when comparing their outcome to a different reference point. Any changes in behavior do not entail first-order welfare effects due to the envelope theorem in this case. If reference dependence reflects a bias, on the other hand, there is no direct effect because reference-dependent payoffs do not enter welfare. However, the envelope theorem fails as reference dependence entails an *internality* (Mullainathan et al., 2012), resulting in a first-order behavioral welfare effect. We then show that under modest additional conditions – in particular, we rule out "diminishing sensitivity"[2] – direct and behavioral welfare effects of changing the reference point are same-signed. This implies, perhaps surprisingly, that normative ambiguity over whether reference dependence should enter welfare calculations is inconsequential for the sign of the welfare effect of changing the reference point. Instead, only the form of reference-dependent payoffs matters for the sign of welfare effects.

Empirical evidence suggests that reference dependence also affects behavioral responses to price instruments, e.g. commodity taxes (Homonoff, 2018). Motivated by such evidence, we analyze the welfare effects of a price change. Changing prices also has first-order direct and behavioral effects. Just as in standard models, a price increase has a negative direct welfare effect (regardless of normative judgments). And as before, when reference dependence is judged to be normative, the change in behavior has no first-order welfare implications. When reference dependence is a bias, however, the change in consumption caused by a price change has a first-order behavioral welfare effect. We show that the overall welfare effect of price changes generally depends both on normative judgments and the form of reference-dependent payoffs.

Next, we examine which features of the reference-dependent payoff formulation matter for welfare. We show in general terms that the nature of the discrepancy between reference-dependent demand and *intrinsic demand* – demand that would materialize in the absence of reference dependent concerns – is key for welfare. For instance, if reference dependence increases demand for the good, both direct and behavioral welfare effects of a higher reference point will be negative. The opposite is true if reference dependence decreases demand: direct and behavioral effects are then positive. This characterization allows us to understand what structure a given payoff formulation imposes on welfare, an understanding we can then apply to the wide variety of payoff formulations proposed in the literature.

First, we consider simple loss aversion models, where the main feature of reference dependence is loss aversion over a single good. Due to its simplicity and because it is able to capture key empirical patterns like bunching at a reference point or non-standard responses to price variation, this type of formulation is commonly used in applications (e.g. DellaVigna et al., 2017; Rees-Jones, 2018; Thakral and Tô, 2021; Seibold, 2021). We find that under simple loss aversion, decreasing the reference point has a positive direct welfare effect because this shrinks utility losses, and a positive behavioral welfare effect through mitigating overconsumption of the good. Our analysis hence reveals that adopting simple loss aversion models imposes significant ex-ante structure on welfare, where lower reference points are always preferred.

Some of the theoretical literature on reference dependence postulates more sophisticated formulations of reference-dependent payoffs, which are less restrictive in terms of welfare. A key example is allowing for

[2]Diminishing sensitivity is a feature of reference dependence typically considered in probabilistic settings, which introduces a specific curvature into reference-dependent payoffs. This can modify welfare effects far away from the reference point. See the discusion in Section 2.1 and Appendix B.5.

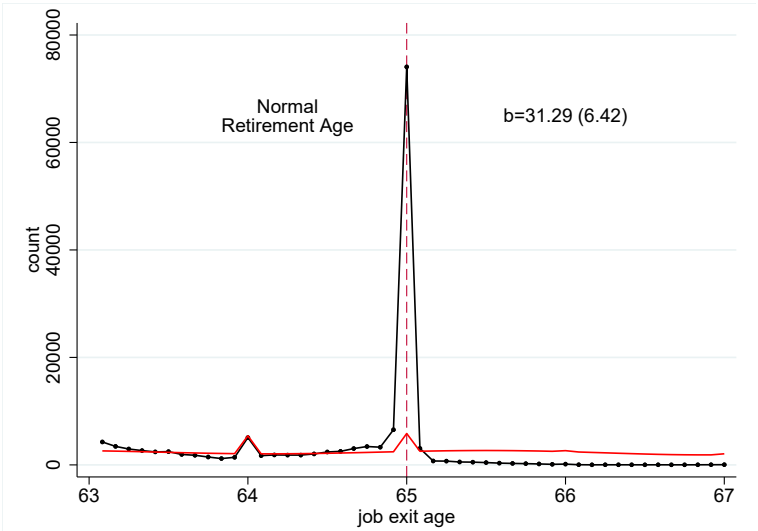
reference dependence over multiple goods. Since reference-dependent losses over one good can correspond to gains over another good, such an extension can change the sign of welfare effects. For instance, in labor supply contexts, reference dependence could be present over both consumption and leisure/labor supply (Crawford and Meng, 2011; Behaghel and Blau, 2012). We propose a flexible reduced form of reference-dependent payoffs that encompasses the key properties of a wide range of formulations in a parsimonious way, while imposing little ex-ante structure on welfare. The flexible reduced form features one parameter governing the size of kink in preferences typically attributed to loss aversion, and a second parameter governing the direction in which reference dependence modifies demand for the good. The additional parameter captures the degree of deviation from simple loss aversion in a reduced-form way, subsuming reference dependence over another good and any other reason why reference dependence modifies payoffs over gains.

Using our flexible reduced form, we derive novel sufficient statistics formulas for the social welfare effects of changes in reference points and prices. To quantify the first-order welfare effect of changing reference points, the key sufficient statistics are simply the two reference dependence parameters capturing the strength of loss aversion and the direction of reference dependence, given a normative judgment over reference dependence. For the welfare effects of price changes, an estimate of the price elasticity of demand for the good is additionally required. Since these parameters can be estimated across many settings in the literature, our sufficient statistics approach to reference dependence will be widely applicable.

We illustrate these theoretical results with an empirical application to old-age pension policy, building on Seibold (2021). The retirement setting has two important advantages for our purposes. First, common policies in this context correspond closely to the types of interventions we analyze theoretically. On the one hand, pension systems typically feature a Normal Retirement Age (NRA), which is presented as a "normal" time to retire and serves as a reference point for retirement decisions. Figure 1 shows that the empirical retirement age distribution exhibits strong bunching precisely at the NRA. As Seibold (2021) shows, this bunching cannot be explained by financial retirement incentives but is well in line with a model of reference dependence over the retirement age/leisure. Note that the figure also suggests a substantial drop in the retirement age distribution above the NRA, which is informative about the direction of reference dependence. Separately, pension systems provide financial retirement incentives which determine the marginal return to working longer (the implicit price of leisure). The second advantage of the empirical setting is that the relevant parameters governing individual behavior and welfare can be transparently estimated. In particular, we use high-quality administrative data on German retirees and exploit the bunching strategy of Seibold (2021) in order to estimate the responsiveness of retirement decisions to financial incentives and to the NRA as a reference point.

Our empirical application yields novel insights into the welfare effects of pension reforms in the presence of reference dependence. We quantify these welfare effects using (i) individual-level simulations of a model of retirement behavior and (ii) our sufficient statistics formulas. We focus on two types of pension reforms often discussed as policy options to induce workers to postpone retirement. The first reform is an increase in the NRA by one year. This reform increases the reference age of retirement, or equivalently lowers the reference point in terms of leisure, the corresponding good. In a simple loss aversion model consistent with empirically observed bunching, we find that such a reform always improves welfare. If reference dependence is judged as a bias, a lower reference point in terms of lifetime leisure counteracts some of the initial sub-optimal early retirement, bringing individuals closer to their optimal retirement age. If reference dependence is judged to be normative, a lower reference point yields direct welfare gains, as individuals

**FIGURE 1: BUNCHING AT THE NORMAL RETIREMENT AGE**




*Notes:* The figure shows the pooled distribution of retirement (job exit) ages around the Normal Retirement Age (NRA) among German workers born in 1946. The dashed vertical line demarcates the location of the NRA. The black connected dots show the actual distribution, while the red line shows the counterfactual density estimated as a seventh-order polynomial excluding the bunching region. The counterfactual density also allows for round-number bunching and features an upward correction to the right of the NRA, where a shift of the retirement age distribution occurs (see Section 4). The parameter $b$ denotes the excess mass at the NRA with its standard error shown in parantheses.

compare their lifetime leisure more favorably to the higher NRA.

The second reform we consider is an increase in the Delayed Retirement Credit (DRC), that is higher actuarial pension adjustment for working beyond the NRA. A higher DRC increases the marginal return to working, implying a higher implicit price of leisure. We find that the welfare effects of such a subsidy for later retirement strongly depend on normative judgments, even within a simple loss aversion model. On the one hand, a higher DRC can improve welfare when reference dependence is judged as a bias, because incentivizing workers to retire later mitigates sub-optimal early retirement. Optimal corrective subsidies for later retirement would be large in this case. If reference dependence is judged as normative, on the other hand, the welfare effects of the DRC are much more muted. Moderate actuarial adjustment can help correct fiscal externalities in the pension system, while an overly large DRC would distort retirement behavior, worsen the fiscal balance of the pension system and ultimately lower welfare.

Our empirical application illustrates that adopting commonly used models of reference dependence can lead to extreme policy implications. In particular, taking our results at face value would imply that the NRA should be set as high as possible. However, there are several factors that can limit such extreme policy recommendations. Most directly in line with our theory, we show that extreme reference points are not necessarily optimal when allowing for the more general structure of reference dependence represented by our flexible reduced-form formulation. For instance, in the retirement context there could be reference dependence over consumption in addition to reference dependence over leisure (Crawford and Meng, 2011; Behaghel and Blau, 2012). Such a two-dimensional framework is one of the canonical cases where our theory shows that the sign of key welfare effects can change. When we allow for two-dimensional reference dependence in the empirical application, we find that it remains locally optimal to increase the NRA under

our preferred parameter estimates. However, if consumption reference dependence was sufficiently strong, the welfare effects of increasing the NRA would turn negative at some point. In addition, we discuss how two further considerations may prevent governments from increasing the NRA to extremely high levels, even without deviating from simple loss aversion. First, pension reforms implemented in practice often combine NRA increases with benefit cuts because of a linkage between the NRA and a "full" pension benefit level. We show that such a reform can reduce private welfare for the average worker, even when the overall social welfare effect of the pension reform is strongly positive. This can make NRA increases unpopular and difficult to implement. Second, if governments were to try and set an extremely high NRA, issues of credibility may arise, and in turn only modest, gradual NRA increases may retain the effectiveness of such reforms.

This paper contributes to the literature on behavioral welfare economics, reviewed by Bernheim and Taubinsky (2018). To our knowledge, we provide the first welfare analysis under reference-dependent preferences, one of the most prominent models in behavioral economics. Our characterization of welfare effects in terms of direct and behavioral effects and sufficient statistics is closely related to existing work studying welfare in the presence of other behavioral biases (e.g. Chetty et al., 2009; Mullainathan et al., 2012; Allcott and Taubinsky, 2015; Allcott et al., 2019; List et al., 2023). In contrast to most of the literature, we allow for normative ambiguity. This is crucial to making progress in settings like reference dependence, where such ambiguity has been recognized as a key obstacle to welfare analysis. Our notion of normative ambiguity builds on foundational work on behavioral revealed preferences by Bernheim and Rangel (2009), and on Goldin and Reck (2022), who use a similar approach to examine the welfare economics of default options.

We connect behavioral welfare economics with the rich literature on reference dependence itself, reviewed by O'Donoghue and Sprenger (2018). Seminal theoretical contributions on modeling reference-dependent preferences include Kahneman and Tversky (1979), Tversky and Kahneman (1991) and Kőszegi and Rabin (2006). A large number of studies document the empirical relevance of reference dependence for individual decision-making. Most closely related to our empirical application is the evidence from field settings described above (e.g. Camerer et al., 1997; DellaVigna et al., 2017; Homonoff, 2018; Rees-Jones, 2018). Existing work on reference dependence largely focuses on positive analysis of behavior and has so far refrained from formal welfare analysis. The main contribution of our paper is to provide this welfare analysis. Our results can be used to derive novel policy implications for the wide range of contexts where reference-dependent preferences have been shown to matter.

Our empirical application relates to a recent literature on retirement behavior, which documents the reference point character of statutory retirement ages (Behaghel and Blau, 2012; Seibold, 2021; Lalive et al., 2022; Gruber et al., 2022) and responses to financial retirement incentives (e.g. Brown, 2013; Manoli and Weber, 2016; Gelber et al., 2020; Duggan et al., 2023). Applying our theoretical findings to the retirement context complements recent approaches to the welfare effects of pension reforms (Haller, 2022; Kolsrud et al., 2023). In particular, we consider how incorporating reference dependence, which is important in explaining real-world retirement behavior, shapes these welfare effects.

The remainder of this paper proceeds as follows. Section 2 introduces the model and characterizes individual welfare, Section 3 turns to the sufficient statistics for social welfare, Section 4 presents the empirical application to retirement behavior, and Section 5 concludes.

# 2 Individual Welfare, Reference Points, and Prices

In this section, we characterize the individual welfare effects of changes in reference points and of price changes. We show how these generally depend on two key factors: the form of reference-dependent payoffs, and whether reference-dependent payoffs are judged to be normative or behavioral. We then examine which properties of reference-dependent payoffs are sufficient to pin down first-order welfare effects, and we consider how these are shaped by various functional form restrictions from prior literature.

## 2.1 Setup

**Behavior.** Our goal is to understand the welfare of an individual with reference-dependent preferences. The individual chooses a good $x \in \mathbb{R}$ and a background good $y \in \mathbb{R}$ subject to a linear budget constraint with income $z$. The exogenous price of $x$ is $p$, and the price of good $y$ is normalized to 1. The individual maximizes a utility function $U(x, y)$, consisting of quasi-linear utility over $x$ and $y$ plus a reference-dependent payoff from consuming $x$ with an exogenous reference point $r \in \mathbb{R}$.

$$\begin{aligned} &\max_{x,y} \; U(x, y, r) = u(x) + y + v(x, r) \\ &\text{subject to } px + y = z. \end{aligned} \tag{1}$$

We label $U(x, y)$ *decision utility* because it generates behavior; this may be distinct from *normative utility* or welfare (Kahneman et al., 1997; Bernheim and Taubinsky, 2018). Following prior work on reference dependence, we call $u(x) + y$ *intrinsic utility*: utility conferred by a consumption bundle separately from any reference-dependent payoffs. Intrinsic utility represents the choices the individual would make in a counterfactual and potentially unobserved frame where reference-dependent payoffs are not present. We assume an interior solution, and that $u' > 0$ and $u'' < 0$ everywhere.

One key difficulty our analysis faces is that the literature has proposed many different functional forms of reference-dependent payoffs $v(x, r)$. The formulation of payoffs often prioritizes explaining particular moments of observed behavior, while maintaining tractability (see O'Donoghue and Sprenger, 2018). However, the welfare consequences of different formulations of reference dependence have not yet been systematically explored. In order to do this, we begin by putting minimal structure on reference-dependent payoffs. We then characterize the features of $v(x, r)$ that shape key welfare effects.

**Assumption 1.** *There are two real functions $\mu$ and $\nu$ such that*

$$v(x, r) = \nu(\mu(x) - \mu(r)), \text{ and} \tag{2}$$

1. *(Differentiability) $\mu(z)$ is everywhere differentiable with $\mu' > 0$; $\nu(z)$ is everywhere continuous, and everywhere differentiable except at $z = 0$,*
2. *(Gain-Loss Utility) $\nu(0) = 0$,*
3. *(Loss Aversion) The left and right derivatives of $\nu(z)$ at $z = 0$ exist, with $\nu'_-(0) > \nu'_+(0)$.*

**Assumption 2.**

1. *(Sub-Domain Monotonicity) $\nu(z)$ is monotone over the domain $(-\infty, 0)$ and over the domain $(0, \infty)$.*

2. *(No Diminishing Sensitivity)* $\nu''(z) = 0$ *for any* $z \neq 0$.

Assumption 1 encompasses the key features held in common by virtually all formulations of reference-dependent payoffs considered in the literature. The individual evaluates an amount of good $x$ according to the gain or loss relative to the reference point $r$. The function $\mu$ governs the units of gains and losses. Prior work considers either gain-loss utility over amounts of the good ($\mu(z) = z$) (Tversky and Kahneman, 1991), or gain-loss utility over utils ($\mu(z) = u(z)$) (Kőszegi and Rabin, 2006). The function $\nu$ determines how much the size of the gain or loss affects the individual's willingness to pay for good $x$. Assumption 1.2 requires that the reference-dependent payoff is zero when there is no gain or loss. Assumption 1.3 captures the empirical phenomenon of *loss aversion*: around the reference point, willingness to pay for $x$ is higher in the loss domain ($x < r$) than in the gain domain ($x > r$). This implies a kink in the utility function, i.e. a discontinuous change in marginal utility, at $x = r$.[3]

Assumption 2 puts additional structure on reference-dependent payoffs, which is not necessary for our general characterization of welfare effects but will help us sign key welfare effects later on. Prior literature typically assumes that payoffs are monotonic over gains and losses globally and that individuals prefer to increase gains and shrink losses (i.e. $\nu' \geq 0$ everywhere). We are slightly more flexible and require in Assumption 2.1 that payoffs are monotonic in the gain and loss domain separately; i.e. $\nu'$ is not necessarily same-signed for $z > 0$ and $z < 0$. This allows us to nest a wider class of potential payoff formulations consistent with behavioral data, including multi-dimensional reference dependence. In Assumption 2.2, we follow much of the literature on reference dependence in deterministic environments and rule out "diminishing sensitivity," which would require $\nu(z)'' < 0$ for $z > 0$ and $\nu''(z) > 0$ for $z < 0$. Diminishing sensitivity is sometimes considered in work on choice under uncertainty, but there is little empirical support for it in the deterministic case (see O'Donoghue and Sprenger, 2018). Nevertheless, we characterize individual welfare under diminishing sensitivity in Appendix B.5. Introducing diminishing sensitivity has no first-order impact on welfare when individuals choose options near the reference point, while it can modify welfare effects under more extreme gains or losses.

**Welfare.** Another key challenge for welfare analysis is, as O'Donoghue and Sprenger (2018) put it, "the question whether gain-loss utility should be given normative weight." We view the answer to this question as a normative judgment the social planner has to make about whether to respect reference-dependent payoffs or regard reference dependence as a bias. In particular, we parameterize the planner's judgment by $\pi \in \{0, 1\}$, and express normative utility as

$$U^*(x, y) = u(x) + y + \pi v(x, r). \tag{3}$$

We denote indirect utility at a given price and reference point by $w(p, r) \equiv U^*(x(p, r), y(p, r))$.

How does our welfare approach relate to revealed preferences? If $\pi = 1$, reference dependence is judged to be normative, $U = U^*$, and the individual's revealed preferences correspond to welfare. If instead $\pi = 0$, reference dependence is judged as a bias, and equation (3) defines a utility function that represents the choices the individual should make in order to maximize welfare. In this case, normative utility function

[3] One alternative type of reference dependence we do not consider here is a utility notch (a discontinuity in the level of utility) rather than a utility kink (a discontinuity in marginal utility) at the reference point (see e.g. Allen et al., 2017). The kink formulation is by far the most commonly adopted one in the literature, and it is better in line with much of the existing empirical evidence. For instance, the retirement age distribution shown in Figure 1 does not exhibit any "missing mass" around the NRA, as would be predicted under a utility notch (see the discussion in Seibold, 2021).

equals intrinsic utility without including reference-dependent payoffs.[4] As welfare under quasi-linearity is money metric by construction, we can directly compare individual welfare under $\pi = 1$, welfare under $\pi = 0$, and intrinsic utility. A key statistic for welfare effects will be *intrinsic willingness to pay*, which simply equals $u'(x)$ in our framework.

This welfare approach to reference dependence is closely related to adaptations of revealed preference analysis in other behavioral settings where individual choices do not necessarily reveal normative preferences (e.g. Chetty et al., 2009; Allcott and Taubinsky, 2015; Allcott et al., 2019; Goldin and Reck, 2020). Our notion of welfare with normative ambiguity encoded in $\pi$ is most similar to Bernheim et al. (2015) and Goldin and Reck (2022) who allow for ambiguity about whether default effects reflect biases or normative preferences. We discuss the relationship of our analysis to the foundational work on revealed preference analysis with behavioral frictions by Bernheim and Rangel (2009) in Appendix D.

## 2.2 Main Individual Welfare Effects

### 2.2.1 Welfare Effects of Changing Reference Points

**Can Policy Influence Reference Points?** We begin by studying the welfare effect of a change in the reference point. Our motivation stems from mounting empirical evidence that reference points can be influenced by policy across a number of contexts. Later on, our empirical application builds on Seibold (2021) who shows that reference dependence explains why many individuals retire precisely at *statutory retirement ages* even in the absence of financial incentives to do so. Moreover, Seibold (2021) documents that reforms to statutory retirement ages have strong behavioral effects, shifting the bunching mass in the retirement age distribution toward the new statutory age, which suggests that such reforms shift individual reference points. Another key example is Rees-Jones (2018), who finds that the distribution of tax liabilities due at the time of tax filing exhibits sharp bunching at zero and missing density above zero. This suggests that income tax withholding creates an arbitrary reference point at zero tax due, with associated loss aversion over the tax liability at the time of filing. Moreover, evidence from laboratory experiments suggests that experimental treatments can shift reference points: for instance, a prominent experiment by Kahneman et al. (1990) shows that individuals' valuation of a good (a coffee mug) depends on whether subjects are told they already own the mug at the start of the experiment. This "endowment effect" suggests that the treatment changes the reference point toward a consumption bundle including the mug.[5]

Given all the evidence that reference points are influenced by policy, we argue that characterizing the welfare effects of changing reference points is valuable. Naturally, real-world policy changes might affect more than just $r$. For example, reforms that shift the Normal Retirement Age (NRA) typically change both reference points and, via an institutional linkage between the NRA and pension benefit schedules, individual budget constraints. For any policy $P$ that affects a reference point $r(P)$, we can express the welfare effect of this policy as $\frac{dW}{dP} = \frac{\partial W}{\partial P} + \frac{\partial W}{\partial r}\frac{\partial r}{\partial P}$. Our theory characterizes the welfare effect of the change in reference point itself, that is $\frac{\partial W}{\partial r}$. Because other potential effects on welfare contained in $\frac{\partial W}{\partial P}$ will depend on the specific policy under consideration, our view is that they are best dealt with in applications. For instance, in our empirical application we are able to quantify the total welfare effects of real-world pension reforms including the linkage to benefits described above. Similarly, while empirical evidence shows that

[4] Note that when $x = r$, normative utility and decision utility coincide for any $\pi$. This is a consequence of Assumption 1.2, which together with equation (3) rules out any other behavioral biases.

[5] There are two potential reasons why willingness to accept may be greater than willingness to pay in such contexts: loss aversion over goods like the mug, and loss aversion over money. We discuss this further in Section 2.3. See O'Donoghue and Sprenger (2018) for a more detailed summary of the experimental literature on reference dependence.

reference points can be influenced by policy, there is little existing guidance on the magnitude of $\frac{\partial r}{\partial P}$, which will likely vary across applications. Thus, our theoretical approach does not rely on any specific magnitude of this term, and it can be applied to any environment where there is a policy that affects reference points by some amount.

**General Characterization: Direct and Behavioral Welfare Effects.** With reference-dependent preferences, individual choices can fall into three *domains*. We call the range of prices and reference points under which $x < r$ the loss domain $L \equiv \{(p,r)|x(p,r) < r\}$, and we similarly define the gain domain $G \equiv \{(p,r)|x(p,r) > r\}$ and the reference domain $R \equiv \{(p,r)|x(p,r) = r\}$. Note that because of the kink in decision utility at $x = r$, the $R$ domain has positive measure, which matches the stylized empirical fact of bunching at reference points (Allen et al., 2017; Rees-Jones, 2018; Seibold, 2021).

Outside the $R$ domain, we can apply the envelope theorem to the consumer maximization problem from equation (1) in order to characterize the welfare effect of a change in $r$. This yields

$$(p,r) \notin R \implies w_r = \underbrace{-(1-\pi)v_x x_r}_{\text{Behavioral Effect}} + \underbrace{\pi v_r}_{\text{Direct Effect}}, \tag{4}$$

where the partial derivatives $v_x$ and $v_r$ are evaluated at observed demand. For the $\pi = 1$ case (reference dependence is normative), equation (4) is the classic envelope result: the first-order welfare effect of a change in $r$ includes only its direct effect on utility, without regard to behavioral responses (Milgrom and Segal, 2002). Intuitively, this direct effect arises because the reference-dependent gain or loss in $v(x,r)$ changes with $r$, holding behavior $x$ fixed. In contrast, when $\pi = 0$ (reference dependence is a bias), there is no direct welfare effect of $r$ because reference-dependent payoffs do not enter welfare. However, the envelope theorem no longer eliminates the effect of behavioral responses on welfare in this case. The behavioral welfare effect equals the impact of the change in $r$ on consumption of $x$ times the *marginal internality*, which is defined as the (money-metric) welfare effect a marginal change in $x$ along the budget constraint, $\frac{dU^*(x,z-px)}{dx}\Big|_{x=x(p,r)}$ (Mullainathan et al., 2012; Allcott and Taubinsky, 2015; Allcott and Kessler, 2019). In our context, the marginal internality equals $-(1-\pi)v_x$, reflecting the consumption distortion due to reference-dependent payoffs present under $\pi = 0$.

An important issue in characterizing welfare effects under reference dependence is that the above result fails to obtain in the $R$ domain. The problem occurs because of the non-differentiability in $v$ at $x(p,r) = r$. Using Assumption 1.2, however, we can derive a similar characterization for the $R$ case:

$$\begin{aligned} v^R(x,r) &\equiv (1-\pi)U(x, z-px) + \pi U(r, z-pr) \\ \text{By Assumption 1.2, } (p,r) \in R &\implies w(p,r) = v^R(x(p,r),r) \\ (p,r) \in \text{Int } R &\implies w_r = \underbrace{(1-\pi)v^R_x x_r}_{\text{Behavioral Effect}} + \underbrace{\pi v^R_r}_{\text{Direct Effect}} \end{aligned} \tag{5}$$

When we evaluate the derivatives, the expression simply becomes

$$(p,r) \in \text{Int } R \implies w_r = u'(r) - p. \tag{6}$$

One could derive equation (6) more directly based on Assumption 1.2, but equation (5) helps us to interpret the overall effect in terms of direct versus behavioral effects. To understand this interpretation, suppose the individual consumes at the reference point because they are trying to avoid incurring a loss. In the interior of

the $R$ domain, this represents a corner solution, whereby increasing $r$ induces them to consume even more of good $x$ to avoid further losses. When $\pi = 0$, this occurs due to bias, and the resulting behavioral welfare effect equals the behavioral response times the marginal internality as in equation (4). Note that in the $R$ domain, $x_r = 1$ and the marginal internality is $\frac{dU^*(x,z-px)}{dx}\Big|_{x=x(p,r)} = (1-\pi)[u'(r) - p] = v_x^R(x(p,r),r)$. On the other hand, when $\pi = 1$, the welfare effect of increasing $r$ occurs due to moving the location of a corner solution for those optimally choosing to consume at the corner, which is a direct welfare effect (cf. Moore, 2022).

Hence, equation (5) allows us to characterize the welfare effect of changing $r$ analogously to equation (4), where the direct effect matters for welfare under $\pi = 1$, while the behavioral effect matters under $\pi = 0$. Since welfare effects are pinned down by the derivatives of $v(x,r)$ outside the $R$ domain, it is useful to compare welfare effects in the $R$ domain to these derivatives. We find that

$$(p,r) \in \text{Int } R \implies v_x(r^-,r) < v_x^R(r,r) < v_x(r^+,r), \tag{7}$$

where $v_x(r^+,r) = \nu'_+(0)\mu'(r)$ is the derivative as $x$ approaches $r$ from the right, and $v_x(r^-,r)$ is the analogous left derivative.

Taken together, equations (4) and (6) provide a characterization of the first-order welfare effect of a shift in the reference point starting from any initial point.[6] To quantify this welfare effect, knowledge of the derivatives $v_x(x(p,r),r)$ and $v_r(x(p,r),r)$ and of $\pi$ is sufficient outside the $R$ domain. In the $R$ case, knowledge of the left and right derivatives of $v(x(p,r),r)$ with respect to $x$ around $x(p,r) = r$ would bound welfare according to (7); fully pinning down welfare in this domain requires knowledge of intrinsic marginal utility $u'(r)$.

**The Sign of Welfare Effects.** The general characterization of welfare effects above is valid under the weak requirements of Assumption 1. Next, we study how the sign of welfare effects depends on the formulation of reference-dependent payoffs $v(x,r)$. Assumption 2 puts some additional structure on the derivatives of this payoff function that helps us sign welfare effects in a straightforward way.

**Lemma 1.** *Under Assumptions 1 and 2.1, at least one of the following must be true:*

- *(Everywhere Increasing)* $v_x \geq 0$ *for all* $x \neq r$.
- *(Everywhere Decreasing)* $v_x \leq 0$ *for all* $x \neq r$.
- *(Single-Peaked)* $v_x \geq 0$ *for all* $x < r$, *and* $v_x \leq 0$ *for all* $x > r$.

In words, Assumptions 1 and 2.1 admit three possibilities for the sign of the derivative $v_x(x,r)$: reference-dependent payoffs are either everywhere increasing in $x$, or everywhere decreasing, or have a maximum at the reference point.

**Proposition 1.** ***Signing the Welfare Effects of Reference Point Variation.***
*Maintain Assumptions 1 and 2 and consider any* $(p,r)$ *that is not on the boundary of* $R$.

**P1.1.** *If* $v$ *is Everywhere Increasing, then* $w_r(p,r) \leq 0$. *If* $v$ *is Everywhere Decreasing, then* $w_r(p,r) \geq 0$.

[6] Technically, welfare effects in the boundary of the set $R$ are not covered by this characterization. In these cases, existence of the derivative $w_r$ is not assured without further assumptions. However, the boundary of $R$ is measure zero and behavior and welfare are continuous around the boundary. Thus, our main welfare results are unaffected by this limitation.

**P1.2.** *Let $r^*$ be the reference point such that $u'(r^*) = p$. If $v$ is Single-Peaked, then $w_r(p, r) \geq 0$ when $r \leq r^*$, and $w_r(p, r) \leq 0$ when $r \geq r^*$. Consequently, $r^*$ is an individually optimal reference point.*

Proposition 1 contains two key results. First, the sign of the welfare effect $w_r$ depends on which of the three cases from Lemma 1 obtains. Proposition 1.1 states that when reference-dependent payoffs increase willingness to pay for good $x$ everywhere, then lowering the reference point weakly increases welfare. When the payoffs decrease the willingness to pay for $x$, on the other hand, a higher reference point weakly increases welfare. As a consequence, in the Everywhere Increasing case, the lowest possible reference point would be individually optimal; in the Everywhere Decreasing case, the highest possible reference point would be individually optimal. In the third case, where the payoff increases willingness to pay for $x$ in the loss domain and decreases it in the gain domain, we find that welfare is maximized where the reference point equals the *intrinsic optimum*, which we define as the optimal choice of $x$ according to intrinsic utility. Proofs of these and all other theoretical results are in Appendix C.

Second, Proposition 1 implies that the sign of the welfare effect $w_r$ does not depend on the judgment about whether reference dependence is normative or a bias, i.e. the choice of $\pi$. This result is perhaps surprising, given that the literature on reference dependence so far has considered ambiguity over $\pi$ as an important obstacle for welfare analysis. Our analysis suggests that whether a change in the reference point increases or decreases welfare instead turns on the properties of reference-dependent payoffs, which in turn determine which of the cases from Lemma 1 apply. Nevertheless, we note that $\pi$ does affect the magnitude of welfare effects, and it determines whether welfare effects are driven by direct or behavioral effects.[7]

### 2.2.2 Welfare Effects of Price Changes

Empirical evidence suggests that reference dependence also matters for responses to price instruments, such as taxes (e.g. Homonoff, 2018). Under Assumption 1, we can characterize the individual welfare effect of a marginal price change in terms of direct and behavioral effects:

$$w_p = \underbrace{-x(p, r)}_{\text{Direct Effect (Roy)}} \underbrace{- (1 - \pi) v_x x_p}_{\text{Behavioral Effect}}. \tag{8}$$

Again, this expression holds everywhere apart from the boundary of the $R$ domain. When $\pi = 1$, equation (8) is Roy's identity (under quasi-linearity). A price increase has a negative direct welfare effect because, holding behavior fixed, the chosen consumption bundle becomes more expensive. Unlike reference point variation, whether a direct effect occurs does not depend on normative judgments because the direct effect of a price change operates through intrinsic utility rather than through reference-dependent payoffs. The behavioral welfare effect, on the other hand, only matters when $\pi = 0$. As before, the behavioral effect equals the behavioral response times the marginal internality $-(1 - \pi)v_x$.[8] Equation (8) implies that the sign of welfare effects of price change are generally ambiguous, as they depend both on the formulation of reference-dependent payoffs *and* on the normative judgment about reference dependence.

Our analysis of price changes builds on prior work on welfare in the presence of internalities (Mullainathan et al., 2012; Allcott and Taubinsky, 2015; Allcott et al., 2019). We apply these insights to a model of

[7] If Assumption 2.2 was relaxed, $\pi$ would also matter for the individually optional reference point. We show this in Appendix B.5 where we allow for diminishing sensitivity.

[8] Note that in the $R$ domain, $x_p = 0$ locally, so the behavioral welfare effect equals zero. An additional technical issue is that $v_x$ does not exist in the $R$ domain because of non-differentiability. In order to evaluate $w_p$ in this case, we can resort to a solution similar to equation (5), replacing $v_x$ with $v_x^R$ whenever $(p, r) \in R$. As $x_p = 0$ in any case, this issue is inconsequential for the magnitude of welfare effects.

reference dependence with a potential internality shaped by normative ambiguity. Our results also have implications for optimal corrective taxation. Equation (8) suggests that setting a marginal tax equal to $(1-\pi)v_x$ everywhere can fully correct any biases in individual decision-making. However, when $\pi = 1$ the optimal corrective tax is zero everywhere. Thus, whether reference-dependent behavior creates scope for corrective taxation depends critically on normative judgments.[9]

## 2.3 Reference-Dependent Payoff Formulations

Next, we analyze how the structure implied by various formulations of reference-dependent payoffs shapes key welfare effects. We consider a wide range of formulations considered in the literature on reference dependence, and we propose a new flexible reduced-form specification that encompasses the key properties of earlier formulations.

Figure 2 illustrates observed demand and intrinsic demand, which are critical for our welfare analysis, under selected formulations of reference-dependent payoffs. We provide a detailed analysis of all payoff formulations discussed in this section in Appendix B. In particular, Appendix Tables B1 and B2 summarize the key features for welfare, welfare effects, and individually optimal reference points for all formulations.

### 2.3.1 Formulations Used in Prior Literature

We begin with a payoff formulation often used in applications of reference dependence, which we call *Simple Loss Aversion*:

$$v(x,r) = 1\{x \leq r\}\Lambda(x-r). \tag{9}$$

Under Simple Loss Aversion, the individual receives a negative payoff in the loss domain, which increases in the size of the loss relative to $r$. The strength of loss aversion is governed by the parameter $\Lambda > 0$. With this formulation, $v_x = 1\{x < r\}\Lambda$. As Panel (a) of Figure 2 shows, observed demand coincides with intrinsic demand in the gain domain, but exceeds intrinsic demand by $\Lambda$ in the loss domain. If we regard reference dependence as a preference $(\pi = 1)$, this reflects a normative willingness to pay to avoid losses. If reference dependence is a bias $(\pi = 0)$, this reflects a marginal internality leading to over-consumption of $x$. Under $\pi = 1$, decreasing the reference point would exert a positive direct effect on welfare, while under $\pi = 0$ decreasing the reference point would mitigate over-consumption of $x$. With Simple Loss Aversion, the Everywhere Increasing case from Lemma 1 obtains. Thus, regardless of the normative judgment encoded in $\pi$, decreasing the reference point weakly increases welfare. However, the Single-Peaked Case also obtains, implying that the intrinsic optimum $r^*$ is individually optimal. In fact, the individual is indifferent between $r^*$ and any lower reference point, and hence the set of individually optimal reference points is $(-\infty, r^*]$.

In Panel (b) of Figure 2, we consider another commonly used formulation where, in addition to loss aversion, the individual also receives gain utility proportional to the size of the gain relative to $r$, which is governed by a parameter $\eta$ (Tversky and Kahneman, 1991).[10] In the case of loss aversion with gain utility, individuals in the gain domain can also experience welfare improvements from a lower reference point because decreasing the reference point increases their gains. Again, the Everywhere Increasing case applies, and thus decreasing the reference point increases welfare regardless of $\pi$. Under $\pi = 0$, any reference point low enough such that consumption falls into the gain domain is individually optimal, and under $\pi = 1$

[9] As an example, we derive the optimal corrective tax schedule under Simple Loss Aversion in Appendix B.1.

[10] Note that we slightly re-parameterize the reference-dependent payoff function for the case of loss aversion with gain utility relative to (Tversky and Kahneman, 1991). The reparameterized version is equivalent in terms of behavior and welfare, and more easily comparable to Simple Loss Aversion.

the lowest possible reference point is individually optimal. Unlike Simple Loss Aversion (nested by $\eta = 0$), with $\eta > 0$, the Single-Peaked case no longer obtains and the intrinsic optimum is a sub-optimal reference point.

Most empirical work on reference-dependent preferences adopts one of these two formulations. In many contexts, this is based on valid empirical or institutional considerations. For instance, our empirical application initially adopts a Simple Loss Aversion model, which can well explain observed empirical retirement patterns and is in line with the institutional framing of the Normal Retirement Age. However, our theoretical results reveal that these common formulations implicitly impose significant ex-ante structure on the direction of key welfare effects. In particular, decreasing reference points always improves welfare.[11] Our results highlight the importance of carefully considering other potential formulations which may carry different welfare implications.

The goal of many empirical studies of reference dependence is identifying the loss aversion parameter $\Lambda$. Panels (a) and (b) of Figure 2 confirm that $\Lambda$ governs the main behavioral phenomenon often associated with reference dependence, namely the extent to which behavior is modified in the loss versus the gain domain. Meanwhile, the gain utility parameter $\eta$ is not essential for predicting behavior and can generally not be identified separately from the parameters of intrinsic utility without further functional form restrictions. This is evident from Figure 2, where observed demand has identical properties in Panels (a) and (b), such that many combinations of $\eta$ and $u'(x)$ would be consistent with observed behavior. We prove this behavioral isomorphism between Simple Loss Aversion and loss aversion with gain utility in Appendix B.2 (see also Barseghyan et al., 2013).

**Two-Dimensional Reference Dependence.** Some of the theoretical literature postulates that reference dependence could be present over multiple goods (e.g. Tversky and Kahneman, 1991; Kőszegi and Rabin, 2006). In part due to difficulties of empirical identification, such multi-dimensional reference dependence is examined less often in applications. Two notable exceptions are Crawford and Meng (2011) and Behaghel and Blau (2012), who consider reference dependence over two goods, consumption and leisure. In an extension of our empirical application, we will allow for a similar type of two-dimensional reference dependence. Multi-dimensional loss aversion is also considered in classic experiments, which test for differences between willingness to accept ($WTA$) and willingness to pay ($WTP$) for a good and the equivalent monetary gain ($EG$) (Kahneman et al., 1990; Tversky and Kahneman, 1991; O'Donoghue and Sprenger, 2018). Experiments tend to find that the difference between $WTA$ and $EG$ exceeds the difference between $EG$ and $WTP$, which suggests that loss aversion over goods is the dominant form of reference dependence (rather than loss aversion over money).

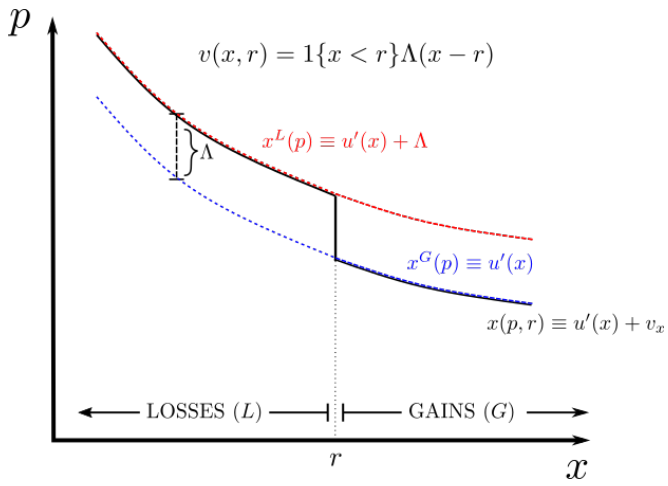
To capture such two-dimensional reference dependence, we suppose in Panel (c) of Figure 2 that the individual is loss averse over both $x$ and $y$, and the two-dimensional reference point lies on the budget constraint.[12] In this case, the gain domain for good $x$ corresponds to the loss domain for good $y$, and vice

[11] While welfare under reference dependence has not been explicitly analyzed in prior literature, the fact that adopting this payoff structure induces individuals to prefer lower reference points is part of the motivation by Kőszegi and Rabin (2006) to assume that the reference point is the expected intrinsic optimum (under rational expectations).

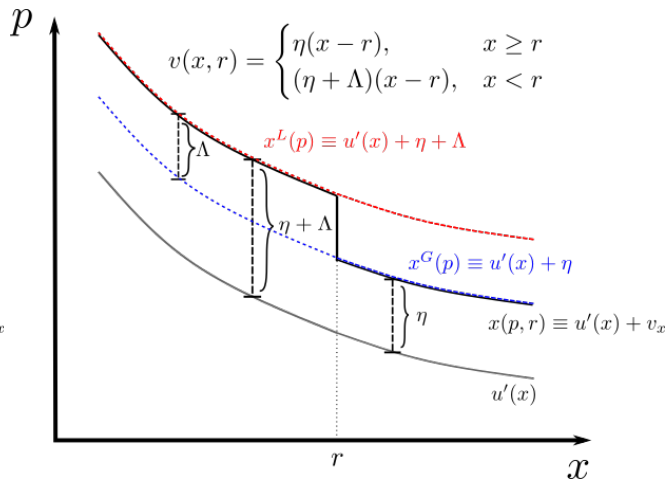
[12] Assuming that the reference point lies on the budget constraint is a useful way to discipline two-dimensional reference dependence, and this assumption is often empirically plausible. For instance, in our empirical application the Normal Retirement Age may function as a reference point both in terms of the retirement age and in terms of consumption. Note that this assumption is weaker than the central assumption about the origins of reference points in Kőszegi and Rabin (2006), which requires that the reference point equals the intrinsic optimum (i.e. $r = r^*$) in the deterministic case. If the two-dimensional reference point is not on the budget constraint, this will generate policy implications similar to one-dimensional reference dependence, such as lowering reference points in all dimensions separately. Also note that two-dimensional reference initially entails a slight abuse of our notation because we specify reference-dependent payoffs as a function of $x$ but not of $y$. This issue is resolved when we re-formulate the two-dimensional

**FIGURE 2: OBSERVED DEMAND AND INTRINSIC DEMAND FOR SELECTED PAYOFF FORMULATIONS**

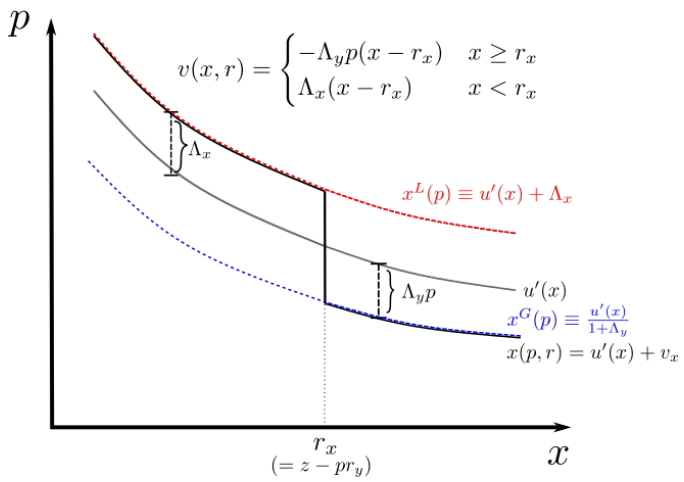
(a) Simple Loss Aversion



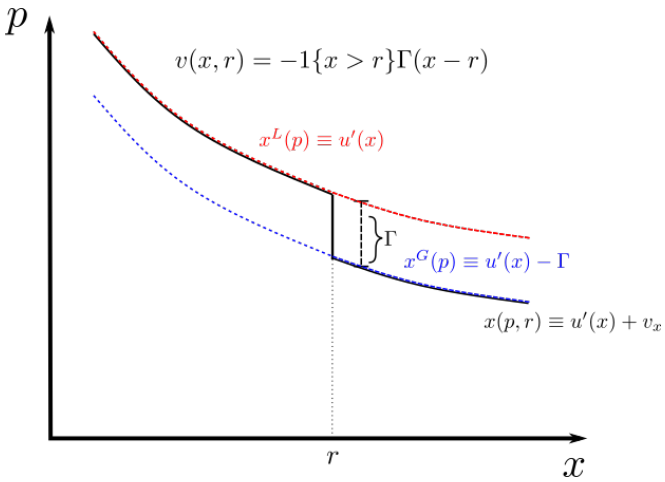

(b) Loss Aversion with Gain Utility



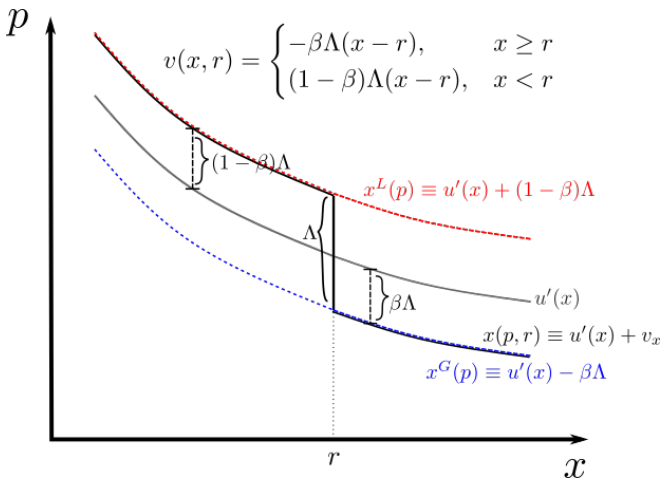

(c) Two-Dimensional Loss Aversion




(d) Gain Discounting




(e) Flexible Reduced Form




*Notes:* The figure illustrates observed and intrinsic demand under different formulations of reference-dependent payoffs. Each panel shows observed demand (in black), gain domain demand (in blue), loss domain demand (in red), and intrinsic demand for the payoff formulation indicated by the panel title. Intrinsic demand coincides with gain domain demand in Panel (a) and with loss domain demand in Panel (d), and is shown in grey in the other panels. The vertical differences between these demand curves are key for welfare analysis; in each panel we illustrate how these relate to parameters of the respective payoff formulation. Note that the reference dependence parameters $\Lambda$, $\eta$, $\Lambda_y$, $\Lambda_x$, and $\Gamma$ are all required to be positive.

versa. Hence, intrinsic demand for a good lies everywhere *between* observed gain and loss domain demand. This generates ambiguous welfare effects of changing the reference point, as opposite-signed direct and behavioral effects occur above and below reference point for good $x$. We return to this issue below.

**Loss Aversion or Gain Discounting?** Finally, a more exotic possibility which would satisfy Assumptions 1 and 2, but has not yet been considered in the literature, is that individuals are not loss averse but instead they discount gains. The literature usually assumes that the kink in revealed preferences at the reference point (Assumption 1.3) results from excess negative payoffs over losses, but in principle such a kink could be driven by discounting payoffs in the gain domain. As Panel (d) of Figure 2 shows, Gain Discounting can generate observed demand curves that look indistinguishable from Simple Loss Aversion. However, Gain Discounting features an inverse relationship between intrinsic demand and observed demand, and as a result the sign of key welfare effects would be reversed. In particular, increasing the reference point would increase welfare. This occurs because Gain Discounting falls into the Everywhere Decreasing case from Lemma 1. While Gain Discounting represents a valid theoretical possibility,[13] the common practice of interpreting reference dependence in terms of loss aversion is rooted in empirical evidence on the psychological and neurological origins of loss aversion.[14]

### 2.3.2 Flexible Reduced-Form Formulation

As the discussion above illustrates, key welfare effects depend on the formulation of reference-dependent payoffs one adopts. We next propose a *Flexible Reduced-Form* specification that encompasses the key properties of prior formulations in a parsimonious way, while imposing less ex-ante structure on welfare than formulations like Simple Loss Aversion. Introducing a new parameter $\beta \in [0, 1]$, we specify

$$v(x, r) = \begin{cases} (1-\beta)\Lambda(x-r), & x \leq r \\ -\beta\Lambda(x-r), & x > r. \end{cases} \tag{10}$$

The parameter $\Lambda$ governs the size of the utility kink implied by Assumption 1.3, similar to equation (9). The parameter $\beta$ can be interpreted as capturing the extent of deviation from Simple Loss Aversion, which is nested by $\beta = 0$. For instance, with two-dimensional reference dependence, $\beta$ reflects the relative magnitude of loss aversion over good $y$ versus over good $x$ with two-dimensional formulation. Moreover, $\beta$ can capture the relative importance of loss aversion versus gain discounting over $x$, where $\beta = 1$ corresponds to pure gain discounting.

More generally, the Flexible Reduced-Form can be thought of as a linear approximation of any payoff formulation that falls into the Single-Peaked case from Lemma 1. Note that this includes formulations

formulation as a function of $x$ only, as shown in Panel (c) of Figure 2. See Appendix B.6 for a detailed discussion of two-dimensional reference dependence.

[13] Under $\pi = 0$, the discussion about loss aversion vs. Gain Discounting can be viewed as an application of the behavioral revealed preference framework of Bernheim and Rangel (2009), where a key question is which "frame" reveals welfare-relevant choices. While loss aversion designates observed demand in the gain domain as normative, Gain Discounting would instead designate observed demand in the loss domain as normative.

[14] The psychological literature suggests that loss aversion has emotional origins (see Rick (2011) for a review). For instance, an influential study by Kermer et al. (2006) finds that loss aversion derives from an affective forecasting error: people wrongly project that they will experience pain if they incur a loss, so they try to avoid losses. More recent evidence suggests that the pain of incurring losses is real rather than a forecasting error, and that emotion-regulation strategies mitigate loss aversion (Sokol-Hessner et al., 2009). This idea is further borne out by neurological research associating activity in the amygdala with loss aversion and the incursion of perceived losses (De Martino et al., 2010; Sokol-Hessner et al., 2013; Sokol-Hessner and Rutledge, 2019). In particular, the analysis of the neurological responses to the incursion of perceived gains and losses in Sokol-Hessner et al. (2013, Figure 4) provides direct evidence in favor of Loss Aversion over Gain Discounting.

falling into more than one case, such as Simple Loss Aversion and Gain Discounting. We formalize this approximation in Appendix B.7. Compared to other potential approximations spanning the same class of payoff formulations, equation (10) is the most tractable as it is linear in $x$ and $r$. As the length of the vertical segment in the demand curve is pinned down by the size of the kink in $v(x, r)$, the approximation is perfectly accurate close to the reference point for every formulation satisfying our assumptions. This includes formulations with non-linearities in $v(x, r)$, such as a utils specification for $\mu(z)$ as suggested by Kőszegi and Rabin (2006), and diminishing sensitivity, which we analyze in detail in Appendices B.3 and B.5, respectively. Further away from the reference point, the Flexible Reduced-Form remains accurate for linear payoff formulations, but the approximation can become less acculate for non-linear formulation. Pinning down how much non-linearities matter for extreme gains or losses would be straightforward in theory, but quantifying these effects empirically would require stronger structural assumptions about the functional form of payoffs.[15]

In summary, a wide range of reference-dependent payoff formulations proposed by prior work are approximated by equation (10). The main exception is one-dimensional loss aversion with gain utility, (Panel (b) of Figure 2), as this formulation falls in the Everywhere Increasing but not the Single-Peaked case. We argue that it is appropriate to abstract from this specific formulation for two reasons.[16] First, as discussed above, loss aversion with gain utility is isomorphic in terms of behavior to Simple Loss Aversion, which is encompassed by the Flexible Reduced-Form formulation. However, the additional parameter $\eta$ is difficult to identify empirically, and there is little direct empirical support for gain utility. Second, other formulations with gain utility can be encompassed by the Flexible Reduced-Form formulation. In particular, we show in Appendix B.6 that two-dimensional loss aversion with gain utility falls in the Single-Peaked case, as long as gain domain payoffs in either dimension are not too strong.

**Welfare Analysis with the Flexible Reduced-Form Formulation.** Panel (e) of Figure 2 depicts observed demand and intrinsic demand under the payoff formulation from equation (10). From Proposition 1, we know that welfare effects depend on whether the reference point is above or below an individual's intrinsic optimum $r^*$. Thus, welfare effects differ across domains, into which individuals with different intrinsic optima sort. Figure 3 illustrates individual welfare effects of changes in the reference point and how these can be decomposed into the direct and behavioral effects from equations (4) and (5). Figure 4 shows a corresponding illustration for price changes and how these relate to the direct and behavioral effects from equation (8).

In the loss domain $L$, welfare effects of reference points and prices resemble the intuition we discussed above for commonly used loss aversion models. In Panel (a) of Figure 3, increasing $r$ increases utility losses if reference dependence is normative ($\pi = 1$), generating a negative direct welfare effect. There is a negative internality of size $-(1-\beta)\Lambda$, but a behavioral welfare effect does not arise in $L$ because there is no behavioral response to a change in $r$ in this case. In Panel (a) of Figure 4, a price increase always leads to a negative direct welfare effect. Additionally, a positive behavioral welfare effect arises under $\pi = 0$ because the individual initially over-consumes $x$, and increasing the price helps correct this internality. Similarly, in Panel (c) of Figure 3, an individual in the "upper part" (where $r > r^*$) of the reference domain $R$ experiences

[15] Our Flexible Reduced-Form formulation departs from Kőszegi and Rabin (2006) in that we do not assume that the strength of loss aversion is the same across dimensions, which would require $\beta = 0.5$ in equation (10). Our subsequent analysis suggests that such an assumption would place significant ex-ante structure on welfare.

[16] We could nest loss aversion with gain utility and other Everywhere Increasing formulations if we allowed $\beta < 0$, and we could nest Everywhere Decreasing formulations with $\beta > 1$. Propositions 2 and 3.1 would hold for these cases, but Proposition 3.2 requires $\beta \in [0, 1]$.

a negative direct welfare effect from increasing $r$ under $\pi = 1$. If $\pi = 0$, a welfare effect of the same magnitude arises, but this can now be interpreted as a behavioral effect (see equation (5)). In Panel (c) of Figure 4, only a negative direct welfare effect occurs in the $R$ domain, as a marginal price change does not change behavior in this case.

Under Simple Loss Aversion, i.e. setting $\beta = 0$ in equation (10), these would be the only welfare effects. Thus, decreasing reference points robustly improves welfare under Simple Loss Aversion, and depending on $\pi$, this is driven by direct or behavioral welfare effects. The welfare effect of a price change is ambiguous, though, as direct and behavioral welfare effects go in opposite directions. Compared to Simple Loss Aversion, allowing for a more flexible structure of reference dependence with $\beta > 0$ in equation (10) modifies these results. In particular, opposite-signed direct and behavioral effects occur in the gain domain $G$. In Panel (b) of Figure 3, a positive direct welfare effect arises from increasing the reference point under $\pi = 1$, and in Panel (d) a positive direct or behavioral welfare effect arises in the "lower part" (where $r < r^*$) of the $R$ domain. There is a positive marginal internality of size $\beta\Lambda > 0$ in $G$, and thus increasing the price generates a negative behavioral welfare effect in Panel (b) of Figure 4. Hence, under the Flexible Reduced-Form formulation, the sign of the total welfare effect of changing the reference point depends on $\beta$. The sign remains robust to the choice of $\pi$, though, as Proposition 1 shows is generally true. The welfare effect of price changes depends both on $\beta$ and $\pi$ under the Flexible Reduced-Form formulation.

# 3 Social Welfare

In many real-world contexts, the planner is constrained to set a uniform policy across all individuals. This is true for both reference points and price-based policies. For instance, in our empirical application the government has to set a Normal Retirement Age and financial retirement incentives applying to all workers. Without this constraint and if each individual's preferences were known, the planner could set individually optimal reference points and prices such that a first-best solution is attained. In this section, we confront the second-best problem and analyze the social welfare effects of changing reference points and prices faced by heterogeneous individuals.[17] We derive sufficient statistics formulas that express welfare effects in terms of estimable quantities.

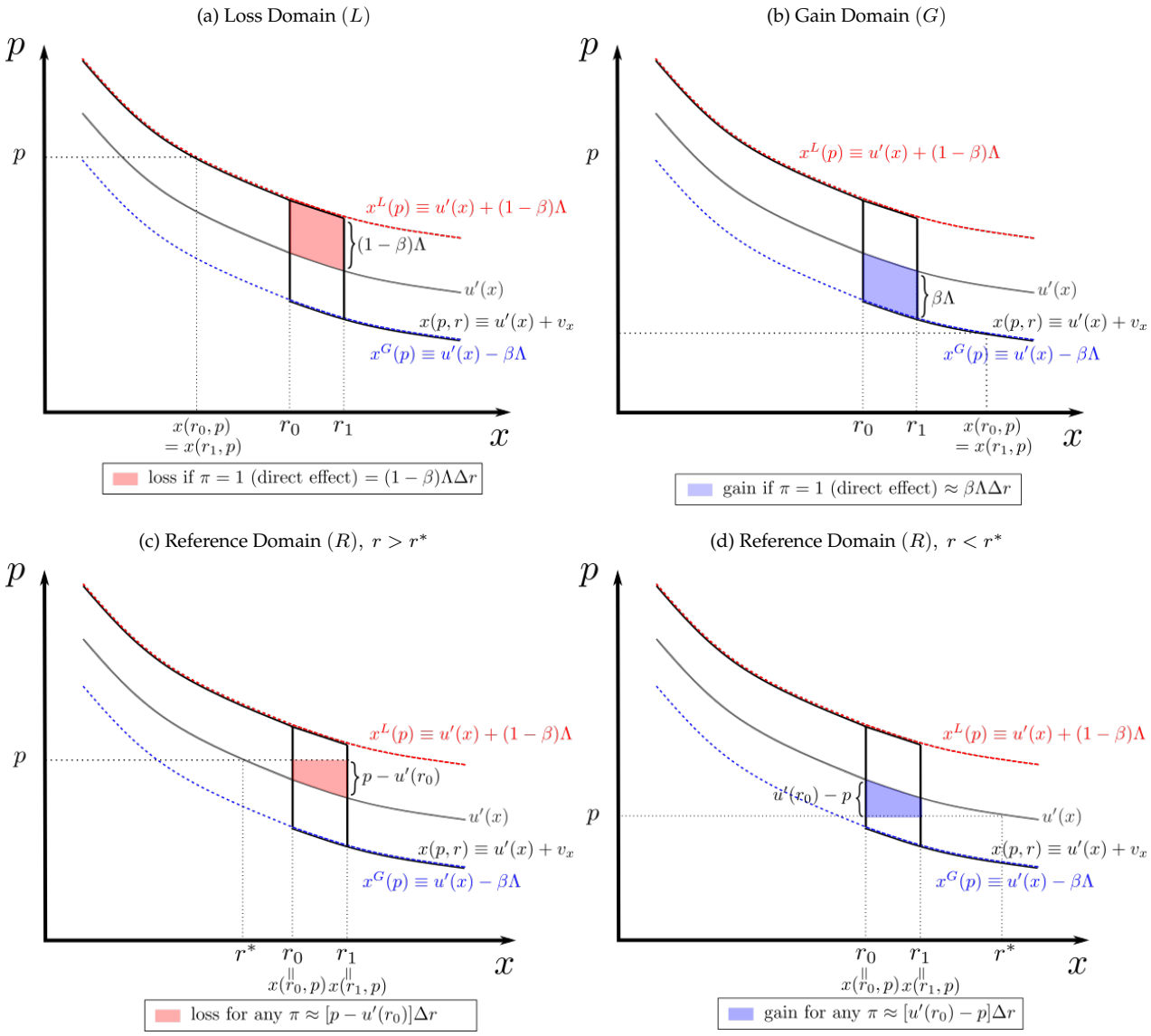
## 3.1 Setup

We consider a population of individuals of measure one. Individual behavior and welfare are characterized as above by equations (1) and (3), and we adopt the Flexible Reduced-Form of reference-dependent payoffs from equation (10). To capture heterogeneity, we suppose that each individual is characterized by preference parameters $i = (\theta_i, \Lambda_i, \beta_i)$, where the vector $\theta_i$ captures heterogeneity in the parameters of intrinsic utility. We write $u_i(x) \equiv u(x, \theta_i)$ and we use similar notation for other individual-specific functions. We specify Utilitarian social welfare as

$$W(p, r) \equiv \int_i w(p, r; \theta_i, \Lambda_i, \beta_i) dF_i(i) \equiv E[w_i(p, r)]. \tag{11}$$

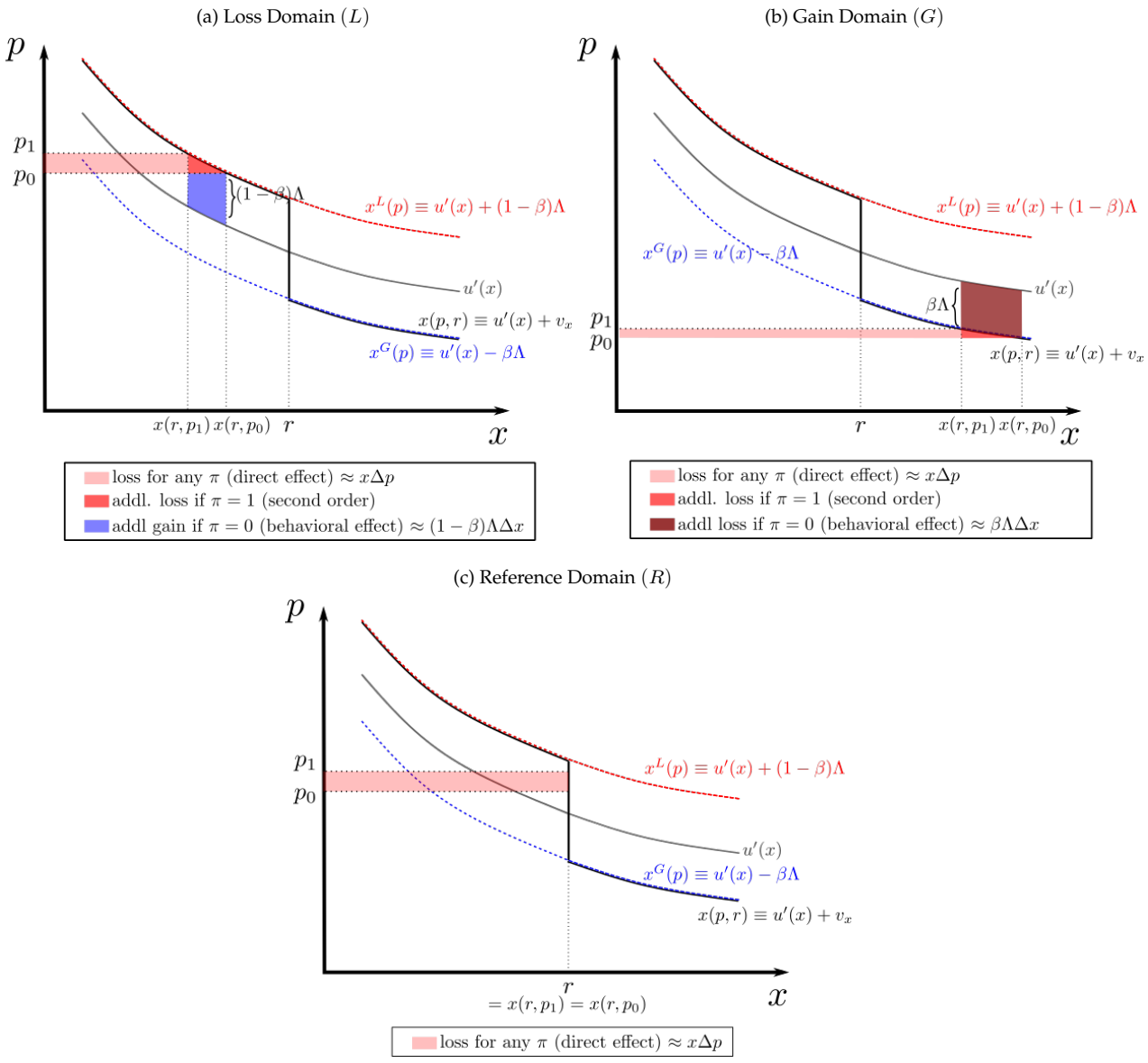
[17] A similar logic is encountered in other settings where the government has to set uniform policies for behavioral agents, such as optimal sin taxes (Allcott and Taubinsky, 2015) or optimal default options (Goldin and Reck, 2022).

**FIGURE 3: WELFARE EFFECTS OF CHANGING THE REFERENCE POINT**




*Notes:* The figure illustrates the welfare effects of changing the reference point under the Flexible Reduced-Form specification from equation (10), for individuals in the loss domain (Panel a), the gain domain (Panel b), and the reference domain (Panels c and d). For the reference domain, Panels (c) and (d) consider the two cases where the reference point is either below or above intrinsic optimum $r^*$. Each panel plots observed demand (in black) and gain domain demand (blue), loss domain demand (red), and intrinsic demand (grey). Red shaded areas denote welfare losses and blue shaded areas denote welfare gains. The legend of each panel provides further interpretation of the main welfare effects.

**FIGURE 4: WELFARE EFFECTS OF CHANGING PRICES**




*Notes:* The figure illustrates the welfare effects of changing the prices under the Flexible Reduced-Form specification from equation (10), for individuals in the loss domain (Panel a), the gain domain (Panel b), and the reference domain (Panel c). Each panel plots observed demand (in black) and gain domain demand (blue), loss domain demand (red), and intrinsic demand (grey). Red shaded areas denote welfare losses and blue shaded areas denote welfare gains. The legend of each panel provides further interpretation of the main welfare effects.

Using Utilitarian social welfare implies that we abstract from distributional concerns, which we defer to future work.[18] Henceforth we use the expectations operator for integrals like equation (11). It is also useful to introduce notation for the set of individuals in the three domains, which we call the $L$, $G$ and $R$ groups. We let $L(p,r) \equiv \{i|x_i(p,r) < r\}$ and define $G(p,r)$ and $R(p,r)$ analogously.[19] To economize on notation, we mostly suppress the $(p,r)$ inputs.

## 3.2 Main Social Welfare Effects

Building on our characterization of individual welfare in the $G$, $L$, and $R$ domains from the previous section, we can express social welfare by aggregating welfare across the $G$, $L$, and $R$ groups. Because individual welfare evolves continuously at the boundary of the three domains, any effects that arise when an individual transitions from one group to another are second-order (a marginal change in welfare for a marginal group of individuals). We provide the following sufficient statistics characterizations of the welfare effects of changes in reference points and prices:

**Proposition 2.** ***Sufficient Statistics Characterizations***

**P2.1.** *Up to a first-order approximation, the social welfare effect of a change in the reference point $\Delta r$ is*

$$\begin{aligned}\Delta W \approx \Delta r \pi \left\{E[\beta_i \Lambda_i \mid i \in G] P[i \in G] - E[(1-\beta_i)\Lambda_i \mid i \in L] P[i \in L]\right\} \\ - \Delta r E\left[u_i'(r) - p \;\middle|\; i \in R\right] P[i \in R].\end{aligned} \tag{12}$$

**P2.2.** *If the distribution of $u_i'(r) - p$ is independent of $(\beta_i, \Lambda_i)$ and locally uniform for $i \in R(p,r)$, the social welfare effect of $\Delta r$ can be further approximated as*

$$\begin{aligned}\Delta W \approx \Delta r \pi \left\{E[\beta_i \Lambda_i \mid i \in G] P[i \in G] - E[(1-\beta_i)\Lambda_i \mid i \in L] P[i \in L]\right\} \\ + \Delta r E\left[\Lambda_i \left(\beta_i - \frac{1}{2}\right) \;\middle|\; i \in R\right] P[i \in R].\end{aligned} \tag{13}$$

**P2.3.** *Up to a first-order approximation, the social welfare effect of a change in price $\Delta p$ is*

$$\begin{aligned}\Delta W &\approx \Delta p (1-\pi) \left\{E\left[\beta_i \Lambda_i x_{p,i} \;\middle|\; i \in G\right] P[i \in G] - E\left[(1-\beta_i)\Lambda_i x_{p,i} \;\middle|\; i \in L\right] P[i \in L]\right\} \\ &\quad - \Delta p E[x_i(p,r)] \qquad (14) \\ &= \Delta p (1-\pi) \left\{E\left[\beta_i \Lambda_i \varepsilon_i \frac{x_i}{p} \;\middle|\; i \in G\right] P[i \in G] - E\left[(1-\beta_i)\Lambda_i \varepsilon_i \frac{x_i}{p} \;\middle|\; i \in L\right] P[i \in L]\right\} \\ &\quad - \Delta p E[x_i], \qquad (15)\end{aligned}$$

*where $\varepsilon_i$ is the price elasticity of demand for good $x$.*

Proposition 2.1 characterizes the first-order social welfare effect of a change in $r$ under arbitrary heterogeneity in preference parameters. Equation (12) shows how normative ambiguity and payoff formulation matter for social welfare in the second-best scenario. The first two terms of the equation correspond to direct welfare effects arising for the $L$ and $G$ groups under $\pi = 1$. These effects depend on $\pi$, $\Lambda$, $\beta$ and the proportions of individuals in the $L$ and $G$ groups, $P[i \in L]$ and $P[i \in G]$. The last term captures the welfare

[18] For instance, one could straightforwardly incorporate distributional concerns by introducing social welfare weights (Saez and Stantcheva, 2016), as in Allcott et al. (2019) and Kolsrud et al. (2023).

[19] To be clear, the $L$ domain is a set of prices and reference points at which an individual chooses $x(p,r) < r$. With individual heterogeneity this domain is individual-specific. The $L$ group is the set of individuals choosing $x(p,r) < r$ at a given price and reference point.

effect for the $R$ group, whose magnitude is independent of $\pi$. This term is more difficult to characterize because $u_i'(r)$ is generally unobserved. Thus, proposition 2.2 provides a further approximation assuming a locally uniform distribution of intrinsic willingness to pay across individuals in the $R$ group.[20] In this case, the mean of $u_i'(r) - p$ in the $R$ group falls halfway between the bounds from equation (7), that is at $\Lambda_i(\beta_i - \frac{1}{2})$. With this approximation, $\pi$, $\Lambda$, $\beta$, $P[i \in L]$, and $P[i \in G]$ are the sufficient statistics for the total social welfare effect of changing the reference point. Setting equation (13) equal to zero yields a first-order condition for a second-best optimal $r$.

In the case of Simple Loss Aversion, which is commonly adopted in applications, the welfare effect of changing the reference point is unambiguously signed. Setting $\beta = 0$ in equation (13), we find $\Delta W / \Delta r \geq 0$, and the inequality is strict as long as there are some individuals in the $L$ or $R$ group. In fact, lowering the reference point generates a Pareto improvement under Simple Loss Aversion, and the planner could achieve the first-best outcome by setting $r \leq \min_i r_i^*$.

Proposition 2.3 provides a sufficient statistics characterization of the social welfare effects of a price change. Compared to equation (14), equation (15) re-casts these effects in terms of the price elasticity of demand for $x$. The first two terms capture behavioral welfare effects due to internalities materializing in the $\pi = 0$ case. Behavioral effects carry opposite signs for the $L$ and $G$ groups, as individuals in $L$ tend to over-consume $x$ while those in $G$ tend to under-consume $x$. The last term of the equation captures the negative direct welfare effect of a price change experienced by all individuals regardless of $\pi$.

In sum, these sufficient statistics formulas capture the welfare effect of reference points and prices in terms of estimable objects. The reference dependence parameter $\Lambda$ is the main estimation target of much of the empirical literature on reference dependence, and demand elasticities $\varepsilon$ are commonly estimated. Moreover, the relative sizes of the $L$, $G$ and $R$ groups can simply be measured in cross-sectional data on individual outcomes. For the parameter $\beta$, there are two possible ways forward. First, one could impose a particular value, for instance if Simple Loss Aversion is the most plausible model of reference dependence in a setting, this would imply $\beta = 0$. Second, one could try to empirically identify $\beta$. In Section 3.3, we show that bunching methods can be used for this purpose. Finally, the parameter $\pi$ of course remains a normative judgment.

**Externalities.** In some settings, including our empirical application, a fiscal or another externality is present and should be incorporated into welfare calculations. For a linear externality valued at $\alpha$ on the margin, the welfare effect of a change in the reference point or a price change is simply the relevant effect from Proposition 2 plus $\alpha E[\Delta x_i]$, that is the marginal externality times the change in demand for good $x$ resulting from the change in $r$ or $p$. We illustrate the derivation of sufficient statistics formulas with a fiscal externality in Appendix C.

### 3.3 Bunching at the Reference Point and Sufficient Statistics for Welfare

Our results above demonstrate that the reference dependence parameters $\Lambda$ and $\beta$ are key sufficient statistics for welfare analysis. We next examine how bunching designs, which are commonly used in empirical studies of reference dependence, can identify these parameters.

[20] Similar local uniformity assumptions are commonly adopted in bunching estimation (Kleven, 2016). This is closely related to our approximation because the $R$ group correspond to the set of bunchers at the reference point, and empirical identification of reference dependence parameters often relies on bunching methods.

**Proposition 3.** ***Identification from Bunching.*** *Define a random variable $\Delta_i = u_i'(r) - p$ and denote its density and cumulative distribution by $f_\Delta$ and $F_\Delta$. Assume that $\Delta$, $\Lambda$ and $\beta$ are mutually independent.*

**P3.1.** *Excess bunching at $x_i = r$ is characterized approximately by*

$$\frac{P[i \in R]}{f_\Delta(0)} \approx E[\Lambda_i] \tag{16}$$

**P 3.2.** *The share of bunching that comes from the right – defined as the share of individuals who would choose to consume more than $r$ in the absence of reference-dependent payoffs – is approximately*

$$P[r_i^* > r | i \in R] \approx E[\beta_i] \tag{17}$$

*These approximations are based on a first-order Taylor approximation of $F_\Delta$ around $\Delta = 0$; they are exact when $f_\Delta$ is locally uniform for $i \in R(p, r)$.*

Proposition 3.1 extends the well-known result that the size of the kink in preferences due to reference dependence, given by $\Lambda$ in our notation, can be identified from observed bunching at the reference point. This insight is exploited by a number of empirical studies quantifying simple loss aversion models, including Allen et al. (2017), Rees-Jones (2018) and Seibold (2021). Proposition 3.1 extends the result to all reference-dependent payoff formulations encompassed by our Flexible Reduced-Form formulation. Note that we characterize bunching in terms of willingness to pay for good $x$, whereby the density of intrinsic willingness to pay plays the role of the counterfactual distribution. Intuitively, $f_\Delta(0)$ reflects how many individuals would choose $x = r$ in the absence of reference-dependent payoffs, and relating the actual share of individuals at the reference point to this counterfactual yields our measure of excess bunching. This result can then be mapped onto bunching in terms of observed choices $x$, which can be readily estimated using standard bunching methods, as we demonstrate in our empirical application.[21]

Proposition 3.2 is a novel result for the literature on reference dependence. The proposition states that the direction of bunching pins down the parameter $\beta$, the other key sufficient statistic for welfare. In equation (17), the right bunching share corresponds to the fraction of individuals whose intrinsic optimum exceeds the reference point ($r_i^* > r$). The right bunching share thus carries information about the direction in which reference dependence modifies behavior relative to intrinsic utility. For instance, if we interpret $\beta$ as the relative strength of loss aversion over good $y$ as opposed to good $x$, then the right bunching share identifies the relative strength of loss aversion in either dimension. Note that empirically, the right bunching share is more challenging to identify than the amount of bunching at a reference point itself. As we show in the empirical application, identifying $\beta$ is possible under some additional assumptions about the counterfactual distribution of choices, but such assumptions might be subject to well-known critiques of bunching methods (Blomquist et al., 2021).[22]

[21] Note that when bunching is defined in terms of $x$, the curvature of $u(x)$ additionally matters. We can show that for a sufficiently small and homogenous $\Lambda$ and a homogenous price elasticity $\varepsilon$, excess bunching in the distribution of $x$ is approximately

$$\frac{P[i \in R]}{f_{r^*}(r)} \approx \varepsilon \frac{\Lambda}{p},$$

where $f_{r^*}$ is the distribution of intrinsic optima. Thus, given an estimate of $\varepsilon$, observed bunching in $x$ can identify $\Lambda$ as a fraction of the price. This result is closely related to how we estimate reference dependence in retirement behavior based on observed bunching of retirement ages in Section 4.2.

[22] There are potential alternative approaches to identifying $\beta$. For example, one could try to directly measure individual choices in the absence of reference dependent payoffs using methods developed by the literature on behavioral welfare economics (Chetty, 2009; Allcott and Taubinsky, 2015; Allcott et al., 2019; Goldin and Reck, 2020). The approach we propose here is simple and closely related to bunching methods commonly used to analyze reference dependence in field settings. We also note that the simplicity of equations

# 4 Empirical Application: Reference Dependence in Retirement Behavior

In this section, we present an empirical application of our theoretical results. Retirement behavior is one of the most important contexts in which reference-dependent preferences have recently been documented. Our empirical setting is that of Seibold (2021), who finds large bunching in the retirement distribution around *statutory retirement ages* in Germany and argues that this phenomenon can be explained by workers perceiving those ages as reference points. In this context, our goal is to characterize the welfare effects of reforms to the Normal Retirement Age, and reforms to financial incentives for delayed retirement. Such pension reforms are often debated in practice and they are closely related to our theoretical results. After describing the setting and mapping it into our theory, we present welfare results based on model simulations, and a complementary set of results based on sufficient statistics.

## 4.1 Institutional Setting and Data

Germany has a pay-as-you-go pension system sharing many of its key characteristics with public pension systems in other developed countries. The vast majority of German workers are covered by public pensions, as enrollment is mandatory for most employees. Pension contributions are levied as a payroll tax on gross earnings. Benefits are defined according to a pension formula based on a worker's lifetime contribution history. Pension benefits are roughly proportional to lifetime income and there is relatively little redistribution. The average net replacement rate is 53% (OECD, 2021), and public pensions are the main source of income for most recipients.

The first key policy dimension for our purposes is given by statutory retirement ages. These are saliently presented age thresholds used as reference points in the framing of retirement and benefit rules. Most importantly, the *Normal Retirement Age (NRA)* is presented to workers as a "normal" age or time to retire in information material, pension statement letters, and other official government communication. This framing translates into a general perception of the NRA as the reference age of retirement: for instance, a pension reform that will increase the NRA to 67 is commonly known as "retirement at 67" in Germany.

The NRA is the most salient and latest statutory retirement age, but there are others in the system. In addition to the NRA, there is a Full Retirement Age (FRA) from which a "full" pension is available. For most workers in our analysis sample, the Normal and Full Retirement Ages coincide, but they can differ for some. Thirdly, the pension system has an Early Retirement Age (ERA), the earliest age from which a pension can be claimed, which we do not analyze directly. Overall, statutory retirement ages induce strong retirement responses. Seibold (2021) documents that 29% of workers retire exactly in the month when they reach a statutory retirement age. As we highlighted before, Figure 1 shows such sharp bunching for the case of the NRA in our sample.

The second key policy dimension is financial retirement incentives. As in many other pension systems, pension benefits are actuarially adjusted as a function of an individual's retirement age. Hence, when a worker chooses to retire later, there is an explicit upward adjustment of pension benefits on top of the increase in their baseline pension due to additional contributions. In Germany, actuarial adjustment is relatively low, however. Pensions increase by 3.6% per year of later retirement below the FRA, and there is

(16) and (17) comes at the cost of an independence assumption about preference parameters, which could be relaxed using tools from this literature.

no explicit adjustment between the FRA and the NRA, should they differ for a worker. The largest actuarial adjustment occurs above the NRA, where a *Delayed Retirement Credit (DRC)* of 6% per year applies.

Two important features of these pension adjustment rules are worth noting here. First, benefit adjustment is generally less than actuarially fair. For instance, Börsch-Supan and Wilke (2004) calculate that pension adjustment around age 65 would have to be between 7% and 8% per year in order to be actuarially fair. Thus, there is a fiscal externality when workers change their retirement decision, whereby later retirement entails a fiscal benefit to the pension system not internalized by workers. Second, the German pension adjustment schedule creates a *non-convex kink* - an increase in the marginal return to work - at the NRA. Notably, a similar kink is present in U.S. Social Security, which also features higher marginal pension adjustment above the NRA due to the DRC.

In the empirical analysis, we use the data set of Seibold (2021), which is based on administrative data covering the universe of German retirees who claim a public pension between 1992 and 2014 provided by the German State Pension Fund (Forschungsdatenzentrum der Rentenversicherung (FDZ-RV), 2015). We apply the same sample restrictions as Seibold (2021) and additionally restrict the sample to birth cohort 1946. The main reason to focus on one birth cohort is to simplify the analysis, as different cohorts face different statutory retirement ages and benefit schedules due to various cohort-based pension reforms. We use the full data for bunching estimation, while simulations are conducted on a 1% sample to facilitate computation.

## 4.2 Model and Parameter Estimation

From our theoretical results, we know that the main determinants of welfare are the strength of loss aversion ($\Lambda$) and the direction of loss aversion ($\beta$), together with the behavioral response to price changes, and the relative number of individuals in the three groups $(G, R, L)$. In the empirical application, we capture these components parsimoniously in a static model of retirement behavior with reference dependence. As we discuss above, a key challenge lies in identifying the parameter $\beta$. We follow two approaches to deal with this. First, we adopt a Simple Loss Aversion model that assumes $\beta = 0$, similar to Seibold (2021). This initial assumption is motivated by the empirical bunching patterns around the NRA from Figure 1. In particular, the clearly visible drop in the density above the NRA suggests that much (if not all) of the bunching comes from above, which is consistent with a model of loss aversion over leisure (see Section 4.6.1 for a detailed discussion). Such a Simple Loss Aversion model is closely in line with most applied work on reference dependence, and helps us build intuition about the mapping between theory and empirics. Of course, the potential downside of the simple model is the ex-ante structure it implies for some welfare results. As a second approach, we thus allow for a more flexible structure of reference dependence later on, where we empirically estimate $\beta$ and examine the robustness of key welfare effects.

We begin by adopting a Simple Loss Aversion model. Preferences of a reference-dependent agent are given by[23]

$$U_i(C, R) = C - \frac{n_i}{1 + \frac{1}{\varepsilon}} \left(\frac{R}{n_i}\right)^{1+\frac{1}{\varepsilon}} - \begin{cases} 0 & R < \hat{R} \\ \widetilde{\Lambda}(R - \hat{R}) & R \geq \hat{R} \end{cases} \tag{18}$$

where $C$ is lifetime consumption and $R$ is the worker's retirement age relative to a career starting age

[23]The simple model we consider here can be interpreted as a reduced form of a more general model of dynamic labor supply. The static version is sufficient to explain the key empirically observed retirement patterns (see e.g. Burtless, 1986; Brown, 2013; Manoli and Weber, 2016) and provides a convenient way to model reference dependence in retirement behavior. Similarly, assuming that utility is quasi-linear in consumption and iso-elastic in labor supply is convenient for the bunching strategy described below, and, though not strictly necessary, it matches our theoretical setup from Section 2.1.

normalized to 0. The parameter $\widetilde{\Lambda}$ captures the strength of loss aversion. The heterogeneous parameter $n_i$ reflects earnings ability at old age, where low ability increases disutility from postponing retirement; for our purposes the distribution of $n_i$ will determine whether an individual is in the $G, R$, or $L$ group under any given policy. The parameter $\varepsilon$ is the elasticity of the retirement age with respect to the implicit net-of-tax rate, which is the relevant elasticity to price changes for our context.

Equation (18) yields a model of loss aversion over *lifetime leisure*. Intuitively, marginal disutility from increasing labor supply beyond the retirement reference point $\hat{R}$ is greater than marginal disutility from approaching $\hat{R}$ from the left, and the parameter $\widetilde{\Lambda}$ determines the size of this kink in the utility function. Such reference dependence in terms of the retirement age can be interpreted as loss aversion over leisure, where workers perceive postponing retirement as a loss relative to a normal time to retire. Alternatively, reference dependence could also be present over consumption, the other good in equation (18) (see e.g. Behaghel and Blau, 2012). In Section 4.6, we allow for more flexible reference dependence along these lines.

Workers face a lifetime budget constraint that expresses consumption $C$ as a function of $R$:

$$C(R) = \sum_{t=0}^{R-1} \delta^t w_t (1 - \tilde{\tau}_t) + \sum_{t=R}^{T} \delta^t B(R), \tag{19}$$

where $w$ is the gross wage per period, $\tilde{\tau}$ is the payroll tax/pension contribution rate, $T$ is the time of death, and $\delta$ is the discount factor.[24] The slope of the budget constraint, that is the marginal gain in lifetime consumption possibilities $C$ from delaying retirement by one period, defines the implicit net wage $w^{net} = dC/dR$. Expressing the consumption gain as a fraction of the gross wage, the *implicit net-of-tax rate* is $1 - \tau = w^{net}/w$.

Bunching methods can be used to transparently identify key parameters of this model.[25] In line with Proposition 3, the model predicts bunching at the NRA when it serves as a reference point. As Seibold (2021) shows, one can identify a marginal bunching individual, whose indifference curve would be tangent to the budget line at some retirement age $R^*$ without reference dependence, and who is tangent exactly at $\hat{R}$ with reference dependence. All workers initially located between $\hat{R}$ and $R^*$ bunch at the reference point, while all individuals initially to the right of $R^*$ retire earlier but stay above the reference point. Individuals to the left of the reference point leave their retirement age unchanged. Hence, the bunching mass $B$ at a retirement age reference point is given by

$$B = \int_{\hat{R}}^{R^*} h_0(R) dR \approx h_0(\hat{R})(R^* - \hat{R})$$

where $h_0(\hat{R})$ is the height of the counterfactual retirement density at $\hat{R}$. Based on the tangency conditions of the marginal bunching individual, the excess mass $b = B/h_0(\hat{R})$ at a statutory retirement age can be expressed as

$$\frac{b}{\hat{R}} = \left( \frac{1 - \tau}{1 - \tau - \Delta\tau - \Lambda} \right)^{\varepsilon} - 1, \tag{20}$$

where $\Lambda = \widetilde{\Lambda}/w$ is the reference dependence parameter normalized by the wage per period and $\Delta\tau$ is the size of the budget constraint kink that may be present at the threshold. Note that equation (20) is closely related to the general bunching approximation from equation (16).[26]

[24] For simplicity, we abstract from the fact that pension benefits can only be claimed from the Early Retirement Age (ERA) onward if the worker retires before the ERA.

[25] See Kleven (2016) for a general overview of bunching methods.

[26] In the absence of a kink at the threshold, we could directly approximate equation (20) via equation (16) (see Appendix E.2). Also note that we assume that $\Lambda$ and $\varepsilon$ are homogeneous across individuals for simplicity. We scale $\Lambda$ by the wage, which corresponds to

We use the identification strategy of Seibold (2021) in order to estimate $\Lambda$ and $\varepsilon$. In particular, we leverage the fact that bunching is observed at the NRA, but also at some standard, "pure" financial incentive discontinuities, i.e. budget constraint kinks or notches without the presence of a statutory age. Indexing these various thresholds by $j$, bunching can be written as

$$\frac{b_j}{\hat{R}_j} = \left( \frac{1-\tau_j}{1-\tau_j-\Delta\tau_j-\Lambda\cdot D_j} \right)^{\varepsilon} - 1 + \xi_j \tag{21}$$

where $D_j$ is an indicator for the NRA and $\xi_j$ is an error term.[27]

Figure 1 shows the empirical retirement age distribution around the NRA among birth cohort 1946. There is sharp, large bunching at age 65, the location of the NRA. The presence of bunching is in line with the NRA serving as a reference point for retirement. While sizable bunching at the NRA has been documented across a number of countries, it is particularly striking in the German case because there is a non-convex kink at the NRA, providing a negative incentive to retire exactly at this age. The figure also shows a counterfactual density fitted as a polynomial to the empirical distribution, excluding the bunching region. Expressing the bunching mass relative to the counterfactual, the overall excess mass at the NRA is around 31, implying that workers are roughly thirty times more likely to retire exactly in the month of the NRA than we would expect from the smooth counterfactual distribution.

Appendix Table A1 shows these bunching estimates and resulting parameter estimates. In Panel A, the average excess mass at the NRA is 31.3, although there is a negative local financial incentive to retire corresponding to a kink size of –0.28. At other, pure financial incentive discontinuities faced by the same workers, the average excess mass of 6.73 is smaller, although these entail sizable financial incentives to retire with an average kink size of 0.47. The bunching observations can be used to estimate equation (21), yielding the estimates of $\Lambda = 0.46$ and $\varepsilon = 0.06$ shown in Panel B of the table. These parameter estimates for birth cohort 1946 are similar to the estimates reported in Seibold (2021) for a broader range of cohorts.

### 4.3 Conceptualizing Pension Reforms

In the light of demographic change and resulting fiscal challenges for pension systems, two types of pension reforms are often considered in order to induce workers to postpone retirement. A first common policy is an increase in the Normal Retirement Age (or similar statutory retirement ages). For example, the NRA will be increased to age 67 in the U.S. by 2027, to 67 in Germany by 2031, and to 68 in the U.K. by 2046. This type of reform entails large effects on retirement behavior (Mastrobuoni, 2009; Staubli and Zweimüller, 2013; Cribb et al., 2016), which is largely driven by shifting individuals' reference points to a higher retirement age (Behaghel and Blau, 2012; Seibold, 2021; Lalive et al., 2022).

Two important aspects are worth noting about NRA reforms. First, while an increase in the NRA sets the reference point at a higher retirement age, such a reform corresponds to decreasing the reference point in terms of lifetime leisure in the model from Section 4.2. Thus, we should conceptually think of a reform that increases the NRA as one that *lowers the reference point* for leisure in the sense of Propositions 1 and 2. Second, while our theoretical analysis considered changes to reference points holding all else fixed, changes to the NRA typically entail some change in individuals' lifetime budget constraints, because pension benefit schedules are linked to the NRA. In the German context, the DRC is only available from the NRA onward. Thus, increasing the NRA also moves the non-convex kink in the budget constraint to the new NRA. More-

the price of leisure. See footnote 21 for a brief discussion of this scaling.

[27] The empirical specification also controls for whether the NRA coincides with the FRA.

over, if the NRA coincides with the FRA, the age from which the "full" pension is available moves upwards with the reform, such that increasing the NRA effectively implies a benefit cut across the board (see e.g. Mastrobuoni, 2009). For instance, a worker retiring at age 64 incurs a penalty of 7.2% relative to full benefits when the NRA is 66, but only a 3.6% penalty when the NRA is 65.

The second type of policy often considered for pension reforms are changes to financial incentives. In particular, a natural way to incentivize workers to retire later is to offer higher marginal pension benefit increases for later retirement. This is often done by increasing the DRC, providing higher actuarial adjustment to workers retiring beyond the NRA. For instance, the U.S. DRC has been gradually increased from 3% to 8% per year over the last decades (Duggan et al., 2023). Conceptually, a higher DRC creates a higher marginal return to work, i.e. a *higher price of lifetime leisure in the loss domain* above the NRA. Whether intentionally or not, the DRC can thus be interpreted as an implicit "corrective tax" on leisure, which incentivizes individuals to move away from the reference point of the NRA by increasing their retirement age.

## 4.4 Welfare Effects of Pension Reforms: Simulation Approach

Our first empirical approach uses the model from Section 4.2 to simulate individual retirement behavior under different policy scenarios and to calculate the welfare effects of pension reforms.

### 4.4.1 Simulation Methods

We simulate the welfare effects of pension reforms, building on Seibold (2021), who calculates effects of similar reforms on behavior and fiscal balances. The first reform is an increase in the NRA from 65 to 66. As discussed above, a change in the NRA not only shifts individuals' retirement reference points, but also entails pension benefit cuts for some workers because of the link between the NRA and pension benefit calculation. We simulate two variants of the NRA reform, without and with associated benefit cuts. While the former is useful in isolating the effect of changing reference points, the latter more accurately captures a "realistic" pension reform. The other type of reform we consider is an increase in the DRC. In order to anchor this change in financial incentives, we increase the credit from the current level of 6% to 10.4% per year, which yields the same effect on the average retirement age as the first reform.

The policy simulations proceed in the following steps. First, we require a counterfactual distribution of retirement ages – a distribution of retirement ages in the absence of reference dependence. We follow the standard approach to obtain this counterfactual distribution and fit a polynomial to the observed distribution, excluding the bunching region around the NRA. In the absence of reference dependence, individuals bunching at the NRA would be distributed across retirement ages above the NRA, and we simulate this un-bunching by distributing the bunching mass across ages 65 and above.[28] We then assign counterfactual retirement ages to individuals in the data based on ranks of actually observed retirement ages.

Second, we simulate optimal retirement ages for each individual under the baseline policy environment where the NRA is 65 and the DRC is 6% per year. Third, we simulate optimal retirement ages under each counterfactual policy scenario. For this, we simulate individual lifetime budget constraints from equation (19) as in Seibold (2021), based on observed individual earnings and contribution histories, and choose

[28]The empirical retirement age distribution offers little information about the counterfactual shape of this upper tail, as few individuals actually retire above the NRA in the data (see Figure 1). In the baseline simulations, we distribute the bunching mass following a fitted Pareto distribution above age 65, corresponding to a moderately decreasing shape above the NRA. Appendix Figure A1 shows the counterfactual density under alternative assumptions about the tail of the distribution, including a uniform and a lognormal distribution above the NRA. Reassuringly, these alternative distributional assumptions have little impact on our simulation results, as Appendix Table A2 shows.

the retirement age that maximizes utility from equation (18) subject to the budget constraint and the reference point given by the NRA.

Fourth, we compute the difference between each counterfactual scenario and the baseline scenario for the following outcomes: contributions to the pension system, benefits paid to workers, and workers' lifetime consumption. Moreover, we calculate the effects on disutility from work and reference-dependent payoffs given the preferences in equation (18). Based on these, we can calculate the effects of each reform on the fiscal balance of the pension system, on the welfare of workers, and on total welfare – the sum of fiscal effects and individual welfare effects. All effects are scaled in terms of net present value at age 65, and in line with Utilitarian social welfare we focus on average effects.

#### 4.4.2 Main Simulation Results

Table 1 summarizes the effects of the two simulated pension reforms.

**Increasing the Normal Retirement Age.** In Column (1), we consider the stylized variant of the NRA reform without associated benefit cuts. Shifting the NRA by one year increases average actual retirement ages by 4.5 months. Such a reform improves the fiscal balance of the pension system, even when mechanical benefit cuts are absent. The positive fiscal effect arises due to a combination of workers paying pension contributions for a longer period and a lower net present value of benefit payments, both of which arise when individuals work longer and postpone retirement. The magnitude of the net fiscal effect is around +€6.4k per worker. Next, the reform affects workers' private welfare. Lifetime consumption increases by around +€4.2k along with later retirement. Workers incur additional disutility from work because increasing the NRA to 66 induces them to work up to one year longer. However, the increase in consumption outweighs extra disutility from work. This reflects the *behavioral welfare effect* of a change in the reference point from equations (4) and (5). The individual is consuming too much leisure when $\pi = 0$, so decreasing the reference point over leisure by increasing the NRA has a corrective effect on behavior. Thus, we find that worker welfare improves in the case of $\pi = 0$. The effect on total welfare is given by the sum of the individual welfare effect and the net fiscal effect. Under $\pi = 0$, we find that total welfare increases by +€7.6k per worker.

In addition, if the planner places normative weight on reference dependence ($\pi = 1$), we should also account for changes in reference-dependent payoffs due to the lower reference point in terms of lifetime leisure. We can conceive of the overall change in reference dependence loss disutility as the sum of two components: a negative component of -€6.8k due to additional disutility from work, and a positive component of +€7.9k from the decrease in the reference point itself.[29] When $\pi = 1$, the first of these modifies the behavioral welfare effect relative to the case when $\pi = 0$. The total behavioral welfare effect when $\pi = 1$ is the sum of worker consumption, disutility from work, and reference-dependent disutility from work, totalling -€5.6k. We observe that this behavioral welfare effect and the net fiscal effect approximately offset one another. This cancellation is a consequence of the envelope theorem, reflecting the theoretical idea that the change in behavior induced by a change in the reference point has no first-order consequences for welfare when $\pi = 1$.[30] With the behavioral effect largely eliminated under $\pi = 1$, the *direct welfare effect*, i.e. the effect on reference-dependent payoffs from the change in reference point itself, becomes the primary determinant of the total welfare effect. We find a total welfare gain under $\pi = 1$ of +€8.7k, which is larger than

[29] See Appendix E.1 for details of this decomposition of reference dependence payoffs.

[30] The two effects only approximately offset each other in the simulation for two reasons. First, there is a small fiscal externality because the pension system is less than actuarially fair. Second, the NRA increase by one year is a discrete reform, such that second-order effects can matter.

under $\pi = 0$, as we would expect from Proposition 2.1.[31] Overall, as we expect from a Simple Loss Aversion model, increasing the NRA improves welfare regardless of $\pi$, but whether this is driven by behavioral or direct welfare effects depends on $\pi$.

Column (2) of Table 1 shows corresponding results for the more realistic variant of this reform, where the NRA increase entails a benefit cut for workers retiring below the NRA. Note that the benefit cut causes a parallel downward shift of budget constraints below the NRA. Under our assumption of quasi-linear utility, this modification does not cause any change in retirement behavior. However, the realistic scenario leads to important differences in the distribution of welfare changes between workers and the government. Due to the benefit cut, the net fiscal effect increases to €10.0k, and the increase in worker consumption becomes only €0.6k. Because this small consumption gain is insufficient to compensate workers for the additional disutility from working longer, worker welfare now decreases under both $\pi = 0$ and $\pi = 1$. However, total welfare effects remains large and positive. This realistic scenario reveals a paradox of NRA reforms in practice: despite large social welfare gains, the average worker's private welfare decreases as long as increasing the NRA is linked to pension benefit cuts.

**TABLE 1: WELFARE EFFECTS OF PENSION REFORMS**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Policy 1: Normal Retirement Age to 66 | | Policy 2: |
| | Stylized scenario: without benefit cut | Realistic scenario: with benefit cut | Delayed Retirement Credit to 10.44% |
| Contributions collected | +2,359 | +2,359 | +2,297 |
| Benefits paid | +3,999 | +7,658 | –4,038 |
| Net fiscal effect | +6,358 | +10,017 | –1,741 |
| Worker consumption | +4,230 | +571 | +12,147 |
| Disutility from work | –2,950 | –2,950 | –2,187 |
| Worker welfare ($\pi = 0$) | +1,280 | –2,379 | +9,960 |
| Ref. dep. disutility from work | –6,835 | –6,835 | –8,743 |
| Ref. dep. utility from ref. point | +7,946 | +7,946 | 0 |
| Worker welfare ($\pi = 1$) | +2,391 | –1,268 | +1,217 |
| Total welfare ($\pi = 0$) | +7,638 | +7,638 | +8,219 |
| Total welfare ($\pi = 1$) | +8,749 | +8,749 | –523 |

*Notes:* The table shows results from simulations of two pension reforms, an increase in the Normal Retirement Age from 65 to 66 and an increase in the Delayed Retirement Credit to 10.44%. Both reforms yield the same effect on the average actual retirement age (+4.5 months). Simulations are conducted for birth cohort 1946. All effects are calculated among workers retiring at age 63 and above, and are in Euros per worker, in terms of net present value at age 65. The signs the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare.

**Increasing the Delayed Retirement Credit.** Column (3) of Table 1 shows the effects of the increase in the DRC to 10.4%. By construction, this policy achieves a sizable increase in the average retirement age like

[31] Note that the quantitative similarity between total welfare under $\pi = 0$ and $\pi = 1$ in our empirical setting is not a generic feature of the simple loss aversion model. Rather, it occurs because the number of individuals retiring exactly at the NRA (the $R$ group) is large. With a larger $L$ group, the additional positive welfare effect occurring only under $\pi = 1$ could be significantly bigger.

the NRA increase. Unlike the NRA reform, the fiscal effect of the DRC increase is negative, at –€1.7k per worker. Workers contribute longer in this scenario, but the positive effect on contributions is more than offset by the large increase in benefit payments.[32] Due to the higher pension benefits and the additional earnings, worker consumption increases strongly. Disutility from work becomes larger too, but less so than under the NRA reform because workers account for their individual marginal disutility of work in deciding how much later to retire. Thus, there is a large positive effect of around +€10.0k on worker welfare under $\pi = 0$.

However, the sizable behavioral response leads to an increase in reference-dependent disutility from work, reducing individual welfare by –€8.7k when this concern carries normative weight under $\pi = 1$. This large negative effect arises because workers increase their retirement ages relative to an unchanged reference point, pushing them further into the loss domain over leisure. Taking this additional welfare effect into account, individual welfare increases only by +€1.2k under $\pi = 1$. Finally, the total welfare effect is positive at +€8.2k under $\pi = 0$, as the large gain in individual welfare strongly dominates the negative fiscal effects. However, the total welfare effect turns slightly negative under $\pi = 1$, when workers experience large disutility from moving away from the reference point.

The large difference in welfare effects between the $\pi = 0$ and the $\pi = 1$ cases is directly related to our theoretical results. When $\pi = 0$, there is an internality from workers consuming too much leisure out of loss aversion. Increasing the DRC acts as a corrective tax on leisure, so this reform has a large positive behavioral welfare effect by (partially) correcting the internality (see equation (8)). In contrast, when $\pi = 1$, the change in worker welfare is much smaller because the internality is not present. Moreover, in this case, the basic intuition of the envelope theorem implies that the effect on worker welfare are virtually entirely offset by the net fiscal effect. Thus, the DRC acts as a distortionary tax on leisure under $\pi = 1$. The initial 6% credit is relatively close to actuarial fairness, so the distortion (and the resulting negative total welfare effect) are relatively small. However, distortions can become large when considering larger changes to the credit, as we find in the extended simulations below.

#### 4.4.3 Extended Simulations

We next extend the simulations to a wider range of policy reforms. This provides further insights into the relationship between the policy simulations and our theoretical results, albeit some additional caution is warranted in interpreting the findings because we are extrapolating further from observed data.

While Table 1 considers a specific change to the Normal Retirement Age, Panel (a) of Appendix Figure A2 shows results for a range of simulated counterfactual NRAs between 65 and 67 in monthly increments. The fiscal balance of the pension system increases monotonically with the NRA. Moreover, the figure confirms the positive welfare effects of increasing the NRA under the loss aversion over leisure. Total welfare increases monotonically with the NRA both under $\pi = 0$ and $\pi = 1$, where the welfare increase is stronger under $\pi = 1$.

Similarly, Panel (b) shows results for a range of simulated values of the DRC. We simulate credits between 6% and 36% per year in half-percentage point increments. The fiscal effects of increasing the DRC tend to be large and negative, because the large increases in pension benefit payments dominate increases in contributions received by the pension system. There is, however, a small range just above the current value of 6% over which the net fiscal effect of increasing the credit is positive, as the pension system moves closer

[32] That increasing the DRC is less fiscally desirable reflects an idea from Loewenstein and O'Donoghue (2006). Policies like increasing the NRA, which they might call a "psychic subsidy" for working, are less fiscally costly than an actual subsidy for working.

to actuarial fairness. Relative to the status quo, any increase in the credit throughout the range we consider we consider improves total welfare under $\pi = 0$, reflecting workers' initial over-consumption of leisure. Under $\pi = 1$, however, the corrective benefits of a higher credit are wiped out by reference-dependent disutility from later retirement, so that total welfare decreases for all but small increases in the credit.

A key difference between increasing the NRA and changing the DRC is that the total welfare effects of the latter reforms are not monotonic. In fact, the extended simulations imply an optimal level of the DRC. The welfare-maximizing credit depends strongly on whether the planner places normative weight on reference dependence. Under $\pi = 0$, total welfare is maximized at a very large DRC of 20.4% p.a., more than three times its current level. This results speaks to a possible role for the DRC to correct inefficiently early retirement under $\pi = 0$. Such a large marginal return to work, or implicit price of leisure, induces workers to retire later and move towards their optimal retirement age. The optimal level of the DRC is much lower under $\pi = 1$. Intuitively, there is no reason for the planner to incentivize workers to move away from the NRA and retire later when reference dependence is not judged as a bias. The only rationale to increase the DRC slightly above its current level is to correct the inefficiency that arises from the fiscal externality, due to less than actuarially fair pension adjustment. Indeed, the optimal DRC of 7.8% p.a. that we find is close to previous calculations of actuarially fair adjustment in the German context (Börsch-Supan and Wilke, 2004).

Overall, these simulations illustrate many of the key ideas from our theoretical results. Using a model of Simple Loss Aversion over leisure, the welfare effects of changing reference points carry an unambiguous sign. Increasing the NRA, which corresponds to lowering reference points in terms of lifetime leisure, yields increases in total welfare regardless of the normative judgment of reference dependence. Increasing the DRC, which corresponds to an implicit tax on leisure in the loss domain, increases total welfare only if reference dependence is judged as a bias. However, increasing the DRC beyond its actuarially fair level decreases welfare if reference dependence carries normative weight.

### 4.5 Using Sufficient Statistics to Calculate Welfare Effects of Pension Reforms

As a second empirical approach, we can use the sufficient statistics formulas from Proposition 2 to approximate the welfare effects of pension reforms. Compared to the full individual-level simulations, this approach is substantially easier to implement. Adapting equation (13) to the retirement context, we can express the first-order welfare effect of a small change in the Normal Retirement Age as

$$\begin{aligned} \Delta W \approx & \Delta \hat{R}\, \pi\, E[\Lambda w_i |\, i \in L]\, P[i \in L] \\ & + \Delta \hat{R}\, E\left[ \left( \frac{\Lambda}{2} + \tau_i \right) w_i \,\middle|\, i \in R \right] P[i \in R] \end{aligned} \tag{22}$$

where, as in Section 4.2, $\hat{R}$ is the retirement age reference point (given by the NRA), $w$ is the gross wage per period, $\tau$ is the implicit tax rate on working for an additional period, and $\Lambda$ is the normalized reference dependence parameter. Building on equation (15), the welfare effect of a small change in the Delayed Retirement Credit can be approximated as

$$\Delta W \approx \left( E\left[ \{-\tau_i - (1-\pi)\Lambda\} w_i \frac{\partial l_i^L}{\partial [w_i(1-\tau_i)]} \Delta\tau_i w_i \,\middle|\, i \in L \right] P[i \in L] \right) \tag{23}$$

where $l^L$ is demand for lifetime leisure in the loss domain and $\Delta\tau$ is the change in the implicit tax rate induced by the reform.

In applying our general sufficient statistics formulas from Section 3.2 to the retirement setting, a few aspects are worth noting. First, we have to take into account the fiscal externality of the pension system in calculating welfare effects. The fiscal externality is captured by the the implicit tax $\tau w$, which leads to a positive effect on government revenue when policies induce workers to retire later. In Appendix C, we provide more details of how the sufficient statistics formulas are modified in the presence of fiscal externalities.[33] Second, because we estimate the reference dependence parameter scaled as a percentage of the gross wage, it enters the formulas as $\widetilde{\Lambda} = \Lambda w$. Third, the relevant price elasticity in equation (23) is the responsiveness of lifetime leisure to the implicit net wage. As we have estimated the retirement age elasticity $\varepsilon$ and $l = T - R$, we can calculate this object as $\frac{\partial l^L}{\partial[w(1-\tau)]} = -\frac{\partial R^L}{\partial[w(1-\tau)]} = -\varepsilon \frac{R^L}{w(1-\tau-\Lambda)}$. Finally, we note that $\tau$, which we define as the implicit tax rate on working for an additional period, enters with the opposite sign compared to the case where the tax is levied on a consumption good.

It is straightforward to implement the sufficient statistics formulas (22) and (23) empirically. We require values of the reference dependence parameter $\Lambda$ and the price elasticity $\varepsilon$, which we have estimated, as well as information on average wages and implicit tax rates, which we directly calculate from the data. Appendix Table A3 summarizes all parameter values used as inputs into the formulas.

Table 2 shows results based on the sufficient statistics approach. For comparison, welfare effects based on the simulation approach are displayed in the lower panel of the table as well. Column (1) shows total welfare effects our main NRA reform, which increases the NRA by one year to 66. The sufficient statistics approach yields a welfare effect of +€6.6k under $\pi = 0$ and +€8.7k under $\pi = 1$, which is very similar to the effects of +€7.6k and +€8.7k, respectively, from the simulation approach. In Column (2), the sufficient statistics formulas yield a welfare effect of +€2.9k under $\pi = 0$ and +€1.1k under $\pi = 1$ for the main DRC reform. Compared to the simulations, we observe that the approximation under-estimates the welfare effect under $\pi = 0$ and even exhibits the wrong sign under $\pi = 1$. Why do these discrepancies occur? In the case of $\pi = 0$, the under-estimation can be mainly explained by the large fraction of workers initially bunching at the NRA in our empirical setting. Indeed, a large share of the total welfare effect of increasing the DRC under $\pi = 0$ is driven by by individuals de-bunching away from the NRA towards older retirement ages. However, the welfare effect of de-bunching is neglected by the sufficient statistics approach because it is a second-order effect. In the case of $\pi = 1$, on the other hand, non-linearity in the welfare effects of increasing the DRC plays a crucial role. Starting from slightly less than actuarially fair pension adjustment, increasing the DRC first increases welfare but quickly reaches a maximum and then begins to fall (see Section 4.4.3). The local approximation of the sufficient statistics approach captures the initial increase in welfare, but cannot account for a large DRC reform lowering welfare.

To shed more light on the nonlinearity issue, Column (3) of Table 2 considers an alternative financial incentive reform featuring a smaller increase in the DRC by only half a percentage point to 6.5%. Reassuringly, sufficient statistics and simulation approaches produce similar results for the small reform. The gap between the welfare effects shrinks to +€0.32k vs. +€0.56k in the case of $\pi = 0$. Under $\pi = 1$, the sufficient statistics approach now correctly yields a small positive effect of +€0.12k, compared to a simulated effect of around +€0.05k.[34]

Compared to the simulation approach, using the sufficient statistics formulas has advantages and disad-

[33] In particular, the sufficient statistics formula for the welfare effects of a price change somewhat simplifies when the price change is induced by a tax change because, as in optimal tax models, the direct revenue effect offsets the direct effect of the tax change on individual welfare.

[34] Remaining discrepancies between sufficient statistics and simulated effects occur due to a non-negligible share of workers de-bunching away from the NRA even for small reforms, and simulated effects exhibiting some non-linearity even locally around the status quo.

**TABLE 2: WELFARE EFFECTS OF PENSION REFORMS: SUFFICIENT STATISTICS VS. SIMULATION APPROACH**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Policy 1: Normal Retirement Age | Policy 2: Delayed Retirement Credit | |
| | Main reform: to 66 | Main reform: to 10.44% | Small reform: to 6.48% |
| Sufficient Statistics Approach | | | |
| Total welfare ($\pi = 0$) | +6,623 | +2,935 | +317 |
| Total welfare ($\pi = 1$) | +8,668 | +1,093 | +118 |
| Simulation Approach | | | |
| Total welfare ($\pi = 0$) | +7,638 | +8,219 | +556 |
| Total welfare ($\pi = 1$) | +8,749 | –523 | +45 |

*Notes:* The table compares the total welfare effects of pension reforms under the sufficient statistics approach and the simulation approach. The first two rows show results from sufficient statistics calculations based on equations (22) and (23), and the last two rows show simulated welfare effects described in Section 4.4. Columns (1) and (2) consider the main reforms from Table 1, namely increasing the Normal Retirement Age from 65 to 66 and increasing the Delayed Retirement Credit (DRC) to 10.44%. Column (3) additionally shows effects of a small change in the DRC to 6.48%. All effects are calculated among workers retiring at age 63 and above, and are in Euros per worker, in terms of net present value at age 65.

vantages. The sufficient statistics approach is substantially easier to implement than full-fledged microsimulations. Estimates of reference dependence parameters and a price elasticity are required, which can be obtained using reduced-form methods in many settings. However, as our results illustrate, the sufficient statistics approach does not always provide accurate approximations for larger reforms. In our empirical application, this issue arises in particular for price changes, where a large change in the DRC beyond its actuarially fair level can lead to the opposite-signed welfare effect compared to a small change. Naturally, such insights are beyond the scope of the local approximation of the sufficient statistics approach.

## 4.6 Two-Dimensional Reference Dependence in the Empirical Application

Our empirical results so far are based on a Simple Loss Aversion model with a reference point over leisure. Next, we allow for a more flexible structure of reference dependence. Specifically, we we consider a model of two-dimensional reference dependence, where loss aversion can also be present over consumption in addition to leisure. This structure captures the main type of deviation from Simple Loss Aversion discussed in the literature. Our analysis of the Flexible Reduced-Form specification from Section 2.3 demonstrates that two-dimensional reference dependence is one reason for the parameter $\beta$ to deviate from one, but once the restriction on $\beta$ is relaxed this implicitly nests a wider range of formulations.

### 4.6.1 Two-Dimensional Reference Dependence and Bunching at the NRA

We specify a model with reference dependence over both leisure and consumption in Appendix E.2. Preferences are identical to the initial specification in equation (18) except that (i) we add a component of utility to capture reference-dependent payoffs over consumption, and (ii) we denote the loss aversion parameter in the leisure dimension by $\Lambda_l$, and in the consumption dimension by $\Lambda_c$. As we show in Appendix B, this

model is one of the formulations approximated by the Flexible Reduced Form. Intuitively, the parameter $\Lambda$ from equation (10) corresponds to the combined strength of loss aversion in the two dimensions, and $\beta$ captures the relative importance of $\Lambda_c$, where a larger value of $\beta$ implies a stronger degree of deviation from loss aversion of leisure only.

With this more flexible structure of reference dependence, a key empirical challenge lies in identifying the relative importance of reference dependence in the two dimensions. Figure 5 illustrates how bunching methods can be used for this purpose. We compare the empirical retirement age distribution around the NRA (Panel a) to stylized predicted distributions in three cases: when reference dependence is present only over leisure (Panel b), only over consumption (Panel c), and over both consumption and leisure (Panel d). Under reference dependence over leisure, the model from Section 4.2 predicts a density shift toward the NRA from above, as individuals retire earlier due to reference dependence. Under consumption reference dependence , on the other hand, a density shift toward the NRA from below is predicted, because workers postpone retirement in order to increase consumption toward the reference point (see Appendix E.2 for details). Thus, a downward shift of the density should occur above the NRA under reference dependence over leisure, whereas there should be such missing density below the NRA under consumption reference dependence. If reference dependence is present in both dimensions, there may be no visible density shift around the NRA, as it occurs simultaneously on both sides.

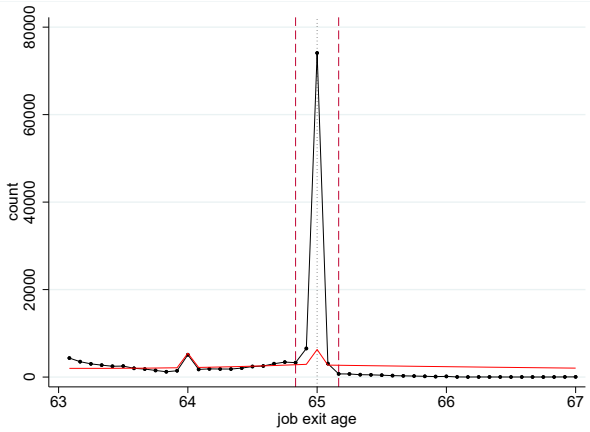
The empirical retirement age distribution exhibits a clear downward shift above the NRA, which closely resembles the prediction under a model with loss aversion over leisure only. This is our main motivation to initially consider a Simple Loss Aversion model for our empirical analysis. However, we cannot necessarily exclude *any* degree of reference dependence over consumption from visual inspection of the empirical distribution. We implement two approaches to allow for some consumption reference dependence. First, we investigate how different assumptions about the relative importance of reference dependence in the two dimensions affect our welfare results. In particular, we calculate a range of combinations of $\Lambda_l$ and $\Lambda_c$ consistent with observed bunching. We obtain these combinations by gradually moving the assumed share of bunching from the left between 0 and 50%. Panel (a) of Appendix Figure A3 shows estimated parameter combinations consistent with the observed amount of excess mass. The higher the assumed share of bunching from the left, the larger the implied $\Lambda_c$, but the smaller the implied $\Lambda_l$.[35] The labeled dots mark parameter combinations corresponding to selected left bunching shares.

The possible range of $\Lambda_c$ shown in Panel (a) of the figure is still relatively wide. Thus, our second approach is to find a preferred estimate of $\Lambda_c$ using the information contained in the observed retirement age distribution around the NRA. In terms of the Flexible Reduced-Form formulation, this corresponds to identifying the parameter $\beta$. Appendix E.2.2 provides more details of the estimation procedure we employ. Intuitively, the counterfactual density is assumed to be continuous around the threshold, and the relative number of bunchers from the left and from the right are inferred from the vertical difference between the counterfactual and the actually observed density on both sides of the threshold. In general, this approach requires a stronger assumption about the true relative density shifts being reasonably well approximated by locally observed relative shifts, which might be subject to critique (Blomquist et al., 2021). Panel (b) of Appendix Figure A3 illustrates the procedure and confirms that the implied density shift is much more substantial above than below the NRA, with a point estimate of the left bunching share of $\beta = 13.3\%$. This yields a consumption reference dependence parameter of $\Lambda_c \approx 0.67$ and a leisure reference dependence parameter of $\Lambda_l \approx 0.46$.

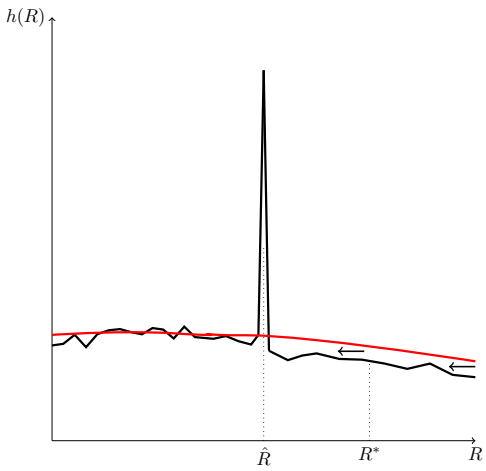
[35] Note that because the main good in the retirement model is leisure, bunching from the left in the figure is analogous to bunching from the right in the sense of Proposition 3.2.

## FIGURE 5: BUNCHING AND THE DIMENSIONS OF REFERENCE DEPENDENCE

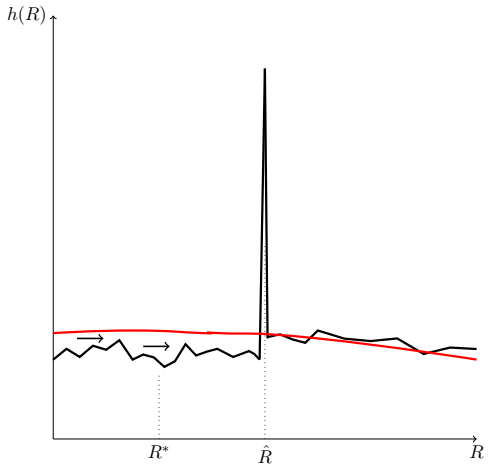
(a) Empirical Retirement Density



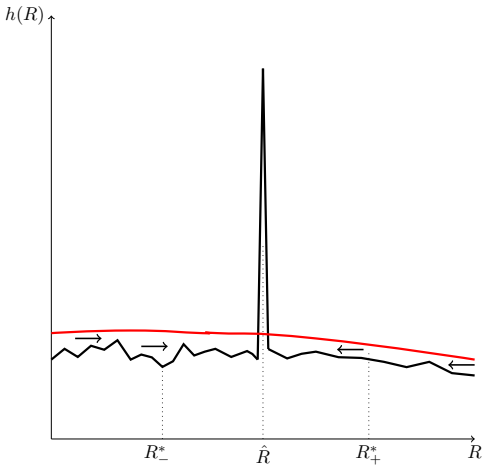

(b) Reference Dependence over Leisure




(c) Reference Dependence over Consumption




(d) Reference Dependence in Both Dimensions




*Notes:* The figure compares the empirical retirement age distribution around the Normal Retirement Age to the predicted distribution under different models of reference dependence. Panel (a) shows the empirical retirement age distribution among German workers born in 1946 as in Figure 1. Panels (b) to (d) show stylized density graphs, illustrating the predicted shape of the density under different reference dependence models, adapted to the shape of the empirical density. Panel (b) corresponds to reference dependence over leisure as described in Section 4.2, Panel (c) corresponds to reference dependence over consumption as described in Appendix E.2, and Panel (d) corresponds to reference dependence in both dimensions.

#### 4.6.2 Policy Simulations with Two-Dimensional Reference Dependence

In line with the two approaches laid out above, we present two sets of results on the welfare effects of pension reforms under two-dimensional reference dependence. First, Table 3 shows simulated welfare effects of the policies considered in Columns (2) and (3) of Table 1 under our preferred two-dimensional reference dependence parameter estimates. The NRA reform can now be interpreted as decreasing the reference point over leisure, while simultaneously increasing the reference point over consumption. The DRC reform still corresponds to a price change in the loss domain over leisure.

**TABLE 3: WELFARE EFFECTS OF PENSION REFORMS UNDER TWO-DIMENSIONAL REFERENCE DEPENDENCE**

| | (1) Policy 1: Normal Retirement Age to 66 | (2) Policy 2: Delayed Retirement Credit to 10.44% |
|---|---|---|
| Contributions collected | +2,885 | +2,327 |
| Benefits paid | +7,800 | –4,105 |
| Net fiscal effect | +10,685 | –1,778 |
| Worker consumption | +2,337 | +12,308 |
| Disutility from work | –5,392 | –2,258 |
| Worker welfare ($\pi = 0$) | –3,055 | +10,050 |
| Ref dep disutility from work | –9,015 | –8,780 |
| Utility from retirement ref point | +10,198 | 0 |
| Ref dep utility from consumption | –1,295 | 0 |
| Disutility from consumption ref point | –4,806 | 0 |
| Worker welfare ($\pi = 1$) | –7,972 | +1,270 |
| Total welfare ($\pi = 0$) | +7,630 | +8,272 |
| Total welfare ($\pi = 1$) | +2,713 | –509 |

*Notes:* The table shows results from simulations of pension reforms under two-dimensional reference dependence. The two pension reforms we consider are an increase in the Normal Retirement Age from 65 to 66 as in Column (2) of Table 1 and an increase in the Delayed Retirement Credit to 10.44% as in Column (3) of Table 1. Simulations are conducted for birth cohort 1946. All effects are calculated among workers retiring at age 63 and above, and are in Euros per worker, in terms of net present value at age 65. The signs the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare.

Fiscal effects of the two reforms remain similar to the baseline simulations: The NRA increase has strong positive fiscal effects, whereas the DRC increase worsens the fiscal balance. More generally, the effects of the DRC increase are similar to the baseline simulations: total welfare increases strongly under $\pi = 0$ but decreases under $\pi = 1$. This occurs because the effects of the DRC on retirement behavior are concentrated among workers at or above the NRA, where consumption reference dependence does not affect utility or behavior.

The magnitude of welfare effects of the NRA reform differ more substantially from the baseline simulation. As before, behavioral and fiscal effects are the first-order determinants of welfare when reference dependence is a bias ($\pi = 0$) and direct effects are the main determinant when reference dependence is judged to be normative ($\pi = 1$). In the $\pi = 0$ case, some of the behavioral effect now comes from workers who are retiring *too late* out of loss aversion over consumption, and increasing the NRA exacerbates this

internality (while still mitigating the internality for those above the NRA who retire too early).[36] Nevertheless, the total welfare effect under $\pi = 0$ remains quite similar. When $\pi = 1$, the main difference between Table 3 and our baseline simulations comes from workers retiring before the NRA, i.e. in the gain domain for leisure. There is a negative direct welfare effect for this group in the two-dimensional model, which counteracts the positive effect on those retiring after the NRA. Intuitively, workers retiring after the NRA face a lower reference point for lifetime leisure, increasing their utility as in Table 1, but workers retiring before the NRA face a higher reference point for lifetime consumption, which reduces their welfare. Because many individuals retire before the NRA, this direct effect substantially reduces the total welfare effect under $\pi = 1$. Yet, the welfare effects of increasing the NRA remain positive both under $\pi = 0$ and $\pi = 1$ in Table 3, which reflects that loss aversion over leisure is the dominant form of loss aversion according to our preferred estimates.

Second, we calculate welfare effects of the NRA reform for a wider range of left bunching shares, corresponding to a wider range of values of $\Lambda_c$, in order to account for the uncertainty that remains in estimating these parameters in Figure A3. Panel (a) of Figure 6 shows welfare effects under $\pi = 0$ and $\pi = 1$ for left bunching shares between 0 and 50%, which we consider an upper bound given the a priori evidence from Figure 5 that most bunching originates from the right. The dashed vertical lines mark the cases where the left bunching share is zero (corresponding to the baseline simulation under Simple Loss Aversion), 13% (our preferred estimate) and 50% (the upper bound). The welfare effect of the NRA reform generally decrease with the left bunching share. Under $\pi = 0$, the effect remains positive up until a large left bunching share of 40%.[37] Under $\pi = 1$, in contrast, the effect already turns negative at a share of 18%. The faster decline of the welfare effect with $\pi = 1$ is due to growing negative direct effects on pre-NRA retirees.[38]

Panel (b) of Figure 6 goes one step further and shows which value of the NRA maximizes social welfare in the simulations. As we know from the previous results, under loss aversion over leisure only, i.e. at a left bunching share of zero, increasing the NRA always increases welfare. However, welfare does not monotonically increase with the NRA when we allow for some consumption reference dependence. At our preferred left bunching share estimate of 13%, increasing the NRA from 65 remains welfare-improving regardless of $\pi$: the optimal NRA is above 68 when $\pi = 0$ and slightly below 66 when $\pi = 1$. However, with sufficiently large consumption reference dependence, this result can change. When the left bunching share exceeds 44% (for $\pi = 0$) or 29% (for $\pi = 1$), it becomes optimal to *decrease* the NRA below 65. Yet, the optimal NRA is only marginally below 65 for any left bunching share.

These results illustrate how the direction of loss aversion matters for the welfare effects of pension reforms. In our empirical setting, increasing the NRA remains welfare-improving under our preferred estimates of two-dimensional reference dependence parameters regardless of the value of $\pi$. However, by how much the NRA should increase beyond 65 depends on both normative judgments and the strength of reference dependence over consumption. Furthermore, if consumption reference dependence was much

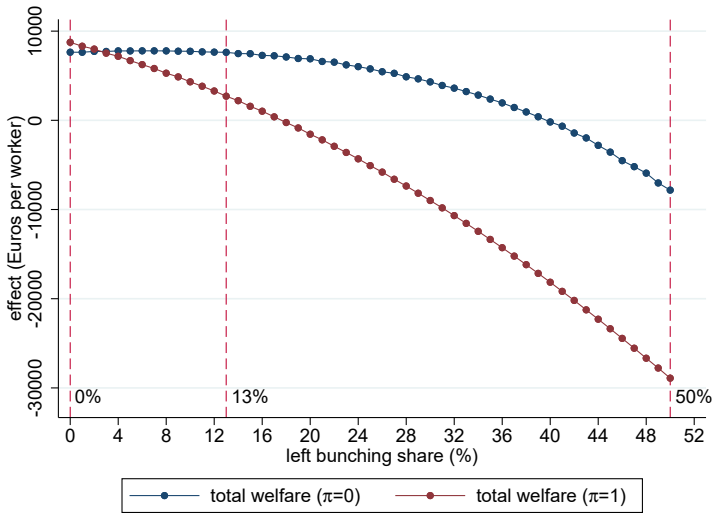
[36] As we focus on the realistic NRA reform scenario, the benefit cut below the NRA exerts a mechanical negative effect on the consumption of workers retiring before the NRA. For this reason, the behavioral component of reference-dependent payoffs over consumption turns negative in the simulation. Under the stylized reform scenario without the benefit cut, this component would be positive.

[37] Equation (13) suggests that under $\pi = 0$, exact offsetting of welfare effects for those moving closer to versus further away from their intrinsic optimum occurs where the left bunching share is 50%, but this equation is based on an approximation assuming a uniform distribution of intrinsic optima for individuals in the reference domain. In our simulated model, this distribution has a modestly negative gradient over the retirement age, so we reach exact offsetting slightly before 50%.

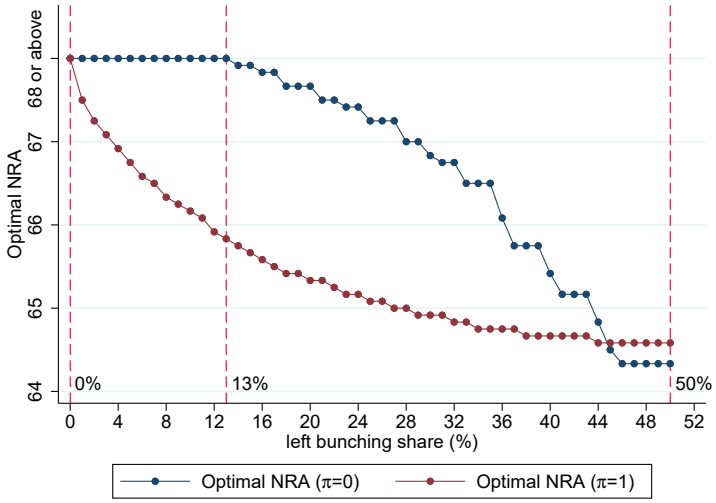
[38] We suppose that all workers retiring at 63 or later use the NRA as a reference point for consumption and leisure in the two-dimensional simulations. If some pre-NRA retirees do not use the NRA as a reference point, but some other reference point like the Early Retirement Age, behavior and welfare would be less affected by a change in the NRA than we find in Table 3, but more closely resemble those from Table 1. Conversely, the difference in welfare effects would be exacerbated if many workers retiring far below the NRA use it as a reference point.

## FIGURE 6: WELFARE RESULTS BY LEFT BUNCHING SHARE

(a) Welfare Effects of Increasing the Normal Retirement Age from 65 to 66




(b) Welfare-Maximizing Normal Retirement Age




*Notes:* The figure shows welfare results as a function of the left bunching share. A higher left bunching share corresponds to stronger consumption reference dependence, i.e. a stronger deviation from Simple Loss Aversion over leisure. Panel (a) shows the total welfare effects of increasing the Normal Retirement Age from 65 to 66, and Panel (b) shows the welfare-maximizing Normal Retirement Age. The results shown in both panels are based on simulations are conducted for birth cohort 1946. The effects are calculated among workers retiring at age 63 and above, and are in Euros per worker, in terms of net present value at age 65. In each panel, the dashed vertical lines denote selected values of the left bunching share, namely zero (no consumption reference dependence), 13% (our preferred estimate, on which the results in Table 3 are based), and 50% (the upper bound).

stronger than our estimates suggest, it could even be optimal to decrease the NRA. The welfare effects of the DRC reform are less impacted by allowing for two-dimensional reference dependence, on the other hand.

### 4.7 Discussion

Our baseline empirical analysis finds that pension reforms increasing the NRA always increase welfare, while financial retirement incentives have ambiguous welfare effects. Taken at face value, this could be interpreted as an extreme policy implication, where the NRA should be increased as much as possible. Our theoretical analysis shows that such extreme policy recommendation can be an inherent feature of Simple Loss Aversion models, which are often used in practice.

Several factors may temper this policy implication. First, allowing for a more flexible structure of reference-dependent preferences can modify the welfare effects of changing reference points. In the retirement context, reference dependence over consumption can lead to opposite-signed welfare effects to reference dependence over leisure. We find that in our empirical setting, because loss aversion over leisure appears to be the dominant form of reference dependence, and it remains welfare-improving to increase the NRA under our preferred two-dimensional estimates. However, as the relative strength of reference dependence over leisure vs. consumption could depend on how retirement and benefit rules are communicated to workers across different pension systems, some caution may be warranted when extrapolating our results to other settings.

Second, an important issue is how policies influencing points can be designed in practice. Our theoretical analysis examines the effects of a change in reference point *ceteris paribus*, but in the retirement context, pension reforms changing the NRA are typically linked to benefit cuts. As we discuss in Section 4.4, this creates large mechanical transfers to the government and corresponding decreases in private welfare. This could be an important factor behind the lack of political and popular support for increasing the NRA seen in many countries. At the same time, our results raise questions about whether it might be possible to design NRA increases without accompanying changes to benefit schedules.

Third, issues of credibility may arise when governments try to influence reference points by shifting the NRA to very high levels. Empirical evidence clearly demonstrates that reforms increasing the NRA or similar statutory retirement ages have large effects on retirement behavior over the observed range (Seibold, 2021; Lalive et al., 2022; Gruber et al., 2022). Yet, it is less clear to what extent governments can stretch these effects. For instance, workers may not find a NRA of 80 a credible "normal" time to retire. These effectiveness considerations may limit the scope of NRA increases, at least in the short run.

## 5 Conclusion

In this paper, we provide a first attempt at studying the welfare economics of reference dependence. We characterize welfare in terms of direct and behavioral effects. For a change in the reference point, we show that the sign of the total welfare effect is unambiguous within a formulation of reference-dependent payoffs. For the welfare effect of a price change, both normative judgments and the payoff formulation matter. We develop a flexible reduced-form formulation of reference-dependent payoffs, which we can use to obtain widely applicable sufficient statistics formulas describing the welfare effects of policies.

Our empirical application highlights the real-world policy relevance of these results. Reference-dependent behavior has been documented in a wide variety of empirical settings, raising important questions about

optimal policy design under such preferences. In the context of retirement, we find that increasing the Normal Retirement is welfare-improving when it serves as a reference point over lifetime leisure. When allowing for a more general structure or reference dependence, this result remains valid under our preferred estimates, but a lower NRA could be optimal under very strong consumption reference dependence. Meanwhile, the welfare effects of subsidies for later retirement are ambiguous and depend critically on normative judgments about reference dependence.

Our analysis suggests that, besides the normative judgment about reference dependence, the form of reference-dependent payoffs shapes key welfare effect. How exactly reference dependence modifies individual choices above and below the reference point for a given good is crucial. We argue that bunching methods are useful in addressing this question. Obtaining more evidence on the nature of reference-dependent payoffs across the many contexts in which reference dependence has been shown to matter will be a fruitful avenue for future research. In addition to bunching methods, one could potentially employ alternative empirical strategies such as survey experiments, which have been used to address similar problems in other behavioral contexts (Chetty et al., 2009; Allcott and Taubinsky, 2015; Allcott et al., 2019; Allcott and Kessler, 2019; Goldin and Reck, 2020). Beyond deterministic settings, such approaches might prove especially useful for analyzing welfare in the case of reference dependence under uncertainty, which is the subject of a rich literature.

More broadly, our results demonstrate that embracing normative ambiguity can provide a way forward for some difficult problems in behavioral economics (Goldin and Reck, 2022). The question of whether behavioral phenomena arise due to biases or non-standard normative preferences has complicated incorporating behavioral economics into welfare analysis in a number of domains. Incorporating normative ambiguity can be productive because it allows us to separate questions that can be empirically analyzed, such as the influence of a change in reference point on behavior, from normative judgments.

# References


Allcott, H. and Kessler, J. B. (2019). The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons. *American Economic Journal: Applied Economics*, 11(1):236–76.

Allcott, H., Lockwood, B. B., and Taubinsky, D. (2019). Regressive Sin Taxes, with an Application to the Optimal Soda Tax. *Quarterly Journal of Economics*, 23(3):1557–1626.

Allcott, H. and Taubinsky, D. (2015). Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market. *American Economic Review*, 105(8):2501–38.

Allen, E. J., Dechow, P. M., Pope, D. G., and Wu, G. (2017). Reference-Dependent Preferences: Evidence from Marathon Runners. *Management Science*, 63(6):1657–72.

Andersen, S., Badarinza, C., Liu, L., Marx, J., and Ramadorai, T. (2022). Reference dependence in the housing market. *American Economic Review*, 112(10):3398–3440.

Barseghyan, L., Molinari, F., O'Donoghue, T., and Teitelbaum, J. C. (2013). The Nature of Risk Preferences: Evidence from Insurance Choices. *American Economic Review*, 103(6):2499–2529.

Behaghel, L. and Blau, D. M. (2012). Framing Social Security Reform: Behavioral Responses to Changes in the Full Retirement Age. *American Economic Journal: Economic Policy*, 4(4):41–67.

Bernheim, B. D. (2009). Behavioral Welfare Economics. *Journal of the European Economic Association*, 7(2-3):267–319.

Bernheim, B. D., Fradkin, A., and Popov, I. (2015). The Welfare Economics of Default Options in 401(k) Plans. *American Economic Review*, 105(9):2798–2837.

Bernheim, B. D. and Rangel, A. (2009). Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. *Quarterly Journal of Economics*, 124(1):51–104.

Bernheim, B. D. and Taubinsky, D. (2018). Behavioral Public Economics. In *Handbook of Behavioral Economics: Applications and Foundations*, volume 1, pages 381–516. Elsevier.

Blomquist, S., Newey, W. K., Kumar, A., and Liang, C.-Y. (2021). On bunching and identification of the taxable income elasticity. *Journal of Political Economy*, 129(8):2320–2343.

Börsch-Supan, A. and Wilke, C. B. (2004). The German Public Pension System: How It Was, How It Will Be. NBER working paper no. 10525.

Brown, K. M. (2013). The Link between Pensions and Retirement Timing: Lessons from California Teachers. *Journal of Public Economics*, 98(1–2):1–14.

Burtless, G. (1986). Social Security, Unanticipated Benefit Increases, and the Timing of Retirement. *Review of Economic Studies*, 53(5):781–805.

Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. (1997). Labor Supply of New York City Cabdrivers: One Day at a Time. *Quarterly Journal of Economics*, 112(2):407–41.

Chetty, R. (2009). Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods. *Annual Review of Economics*, 1(1):451–488.

Chetty, R., Looney, A., and Kroft, K. (2009). Salience and Taxation: Theory and Evidence. *American Economic Review*, 99(4):1145–1177.

Crawford, V. P. and Meng, J. (2011). New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income. *American Economic Review*, 101(5):1912–32.

Cribb, J., Emmerson, C., and Tetlow, G. (2016). Signals Matter? Large Retirement Responses to Limited Financial Incentives. *Labour Economics*, 42:203–12.

De Martino, B., Camerer, C. F., and Adolphs, R. (2010). Amygdala Damage Eliminates Monetary Loss Aversion. *Proceedings of the National Academy of Sciences*, 107(8):3788–92.

DellaVigna, S. (2018). Structural Behavioral Economics. In *Handbook of Behavioral Economics: Applications and Foundations*, volume 1, pages 613–723. Elsevier.

DellaVigna, S., Lindner, A., Reizer, B., and Schmieder, J. F. (2017). Reference-Dependent Job Search: Evidence from Hungary. *Quarterly Journal of Economics*, 132(4):1969–2018.

Duggan, M., Dushi, I., Jeong, S., and Li, G. (2023). The Effect of Changes in Social Security's Delayed Retirement Credit: Evidence from Administrative Data. *Journal of Public Economics*, 223:104899.

Fehr, E. and Goette, L. (2007). Do Workers Work More if Wages are High? Evidence from a Randomized Field Experiment. *American Economic Review*, 97(1):298–317.

Forschungsdatenzentrum der Rentenversicherung (FDZ-RV) (2015). Versichertenrentenzugang 1992 – 2014. Research Data Center of the German State Pension Fund.

Gelber, A. M., Jones, D., and Sacks, D. W. (2020). Estimating adjustment frictions using nonlinear budget sets: Method and evidence from the earnings test. *American Economic Journal: Applied Economics*, 12(1):1–31.

Goldin, J. and Reck, D. (2020). Revealed Preference Analysis with Framing Effects. *Journal of Political Economy*, 126(7):2759–95.

Goldin, J. and Reck, D. (2022). Optimal Defaults with Normative Ambiguity. *Review of Economics and Statistics*, 104(1):17–33.

Gruber, J., Kanninen, O., and Ravaska, T. (2022). Relabeling, Retirement and Regret. *Journal of Public Economics*, 211:104677.

Haller, A. (2022). Welfare Effects of Pension Reforms. Working paper.

Homonoff, T. A. (2018). Can Small Incentives Have Large Effects? The Impact of Taxes Versus Bonuses on Disposable Bag Use. *American Economic Journal: Economic Policy*, 10(4):177–210.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, 98(6):1325–1348.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, 47(2):263–92.

Kahneman, D., Wakker, P. P., and Sarin, R. (1997). Back to Bentham? Explorations of Experienced Utility. *Quarterly Journal of Economics*, 112(2):375–406.

Kermer, D. A., Driver-Linn, E., Wilson, T. D., and Gilbert, D. T. (2006). Loss Aversion is an Affective Forecasting Error. *Psychological Science*, 17(8):649–53.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8:435–64.

Kolsrud, J., Landais, C., Reck, D., and Spinnewijn, J. (2023). Retirement Consumption and Pension Design. *forthcoming, American Economic Review*.

Kőszegi, B. and Rabin, M. (2006). A Model of Reference-Dependent Preferences. *Quarterly Journal of Economics*, 121(4):1133–65.

Lalive, R., Magesan, A., and Staubli, S. (2022). How Social Security Reform Affects Retirement and Pension Claiming. *American Economic Journal: Economic Policy*, forthcoming.

List, J. A., Rodemeier, M., Roy, S., and Sun, G. K. (2023). Judging nudging: Understanding the welfare effects of nudges versus taxes. NBER working paper no. 31152.

Loewenstein, G. and O'Donoghue, T. (2006). "We Can Do This the Easy Way or the Hard Way": Negative Emotions, Self-Regulation, and the Law. *University of Chicago Law Review*, 73(1):183–206.

Manoli, D. and Weber, A. (2016). Nonparametric Evidence on the Effects of Financial Incentives on Retirement Decisions. *American Economic Journal: Economic Policy*, 8(4):160–182.

Mastrobuoni, G. (2009). Labor Supply Effects of the Recent Social Security Benefit Cuts: Empirical Estimates Using Cohort Discontinuities. *Journal of Public Economics*, 93(11-12):1224–1233.

Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601.

Moore, D. T. (2022). Evaluating tax reforms without elasticities: What bunching *Can* identify. Working paper.

Mullainathan, S., Schwartzstein, J., and Congdon, W. J. (2012). A Reduced-Form Approach to Behavioral Public Finance. *Annual Review of Economics*, 4:511–540.

O'Donoghue, T. and Sprenger, C. (2018). Reference-Dependent Preferences. In *Handbook of Behavioral Economics: Applications and Foundations*, volume 1, pages 1–77. Elsevier.

OECD (2021). Pensions at a Glance 2021. OECD database.

Rees-Jones, A. (2018). Quantifying Loss-Averse Tax Manipulation. *Review of Economic Studies*, 85(2):1251–78.

Rick, S. (2011). Losses, Gains, and Brains: Neuroeconomics Can Help to Answer Open Questions about Loss Aversion. *Journal of Consumer Psychology*, 21(4):453–63.

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., Davison, C., Demić, E., Esteban-Serna, C., Friedemann, M., et al. (2020). Replicating Patterns of Prospect Theory for Decision under Risk. *Nature Human Behavior*, 4(6):622–633.

Saez, E. and Stantcheva, S. (2016). Generalized Social Marginal Welfare Weights for Optimal Tax Theory. *American Economic Review*, 106(1):24–45.

Seibold, A. (2021). Reference Points for Retirement Behavior: Evidence from German Pension Discontinuities. *American Economic Review*, 111(4):1126–65.

Sokol-Hessner, P., Camerer, C. F., and Phelps, E. A. (2013). Emotion Regulation Reduces Loss Aversion and Decreases Amygdala Responses to Losses. *Social Cognitive and Affective Neuroscience*, 8(3):341–50.

Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F., and Phelps, E. A. (2009). Thinking Like a Trader Selectively Reduces Individuals' Loss Aversion. *Proceedings of the National Academy of Sciences*, 106(13):5035–40.

Sokol-Hessner, P. and Rutledge, R. B. (2019). The Psychological and Neural Basis of Loss Aversion. *Current Directions in Psychological Science*, 28(1):20–27.

Staubli, S. and Zweimüller, J. (2013). Does Raising the Early Retirement Age Increase Employment of Older Workers? *Journal of Public Economics*, 108:17–32.

Thakral, N. and Tô, L. T. (2021). Daily Labor Supply and Adaptive Reference Points. *American Economic Review*, 111(8):2417–43.

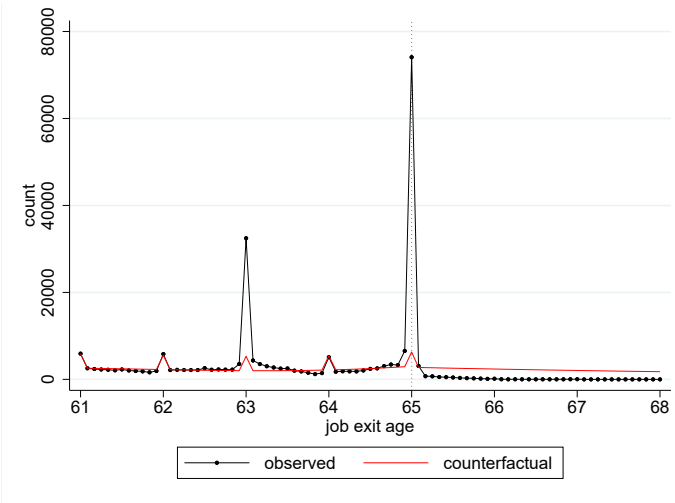
Tversky, A. and Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *Quarterly Journal of Economics*, 106(4):1039–61.

# Online Appendix (Not For Print Publication)

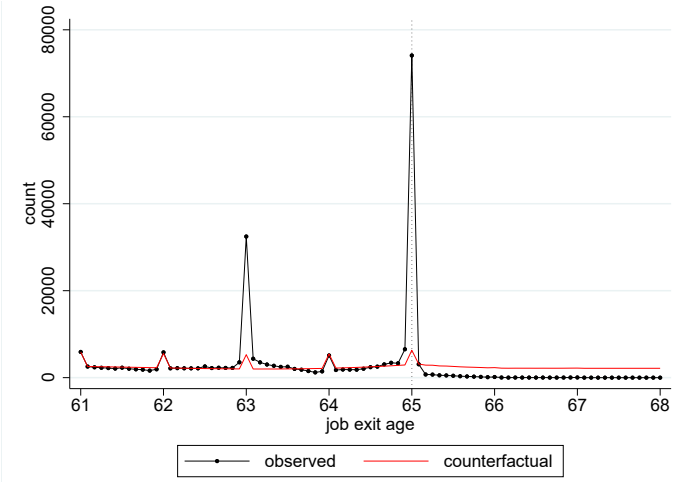
## A Additional Figures and Tables

## FIGURE A1: COUNTERFACTUAL RETIREMENT AGE DISTRIBUTION

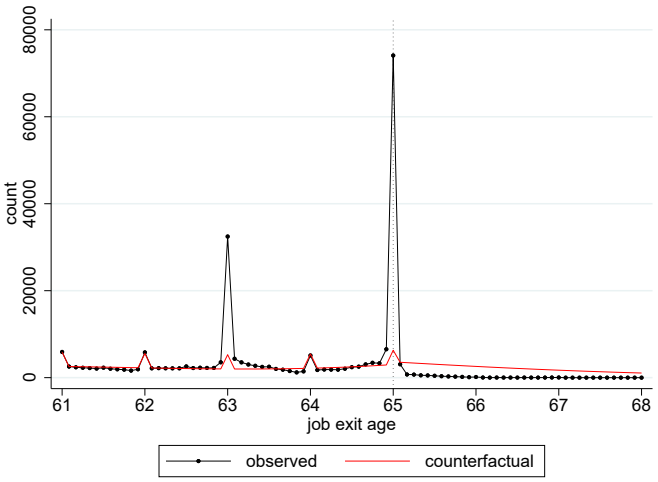
(a) Baseline: Pareto Tail




(b) Alternative 1: Uniform Tail



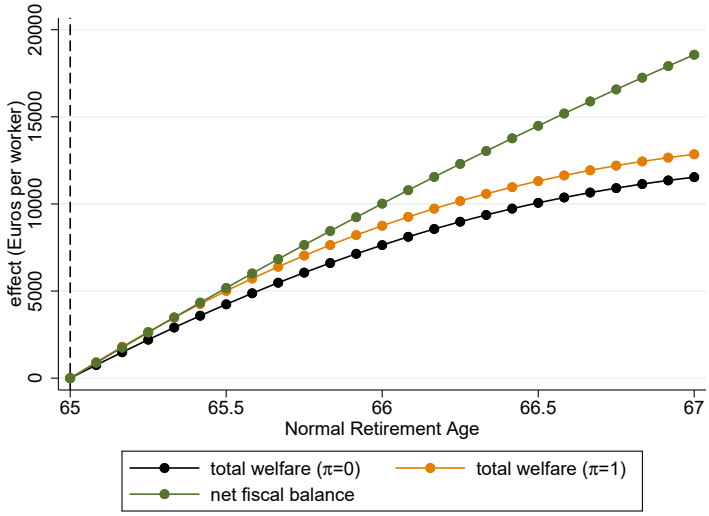

(c) Alternative 2: Lognormal Tail



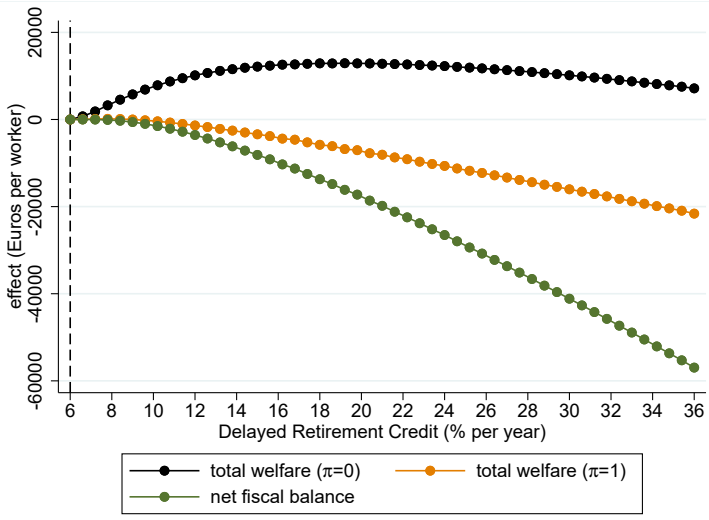

*Notes:* The figure shows counterfactual retirement distributions under different assumptions about the shape of the upper tail of the distribution. In all panels, the counterfactual distribution up until the Normal Retirement Age (age 65) is obtained by fitting a seventh-order polynomial to the observed retirement age distribution, allowing for round-age effects. Panel (a) shows the baseline distribution we use in the simulations, where the upper tail is given by a fitted Pareto distribution. Panels (b) and (c) show alternative counterfactual distributions, where the upper tail is is given by a uniform and lognormal distribution, respectively. Appendix Table A2 shows that our simulation results are robust to the shape of the upper tail of the counterfactual distribution.

## FIGURE A2: EXTENDED POLICY SIMULATIONS

(a) Increasing the Normal Retirement Age



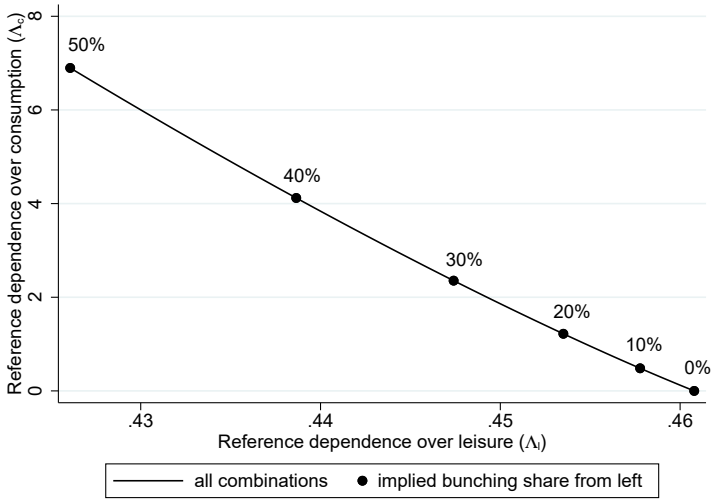

(b) Increasing the Delayed Retirement Credit



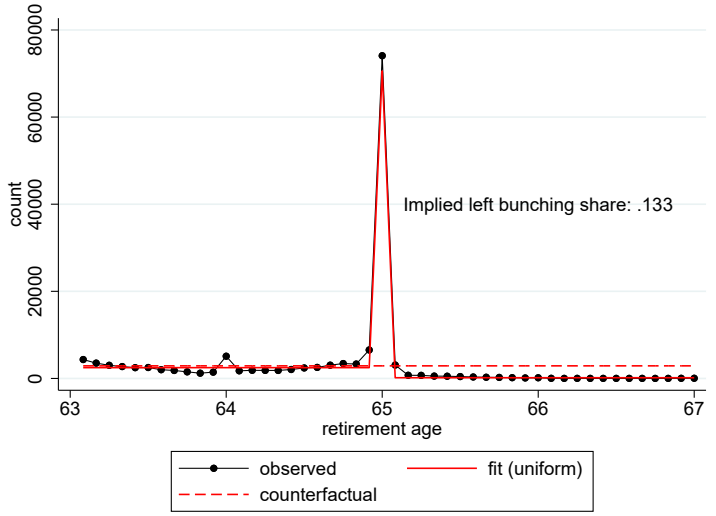

*Notes:* The figure shows simulated fiscal and welfare effects of pension reforms over an extended range of policies. Panel (a) shows the effects of increasing the Normal Retirement Age to ages between 65 and 67 in monthly increments. Panel (b) shows the effects of increasing the Delayed Retirement Credit to values between 6% and 36% per year in half-percentage point increments.Simulations are conducted for birth cohort 1946. All effects are calculated among workers retiring at age 65 and above, and are in Euros per worker, in terms of net present value at age 65. Total welfare is the sum of net fiscal effect and change in worker welfare.

## FIGURE A3: TWO-DIMENSIONAL REFERENCE DEPENDENCE

(a) Bunching at the NRA Identifies Combinations of $\Lambda_l, \Lambda_c$




(b) Preferred Bunching Share Estimate




*Notes:* Panel (a) of the figure shows a simulated range of combinations of reference dependence over leisure $\Lambda_l$ and reference dependence over consumption $\Lambda_c$. Parameter combinations are obtained by gradually moving the left bunching share from zero to 50% as described in Appendix E.2. Labeled dots mark parameter combinations implied by selected left bunching shares between 0 and 50%. Panel (b) illustrates how we obtain our preferred estimate of $\Lambda_c$. The black connected dots show the observed retirement age distribution around the NRA among workers born in 1946. The solid red line denotes the average empirical retirement age density on each side of the threshold, and the dashed red line denotes the implied counterfactual density.

## TABLE A1: BUNCHING AND PARAMETER ESTIMATES

**Panel A: Bunching Estimates**

| | (1) Excess mass | (2) Kink size | (3) Number of bunching observations |
|---|---|---|---|
| Normal Retirement Age (NRA) | 31.29 (6.42) | -0.28 | 5 |
| Pure financial incentive discontinuities | 6.73 (2.09) | 0.47 | 15 |

**Panel B: Parameter Estimates**

| | |
|---|---|
| Reference dependence w.r.t. NRA $\Lambda$ | 0.461 (0.000) |
| Retirement age elasticity $\varepsilon$ | 0.057 (0.014) |

*Notes:* Panel A of the table summarizes bunching estimates at the Normal Retirement Age and at pure financial incentive discontinuities. The excess mass figures shown represent the average excess mass estimates at the respective type of threshold among the subset of group-level bunching observations from Seibold (2021) applying to workers in birth cohort 1946, with standard errors in parantheses. The table also shows the average kink size at each type of threshold as well as the number of bunching observations the average estimate is based on. Panel B presents the parameter estimates based on estimating equation (21), using the bunching observations summarized in Panel A.

**TABLE A2: WELFARE EFFECTS OF PENSION REFORMS: ALTERNATIVE COUNTERFACTUAL DISTRIBUTIONS**

| | (1) | (2) |
|---|---|---|
| | **Panel A: Uniform Tail** | |
| | Policy 1: Normal Retirement Age to 66 | Policy 2: Delayed Retirement Credit to 10.20% |
| Contributions collected | +2,363 | +2,278 |
| Benefits paid | +7,678 | –3,879 |
| Net fiscal effect | +10,041 | –1,601 |
| Worker consumption | +562 | +11,937 |
| Disutility from work | –2,901 | –2,094 |
| Worker welfare ($\pi = 0$) | –2,339 | +9,843 |
| Ref. dep. disutility from work | –6,900 | –8,670 |
| Ref. dep. utility from ref. point | +8,121 | 0 |
| Worker welfare ($\pi = 1$) | –1,118 | +1,173 |
| Total welfare ($\pi = 0$) | +7,702 | +8,242 |
| Total welfare ($\pi = 1$) | +8,923 | –428 |
| | **Panel B: Lognormal Tail** | |
| | Policy 1: Normal Retirement Age to 66 | Policy 2: Delayed Retirement Credit to 11.28% |
| Contributions collected | +2,261 | +2,173 |
| Benefits paid | +7,531 | –4,386 |
| Net fiscal effect | +9,792 | –2,213 |
| Worker consumption | +376 | +12,103 |
| Disutility from work | –3,198 | –2,434 |
| Worker welfare ($\pi = 0$) | –2,822 | +9,669 |
| Ref. dep. disutility from work | –6,027 | –8,268 |
| Ref. dep. utility from ref. point | +6,859 | 0 |
| Worker welfare ($\pi = 1$) | –1,990 | +1,401 |
| Total welfare ($\pi = 0$) | +6,970 | +7,456 |
| Total welfare ($\pi = 1$) | +7,802 | –811 |

*Notes:* The table shows results from pension reform simulations as in Columns (2) and (3) of Table 1 under alternative assumptions about the upper tail of the retirement age distribution as indicated in the panel titles. Each panel considers two reforms, an increase in the Normal Retirement Age (NRA) from 65 to 66, and an increase in the Delayed Retirement Credit yielding the same effect on the average retirement age as the NRA reform, given the respective assumption about the retirement age distribution. Simulations are conducted for birth cohort 1946. All effects in Euros per worker, in terms of net present value at age 65. The signs the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare.

**TABLE A3: PARAMETERS FOR SUFFICIENT STATISTICS CALCULATIONS**

| Parameter | Value |
|---|---|
| Loss aversion parameter $\Lambda$ | 0.461 |
| Average monthly wage $E(w_i)$ | 2,400.639 |
| Average implicit tax rate (worker) | 0.178 |
| Employer contribution rate | 0.095 |
| Total fiscal externality $E(\tau_i)$ | 0.273 |
| Fraction in $L$ group $P(i \in L)$ | 0.154 |
| Fraction in $R$ group $P(i \in R)$ | 0.456 |
| Leisure demand responsiveness $E\left[\frac{\partial l_i^L}{\partial[w_i(1-\tau_i)]}\right]$ | -0.017 |
| Average change in implicit tax rate $E(\Delta\tau_i)$ (main DRC reform) | -0.264 |
| Average change in implicit tax rate $E(\Delta\tau_i)$ (small DRC reform) | -0.029 |

*Notes:* The table shows the parameter values entering the sufficient statistics calculations in Section 4.5.

# B Detailed Analysis of Reference-Dependent Payoff Formulations

In this Appendix, we examine how the welfare effects of changing reference points and prices are shaped by the form of reference-dependent payoffs. In particular, we apply our general characterization of these welfare effects from Proposition 1 and equations (4), (5), and (8) to an exhaustive list of payoff formulations. Tables B1 and B2 provide an overview of payoff formulations and summarize key results.

This Appendix is structured as follows: Sections B.1 and B.2 analyze the most commonly used formulations of reference-dependent payoffs, namely Simple Loss Aversion and Loss Aversion with Gain Utility. Section B.3 considers Kőszegi and Rabin (2006)-type reference dependence over utils. Section B.4 examines an alternative type of reference dependence we label Gain Discounting. Section B.5 investigates the impact of incorporating Diminishing Sensitivity on key results. Section B.6 analyzes two-dimensional reference dependence. Finally, Section B.7 demonstrates how our Flexible Reduced-Form specification can approximate a broad set of reference-dependent payoff formulations.

**TABLE B1: REFERENCE-DEPENDENT PAYOFF FORMULATIONS**

| Description | (1) Reference-Dependent Payoff | (2) Assumptions A1 & A2 | (3) Lemma 1 Case |
|---|---|---|---|
| Simple Loss Aversion | $1\{x < r\}\Lambda(x-r)$ | Yes | everywhere increasing + single-peaked |
| Loss Aversion with Gain Utility | $(\eta + 1\{x < r\}\Lambda)(x-r)$ | Yes | everywhere increasing |
| Utils Formulation (Köszegi-Rabin) | $(\eta + 1\{x < r\}\Lambda)(u(x)-u(r))$ | Yes | everywhere increasing |
| Gain Discounting | $1\{x > r\}\Gamma(x-r)$ | Yes | everywhere decreasing + single-peaked |
| Simple Loss Aversion with Diminishing Sensitivity | $-\alpha^{-1}(1\{x < r\}\Lambda)(r-x)^{\alpha}$ | 2.2 Fails | N/A |
| Loss Aversion with Gain Utility & Diminishing Sensitivity | $\alpha^{-1}(\eta)(x-r)^{\alpha}$, if $x \geq r$<br>$-\alpha^{-1}(\eta+\Lambda)(r-x)^{\alpha}$, if $x < r$ | 2.2 Fails | N/A |
| Two-Dimensional Loss Aversion, $(r_x, r_y)$ on budget constraint | $1\{x < r_x\}\Lambda_x(x-r_x)$<br>$+1\{y < r_y\}\Lambda_y(y-r_y)$ | Yes | single-peaked |
| Two-Dimensional Loss Aversion with Gain Utility, $(r_x, r_y)$ on budget constraint | $(\eta_x + 1\{x < r_x\}\Lambda_x)(x-r_x)+$<br>$(\eta_y + 1\{y < r_y\}\Lambda_y)(y-r_y)$ | Yes | depends on parameters |
| Two-Dimensional Loss Aversion, any $(r_x, r_y)$ | $1\{x < r_x\}\Lambda_x(x-r_x)$<br>$+1\{y < r_y\}\Lambda_y(y-r_y)$ | 1.2 Fails | N/A |

*Notes:* The table summarizes the formulations of reference-dependent payoffs considered in the appendix. Column (1) shows the functional form of reference-dependent payoffs for each formulation. Columns (2) and (3) describe the features of each formulation that pin down the sign of key welfare effects: whether the formulation satisfies Assumptions 1 and 2, and the which of the three cases from Lemma 1 obtains.

**TABLE B2: PAYOFF FORMULATIONS AND THE WELFARE EFFECT OF CHANGING REFERENCE POINTS**

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Welfare Effect $w_r(p,r)$ by Domain | | | |
| Description | Gain Domain $(x > r)$ | Reference Domain $(x = r)$ | Loss Domain $(x < r)$ | Individually Optimal Reference Points |
| Simple Loss Aversion | $0$ | $u'(r) - p$ | $-\pi\Lambda$ | $(-\infty, r^*]$ |
| Loss Aversion with Gain Utility | $-\pi\eta$ | $u'(r) - p$ | $-\pi(\eta + \Lambda)$ | $\pi = 0 : (-\infty, \tilde{r}]$<br>$\pi = 1 : -\infty$ |
| Utils Formulation (Köszegi-Rabin) | $-\pi\eta u'(r)$ | $u'(r) - p$ | $-\pi(\eta + \Lambda)u'(r)$ | $\pi = 0 : (-\infty, \tilde{r}]$<br>$\pi = 1 : -\infty$ |
| Gain Discounting | $\pi\Gamma$ | $u'(r) - p$ | $0$ | $[r^*, \infty)$ |
| Simple Loss Aversion with Diminishing Sensitivity | $0$ | $u'(r) - p$ | $-\pi\Lambda(r-x)^{\alpha-1}$<br>$+(1-\pi)\Lambda(r-x)^{\alpha-1}x_r$ | $\pi = 0$: $(-\infty, r^*), +\infty$<br>$\pi = 1$: $(-\infty, r^*)$ |
| Loss Aversion with Gain Utility & Diminishing Sensitivity | $-\pi\eta(x-r)^{\alpha-1}$<br>$+(1-\pi)\eta(x-r)^{\alpha-1}x_r$ | $u'(r) - p$ | $-\pi(\eta + \Lambda)(r-x)^{\alpha-1}$<br>$+(1-\pi)(\eta + \Lambda)(r-x)^{\alpha-1}x_r$ | $\pi = 0$: $-\infty, +\infty$<br>$\pi = 1 : -\infty$ |
| Two-Dimensional Loss Aversion, $(r_x, r_y)$ on budget constraint | $\pi\Lambda_y p$ | $u'(r) - p$ | $-\pi\Lambda_x$ | $r_x = r_x^*$<br>$r_y = r_y^*$ |
| Two-Dimensional Loss Aversion with Gain Utility, $(r_x, r_y)$ on budget constraint | $\pi(\eta_y p - \eta_x + \Lambda_y p)$ | $u'(r) - p$ | $\pi(p\eta_y - \eta_x - \Lambda_x)$ | See Appendix B.6.2 |
| Two-Dimensional Loss Aversion, any $(r_x, r_y)$ | $0$ | $u'(r) - p$<br>$-1\{y < r_y\}\pi p\Lambda_y$ | $-\pi\Lambda_x$ | $r_x \in (-\infty, r_x^*]$<br>$r_y \in (-\infty, r_y^*]$ |

*Notes:* The table evaluates welfare effects of changes in the reference point and describes individually optimal reference points for the payoff formulations from Table B1. Columns (1) to (3) evaluates the marginal welfare effect of changing the reference point $w_r(p,r)$ in the Gain, Reference, and Loss Domains. Note that for specifications with diminishing sensitivity, we do not express the behavioral response $x_r$ in terms of primitives in the table due to space constraints. See Appendix B.5 for details. Column (4) shows the set of individually optimal reference points under each formulation, where $r^*$ is the intrinsic optimum characterized by $u'(r^*) = p$ ($u'(r_x^*) = p; r_y^* = z - pr_x^*$ in the two-dimensional case), and $\tilde{r}$ is the reference point at the boundary between the gain and reference domain. Under two-dimensional loss aversion with gain utility and a reference point on the budget constraint, any of the cases from Lemma 1 could apply (see Table B1), and due to space constraints we defer the characterization of optimal reference points in this case to Appendix B.6.2.

## B.1 Simple Loss Aversion

We begin with the formulation we refer to as Simple Loss Aversion in the main text. Reference-dependent payoffs $v(x, r)$ are given by

$$v(x, r) = 1\{x \leq r\}\Lambda(x - r). \tag{24}$$

Thus, reference dependence makes the individual averse to losses over good $x$; the strength of this motive is governed by $\Lambda$. With this formulation, $v_x = \Lambda 1\{x < r\}$. This is weakly positive everywhere, so we are in the Everywhere Increasing case from Lemma 1. Since $v_x$ is also weakly positive in the loss domain and weakly negative in the gain domain, the Single-Peaked case also obtains. Hence, both Propositions 1.1 and 1.2 apply. The welfare effects of increasing $r$ are weakly negative everywhere, but they are zero in the gain domain, so the set of individually optimal reference points is $(-\infty, r^*]$. These results essentially follow from Proposition 1, given the properties of Simple Loss Aversion. Nevertheless, we work through a characterization of behavior and welfare in more detail. Unlike the main text, we will allow for heterogeneity across individuals indexed by $i$ from the outset.

**Demand.** We begin by describing demand $x_i(p, r)$ under Simple Loss Aversion. We first characterize potentially optimal choices in the gain domain ($x_i^G$) and in the loss domain ($x_i^L$) as follows:

$$u_i'(x_i^G(p)) = p, \tag{25}$$

$$u_i'(x_i^L(p)) + \Lambda_i = p. \tag{26}$$

Because $u_i'' < 0$ and $\Lambda_i > 0$, $x_i^G(p) < x_i^L(p)$, i.e. loss aversion increases demand in the loss domain relative to demand in the gain domain. Demand of a given individual is

$$x_i(p, r) = \begin{cases} x_i^G(p), \text{ if } x_i^G(p) > r & (G) \\ x_i^L(p), \text{ if } x_i^L(p) < r & (L) \\ r, \qquad \text{otherwise.} & (R) \end{cases} \tag{27}$$

Thus, at any given price and reference point, there are three groups of individuals, namely those whose demand is in the gain domain ($G$), in the loss domain ($L$), or at the reference point ($R$):

$$\begin{aligned} G(p, r) &\equiv \{i | x_i^G(p) > r\} = \{i | u_i'(r) > p\} \\ L(p, r) &\equiv \{i | x_i^L(p) < r\} = \{i | u_i'(r) + \Lambda_i < p\} \\ R(p, r) &\equiv \{i | x_i^G(p) \leq r \leq x_i^L(p)\} = \{i | u_i'(r) < p < u_i'(r) + \Lambda_i\}. \end{aligned}$$

**The Marginal Internality.** As we discuss in Section 2.2, a key statistic for welfare is the marginal internality, which is defined as the money metric welfare effect of a marginal change in $x$ along the budget constraint, $m_i(p, r; \pi) \equiv \frac{dU_i^*(x, z_i - px)}{dx}\Big|_{x = x_i(p,r)}$. Using the first-order conditions in equations (25) and (26) and the behavioral characterization in (27), it is straightforward to derive the following:

- If $x_i(p, r) > r$, $m_i(p, r; \pi) = 0$.
- If $x_i(p, r) < r$, $m_i(p, r; \pi) = -(1 - \pi)\Lambda_i$
- If $x_i(p, r) = r$,
  - $m_i(p, r; \pi)$ is undefined when $\pi = 1$.

– $m_i(p, r; \pi) = u_i'(r) - p$ when $\pi = 0$, with $-\Lambda_i \leq m_i \leq 0$

When the planner judges that observed demand is welfare-maximizing ($\pi = 1$), there is no marginal internality as a consequence of the envelope theorem. The marginal internality is undefined when $x = r$ in this case because of the kink in utility at $x = r$, but it remains the case that no deviation from observed behavior would improve welfare. When $\pi = 0$, in contrast, individuals with $x_i \leq r$ are over-consuming good $x$ out of loss aversion, so the marginal internality is negative.

**Main Welfare Effects.** Panel (a) of Figure 2 describes observed demand and intrinsic demand $u'(x)$ under Simple Loss Aversion. Under $\pi = 0$ the marginal internality is the vertical distance between marginal utility and the price at observed demand. The first row of Table B2 shows the welfare effects of changing reference points, which follow directly from equations (4) and (6). One can also derive them from first principles using the same set of steps used in equations (4) to (6). The welfare effects of price changes follow from equation (8):

$$w_{i,p} = -x_i - (1 - \pi)1\{x_i < r\}\Lambda x_{i,p} \tag{28}$$

Figure B1 provides a detailed illustration of the welfare effects of changes in references points and prices under Simple Loss Aversion, building on Panel (a) of Figure 2. In this model, individuals generally prefer lower reference points because they shrink losses. In the loss domain, changing $r$ has no effect on behavior but there is a direct welfare effect that matters under $\pi = 1$: increasing $r$ increases the individual's reference-dependent losses. When $\pi = 0$, increasing $r$ worsens over-consumption of good $x$ out of loss aversion, generating a negative behavioral welfare effect. The behavioral effect only materializes in the reference domain. Elsewhere, changing $r$ does not affect behavior. When $r \leq r^*$ at price $p$, all direct and behavioral effects are zero in this model, so any reference point at or below $r^*$ is individually optimal. In summary, lowering reference points robustly increases welfare regardless under Simple Loss Aversion.

Figure B2 illustrates the welfare effects of price changes. When $\pi = 0$, over-consumption of good $x$ generates a negative internality in the loss domain, and because increasing the price decreases consumption of good $x$, we obtain a positive behavioral welfare effect. In addition, there is always a standard negative direct welfare effect. Note that in the $R$ domain, demand is locally inelastic, so we find only a direct effect.

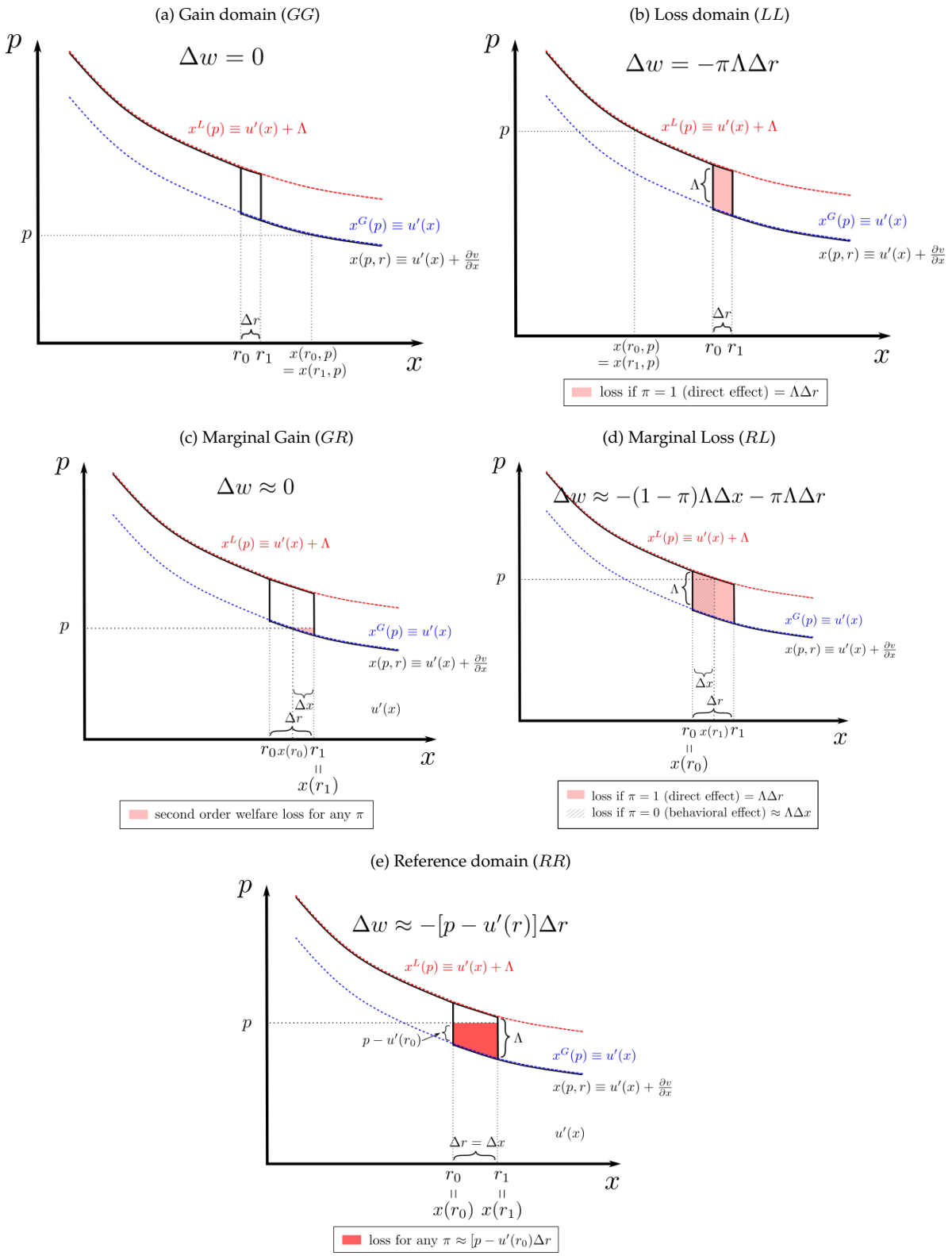
**Optimal Corrective Taxes.** The corrective tax schedule for good $x$ that maximizes social welfare for a given a reference point $r$ is characterized by

$$T(x, p, r) = \begin{cases} 0 & x \geq r \\ t^*(p, r)(x - r) & x < r; \end{cases} \tag{29}$$

$$t^*(p, r) = (1 - \pi)\frac{E\left[\Lambda_i \frac{\partial x_i^L}{\partial p} \middle| i \in L(p + t^*(p, r), r)\right]}{E\left[\frac{\partial x_i^L}{\partial p} \middle| i \in L(p + t^*(p, r), r)\right]}. \tag{30}$$

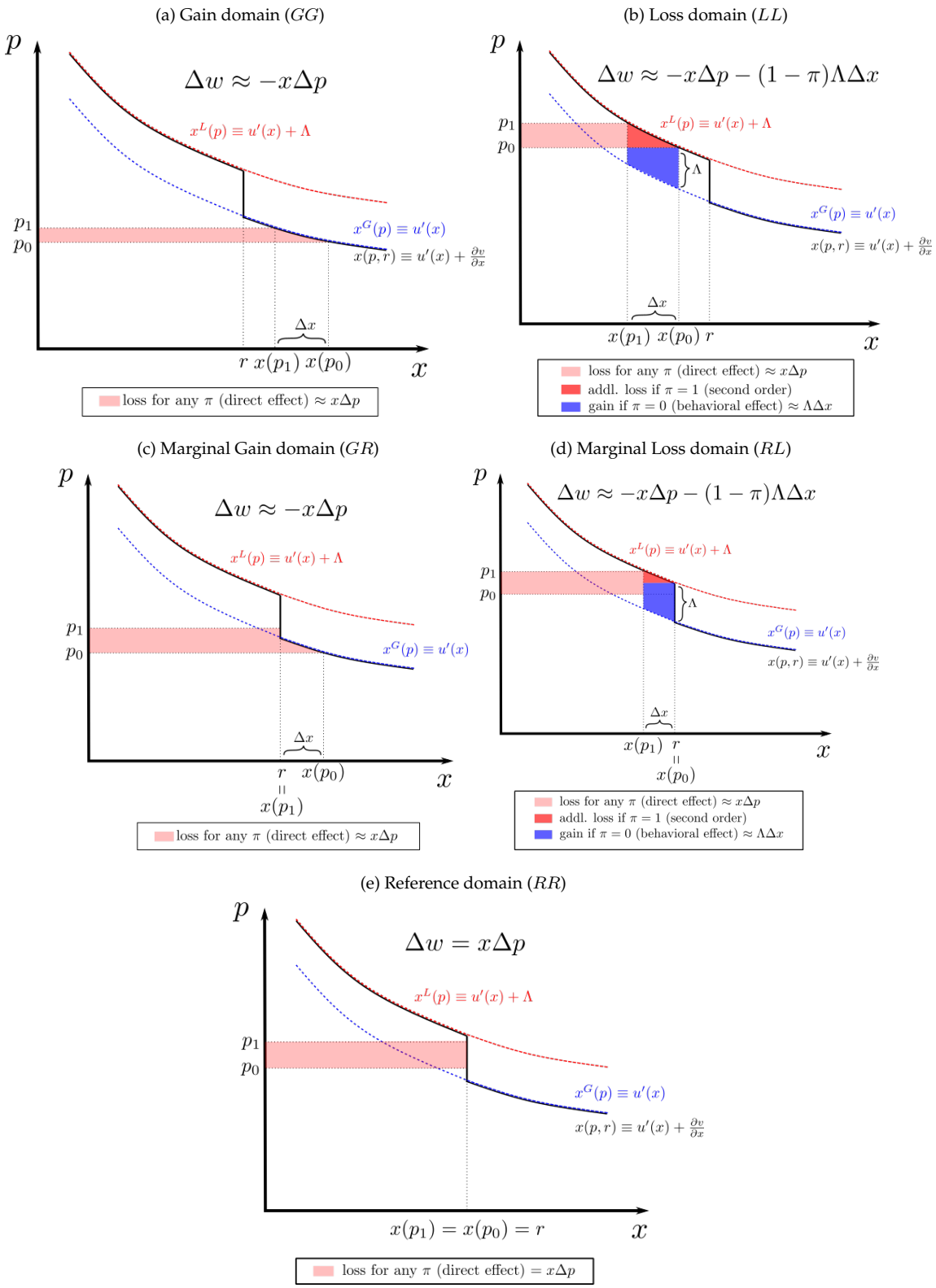
When reference dependence carries full normative weight ($\pi = 1$), there is no scope for corrective taxation, as individuals are making optimal choices in this case. When reference dependence is judged as a bias ($\pi = 0$), on the other hand, it is efficient to tax losses, i.e. to tax consumption of $x$ in the loss domain, because the tax should be set proportionally to marginal internalities (Mullainathan et al., 2012; Allcott and Taubinsky, 2015).Equation (30) quantifies the optimal corrective tax in the loss domain. The expression corresponds to what Allcott and Taubinsky (2015) call the *average marginal bias*. When the strength of reference

**FIGURE B1: WELFARE EFFECTS OF CHANGING THE REFERENCE POINT UNDER SIMPLE LOSS AVERSION**




*Notes:* The figure illustrates the welfare effects of changing the reference point under Simple Loss Aversion, in the domains indicated by the panel titles. We denote observed demand in black and gain and loss domain demand in blue and red, respectively, as in Panel (a) of Figure 2. All welfare changes are losses given by the areas shaded in red, reflecting the result that increasing the reference point unambiguously decreases welfare. Welfare losses due to direct effects are depicted in light red shaded areas, while losses due to behavioral effects are shaded with diagonal hatching. In panel (e), the change in welfare in the $RR$ case is the same regardless of $\pi$, but whether the depicted welfare loss represents a behavioral welfare effect or a direct welfare effect depends on $\pi$, so we use dark red shading. The legend of each panel provides further interpretation of the main welfare effects.

## FIGURE B2: WELFARE EFFECTS OF CHANGING PRICES UNDER SIMPLE LOSS AVERSION




*Notes*: The figure illustrates the welfare effects changing prices under Simple Loss Aversion, in the domains indicated by the panel titles. We denote observed demand in black and gain and loss domain demand in blue and red, respectively, as in Panel (a) of Figure 2. Red shaded areas denote welfare losses and blue shaded areas denote welfare gains. The legend of each panel provides further interpretation of the main welfare effects.

dependence and the demand response to a price change are independent, the optimal corrective tax simplifies to the average value of $\Lambda_i$ among individuals in the loss domain. Otherwise, the covariance between $\Lambda_i$ and the demand response has to be taken into account.[39]

## B.2 Loss Aversion with Gain Utility

The Simple Loss Aversion formulation is based on the model of reference dependence in riskless choice by Tversky and Kahneman (1991), but their specification incorporates an additional feature: a reference-dependent payoff over gains. In the case of Loss Aversion with Gain Utility, reference-dependent payoffs are given by

$$v_i(x,r) = \begin{cases} \eta_i(x-r) & x > r \\ \eta_i\lambda_i(x-r) & x \leq r, \end{cases} \tag{31}$$

The parameter $\eta_i$ can be interpreted as governing the overall importance of reference dependence, while $\lambda_i$ governs the strength of loss aversion. Incorporating $\eta_i$ makes the individual consume more $x$ by virtue of comparing their consumption to the reference point both in the gain and the loss domain.

**Behavioral Isomorphism to Simple Loss Aversion.** A key reason why we mainly discuss Simple Loss Aversion as an example of simple models of reference dependence is that Loss Aversion with Gain Utility is behaviorally behaviorally indistinguishable from Simple Loss Aversion. We establish this result formally here.

Consider a demand function $x(p,r,z)$, which describes the choice of $x$ the consumer makes for any $(p,r,z)$. Note that we drop $i$ subscripts and focus on one individual. We say $x(p,r,z)$ is *rationalizable* by a model if there are utility functions and parameters such that the optimization problem the model describes generates the observed behavior for any $(p,r,z)$. That is, $x(p,r,z)$ is rationalizable by Simple Loss Aversion if and only if there is a utility function $u(x)$ with $u' > 0, u'' < 0$ and a parameter $\Lambda > 0$ such that for any $(p,r,z)$ the solution to the consumer decision problem from equation (1) is $x(p,r,z)$. On the other hand, $x(p,r,z)$ is rationalizable by Loss Aversion with Gain Utility under analogous conditions, using $\tilde{u}$ to denote utility over good $x$ with this model when we compare across formulations.

We need one modest technical assumption for our result to obtain, which is that the domain of good $x$ is compact. For Simple Loss Aversion, this ensures that $u'(x)$ has a strictly positive minimum for all values of $x$, which we denote $\epsilon \equiv \min u'(x)$. The assumption ensures $\epsilon > 0$ exists.

**Proposition 4.** ***Behavioral Equivalence of Simple Loss Aversion and Loss Aversion with Gain Utility.*** *A demand function $x(p,r,z)$ is rationalizable by Simple Loss Aversion if and only if it is rationalizable by Loss Aversion with Gain Utility.*

**Corollary 4.1.** ***The Isomorphism.*** *If $x(p,r,z)$ is rationalizable by Simple Loss Aversion with utility $u(x)$ and parameter $\Lambda$ and rationalizable by Loss Aversion with Gain Utility with $\tilde{u}(x)$ and parameters $\eta$, $\lambda$, then we must have*

$$u(x) = \tilde{u}(x) + \eta x. \tag{32}$$

[39] To see how the covariance matters, we can re-write equation (30) as

$$t^*(p,r) = (1-\pi)\left\{ E\left[\Lambda_i \middle| i \in L(p + t^*(p,r), r)\right] + \frac{Cov\left[\Lambda_i, \frac{\partial x_i^L}{\partial p} \middle| i \in L\right]}{E\left[\frac{\partial x_i^L}{\partial p} \middle| i \in L\right]} \right\}.$$

$$\Lambda = \eta(\lambda - 1). \tag{33}$$

*Proof.* First suppose that $x_i(p, r, z)$ is rationalizable by Simple Loss Aversion with some utility function $u(x)$ and parameter $\Lambda$.

Set any $\eta$ such that $0 < \eta < \epsilon$.[40] Specify $\tilde{u}$ according to equation (32), i.e. $\tilde{u} = u(x) - \eta x$. Specify $\lambda_i$ according to equation (33), i.e. $\lambda_i = \frac{\Lambda_i + \eta_i}{\eta_i}$.

Because $u' > \eta$ for any $x$ by construction, we know that $\tilde{u}' = u' - \eta >= u' - \varepsilon > 0$, and $u'' < 0 \implies \tilde{u}_i'' < 0$. Further, by construction $\eta > 0$ and $\lambda > 1$. With the necessary restrictions satisfied, we only need to show that with these specifications, the optimization problem under Simple Loss Aversion is equivalent to the optimization problem under Loss Aversion with Gain Utility. As we have guaranteed equations (32) and (33) hold, we can re-express decision utility under Simple Loss Aversion as:

$$U(x) = \tilde{u}(x) + \eta x + z - px + \mathbb{1}\{x < r\}\eta(\lambda - 1)(x - r), \tag{34}$$

Next note that as it has no effect on the optimal $x$, we may freely eliminate $-\eta r$ from the maximand. Doing so and re-arranging yields the objective under Loss Aversion with Gain Utility.

For the converse, suppose that $x(p, r, z)$ is rationalizable by Loss Aversion with Gain Utility with utility function $\tilde{u}(x)$ and parameters $\eta > 0$, and $\lambda > 1$. Specify $u(x)$ using equation (32) and set $\Lambda$ using (33). Checking the restrictions, we know that $\tilde{u}' > 0$ and $\eta > 0$, implying that $u' = \tilde{u}' + \eta > 0$, and $u'' = \tilde{u}'' < 0$. And we know that $\Lambda > 0$ by $\eta > 0$ and $\lambda > 1$. We can re-express the optimization problem in Loss Aversion with Gain Utility as

$$U(x) = \tilde{u}(x) + \eta x + z - px + 1x > r\eta(\lambda - 1)(x - r) - \eta r. \tag{35}$$

The last term has no bearing on the optimum so we can eliminate it. Applying our constructed $u_i(x)$ and $\Lambda_i$ then yields the objective under Simple Loss Aversion. □

**Reparameterization.** Before we characterize demand and welfare under Loss Aversion with Gain Utility, we note that we can re-parameterize the payoff function from equation (31) as follows:

$$\tilde{U}_i(x, y) = \tilde{u}_i(x) + y + \tilde{v}_i(x|r), \tag{36}$$

$$\tilde{v}_i(x|r) = \begin{cases} \eta_i(x - r), & x > r \\ [\eta_i + \Lambda_i](x - r), & x < r, \end{cases} \tag{37}$$

The reparameterized version of the model is fully equivalent both in terms of behavior and welfare to the original Tversky and Kahneman (1991) formulation from equation (31), but slightly more convenient to work with below. As such, it is of course still behaviorally isomorphic to Simple Loss Aversion.

**Demand.** Panel (b) of Figure 2 illustrates demand in the reparameterized Loss Aversion with Gain Utility model. Given the behavioral equivalence result above, it is not surprising that the same basic characterization of demand arises. Due to the different parametric structure, first-order conditions are modified, though:

$$u'(x_i^G(p)) + \eta_i = p, \tag{38}$$

$$u'(x_i^L(p)) + (\eta_i + \Lambda_i) = p. \tag{39}$$

[40] The fact that an arbitrary $\eta$ can be chosen in this step is directly related to the fact that $\eta$ is typically unidentified from observations of observed demand.

Again, because $u_i'' < 0$ and $\eta_i + \Lambda_i > \eta_i$, $x_i^G(p) < x_i^L(p)$, i.e. loss aversion increases demand in the loss domain relative to demand in the gain domain. An analogue to equation (27) obtains but with the modified gain- and loss-domain demand curves from equations (38) and (39).

**Welfare.** Note that $v_x = \eta$ in the gain domain and $v_x = \eta + \Lambda$ in the loss domain. Both of these effects are positive, so we are in the Everywhere Increasing case from Lemma 1; unlike Simple Loss Aversion the Single-Peaked case does not apply, though. Proposition 1.1 then implies that decreasing $r$ is welfare-improving, and when $\pi = 1$ the inequality is strict: increasing $r$ has the direct effect of making reference-dependent losses larger and gains smaller, and this has a non-zero effect in all domains. Regarding behavioral welfare effects, when $\pi = 0$, we also find negative internalities in both the gain and loss domain. However, decreasing $r$ outside the reference domain has no effect on behavior and thus no effect on welfare under $\pi = 0$. Letting $\tilde{r}$ denote the lowest possible reference point in the reference domain, which is characterized by $u_i'(\tilde{r}) + \eta = p$, we have that any $r \in (-\infty, \tilde{r}]$ is individually optimal.

Figures B3 and B4 unpack the welfare effects of changing reference points and prices under Loss Aversion with Gain Utility. Comparing Figures B3 and B1, and the analytic expressions in Table B2, we observe that our main welfare results are qualitatively similar under Loss Aversion with Gain Utility and Simple Loss Aversion. The sign of key welfare effects remains the same, and if anything, magnitudes become larger under Loss Aversion with Gain Utility. Under $\pi = 0$, welfare effects are exacerbated because negative internalities from over-consumption of $x$ are larger in the loss and reference domain and additionally present in the gain domain. Under $\pi = 1$, negative direct effects of increasing $r$ are also larger in the loss and reference domains and additionally present in the gain domain.

## B.3 Reference Dependence over Utils

Kőszegi and Rabin (2006) introduce a different formulation of reference-dependent payoffs where individuals compare utility from their consumption of $x$ to utility at the reference point, rather than comparing the amount of $x$ directly to $r$. This modification is in part motivated by the fact that the scaling of reference dependence parameters such as $\Lambda$ otherwise depends on the units of $x$, which can make comparisons of these parameters across dimensions of the menu space less intuitive. In terms of equation (2), a Köszegi-Rabin type formulation thus implies $\mu(z) = u(z)$ instead of $\mu(z) = z$ as we consider so far ($\nu$ remains the same).

**Setup.** With reference dependence over utils, payoffs $v_i(x, r)$ are

$$v_i(x, r) = \begin{cases} \eta_i[u_i(x) - u_i(r)] & x \geq r \\ (\eta_i + \Lambda_i)[u_i(x) - u_i(r)] & x < r \end{cases} \tag{40}$$

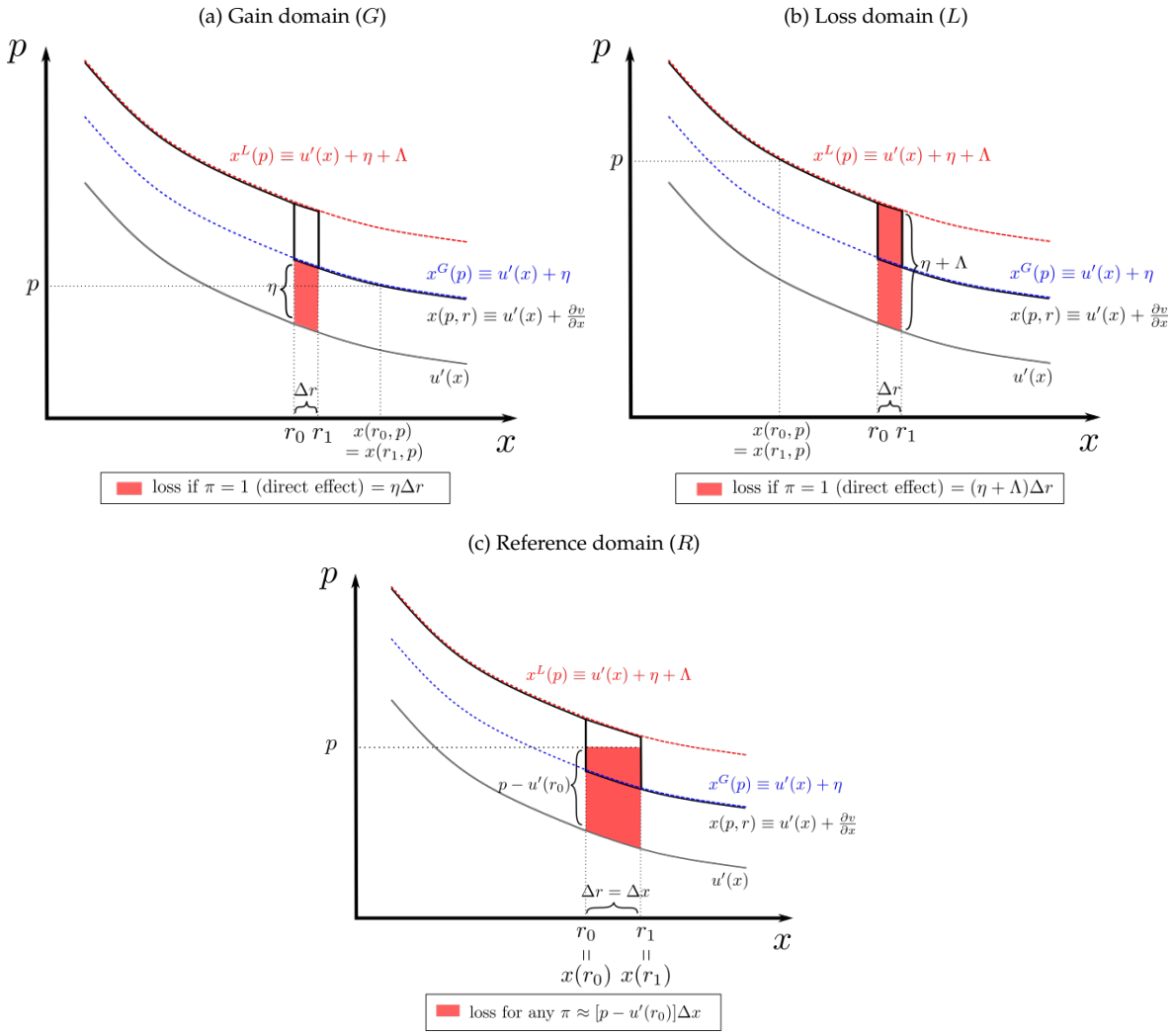
Note that we adopt a structure analogous to Loss Aversion with Gain Utility here, which is in line with Kőszegi and Rabin (2006). Alternatively, a version of Simple Loss Aversion over utils would also be straightforward to analyze.

**Demand.** We obtain a characterization of demand similar to Section B.2. The first-order conditions are

$$u_i'(x_i^G(p))(1 + \eta_i) = p, \tag{41}$$

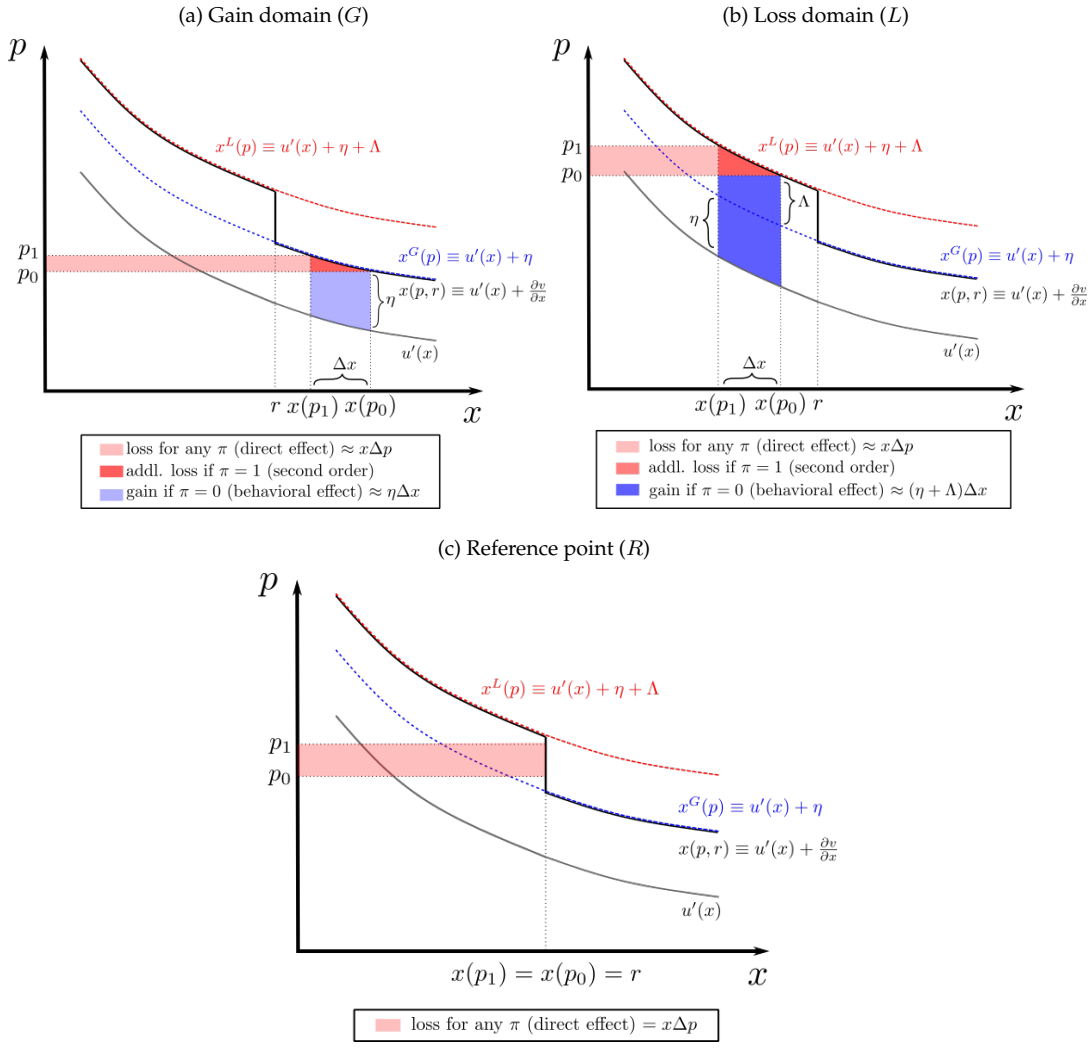
$$u_i'(x_i^L(p))(1 + \eta_i + \Lambda_i) = p. \tag{42}$$

**FIGURE B3: WELFARE EFFECTS OF CHANGING THE REFERENCE POINT UNDER LOSS AVERSION WITH GAIN UTILITY**




*Notes:* The figure illustrates the welfare effects of changing the reference point under Loss Aversion with Gain Utility, in the domains indicated by the panel titles. We denote observed demand in black, intrinsic demand in grey, and gain and loss domain demand in blue and red, respectively, as in Panel (b) of Figure 2. Red shaded areas denote welfare losses. The legend of each panel provides further interpretation of the main welfare effects.

**FIGURE B4: WELFARE EFFECTS OF PRICE CHANGES UNDER LOSS AVERSION WITH GAIN UTILITY**




*Notes*: The figure illustrates the welfare effects of changing prices under Loss Aversion with Gain Utility, in the domains indicated by the panel titles. We denote observed demand in black, intrinsic demand in grey, and gain and loss domain demand in blue and red, respectively, as in Panel (b) of Figure 2. Red shaded areas denote welfare losses and blue shaded areas denote welfare gains. The legend of each panel provides further interpretation of the main welfare effects.

Demand of a given individual once again falls into one of the three domains from equation (27), where the gain- and loss-domain demand curves are pinned down by equations (41) and (42). As Table B2 shows, the key properties of this formulation are very similar to Loss Aversion with Gain Utility, and the welfare effects of changing reference points and prices are qualitatively the same. Quantitative magnitudes can differ because unlike before, observed demand in the gain and loss domains and intrinsic demand are not parallel any more. They are locally parallel around the reference point, which reflects the approximation result from Proposition 2 of Kőszegi and Rabin (2006). These nonlinearities matter mainly for the direct welfare effects of changing reference points (i.e. when $\pi = 1$) for individuals far away from the reference point. Behavioral effects, which occur around the reference point, are less affected. Since welfare effects are similar to Figure 2b, we do not include a separate graphical illustration of reference dependence over utils.

## B.4 Gain Discounting

The literature on reference-dependent preferences typically interprets empirical patterns such as bunching at the reference point to loss aversion, which modifies payoffs over consumption of $x$ in the gain domain. Accordingly, the formulations of reference-dependent payoffs we considered so far fall in the Everywhere Increasing case from Lemma 1. However, in principle, these behavioral patterns could also be explained by an opposite-signed modification of payoffs over consumption in the gain domain. In other words, rather than consuming more of good $x$ in the loss domain in order to reduce losses, individuals could be consuming less of good $x$ in the gain domain because they discount gains. In this section, we lay out a possible formulation along these lines, which we call Gain Discounting.

$$v(x,r) = \begin{cases} -\Gamma(x-r), & x \geq r \\ 0, & x < r. \end{cases} \tag{43}$$

where the parameter $\Gamma$ governs the strength of gain discounting similarly to $\Lambda$ in the Simple Loss Aversion model. It is straightforward to verify that this payoff formulation satisfies Assumption 1 and 2. As before, the case-wise characterization of behavior in Equation (27) obtains. The first-order conditions in the gain and loss domains are given by

$$u_i'(x_i^G(p)) - \Gamma = p, \tag{44}$$

$$u_i'(x_i^L(p)) = p. \tag{45}$$

Comparing first-order conditions suggests that Gain Discounting model is behaviorally indistinguishable from Simple Loss Aversion. The formal proof is very similar to the one in Section B.3.

Observed demand and intrinsic demand under Gain Discounting are illustrated in Panel (d) of Figure 2. Perhaps unsurprisingly, adopting this formulation reverses the signs of all key welfare effects: we find positive direct welfare effects of increasing $r$ when $\pi = 1$ and positive behavioral welfare effects when $\pi = 0$. These effects now appear in the gain domain rather than the loss domain. Positive behavioral effects are driven by a positive marginal internality $(1-\pi)\Gamma$ in the gain domain, which reflects under-consumption of $x$ due to gain discounting.

Proposition 1.1 can be applied to Gain Discounting. But because the Everywhere Decreasing case from Lemma 1 obtains, the Proposition now implies that increasing $r$ improves welfare. Table B2 reports the welfare effects of changes in $r$ in detail and shows that any $r \geq r^*$ is individually optimal. The welfare

effects of price change is given by

$$w_{i,p} = -x_i + 1\{x_i > r\}(1-\pi)\Gamma x_p. \tag{46}$$

Again the sign of the behavioral welfare effect is reversed in equation (46), such that a price increase now lowers welfare.

The discussion about loss aversion vs. gain discounting is closely related to the framework by Bernheim (2009). In particular, one could view observed demand in the gain domain vs. the loss domain as demand under two different "frames". Thus, one could consider Simple Loss Aversion and Gain Discounting as two potential forms of preferences over $x$, where either demand in the gain domain or demand in the loss domain is judged to be normative. However, in terms of our framework, such an interpretation would impose $\pi = 0$ ex-ante. We provide a detailed discussion of our work and Bernheim (2009) in Appendix D.

## B.5 Incorporating Diminishing Sensitivity

Assumption 2.2 rules out diminishing sensitivity in our main analysis. This is motivated by the fact that empirical support for diminishing sensitivity in deterministic environments is limited (O'Donoghue and Sprenger, 2018). In this section, we describe how relaxing this assumption changes our welfare effects. Proposition 1 does not apply in this case, as the sign of direct and behavioral welfare effects can differ. Nevertheless, we can use similar steps to characterize welfare, and the characterization of optimal policy turns out not to be very different from other formulations. We specify the following formulation of reference-dependent payoffs:

$$v(x,r) = \begin{cases} \frac{1}{\alpha}\eta(x-r)^\alpha & x \geq r \\ -\frac{1}{\alpha}(\eta+\Lambda)(r-x)^\alpha & x < r \end{cases} \tag{47}$$

This specification adds diminishing sensitivity to the Loss Aversion with Gain Utility formulation from equation (31), whereby the previous formulation without diminishing sensitivity would be nested by $\alpha = 1$. In the following, we instead consider $\alpha \in (0,1)$. Compared to prior literature, we scale reference-dependent payoffs by $1/\alpha$, which does not matter for behavior and welfare and allows us to maintain the same interpretation of the $\Lambda$ and $\eta$ parameters as in the other formulations. Equation (47) has the key properties by which diminishing sensitivity is typically defined: $\nu' > 0$ everywhere, $\nu'' > 0$ when $x < r$ and $\nu'' < 0$ when $x > r$. As an alternative formulation, we could consider a variant of Simple Loss Aversion with diminishing sensitivity at the end of this section.

With this formulation, we continue to have case-wise demand in the gain, loss and reference domains. However, demand in the gain and loss domains now depends on both the price and the reference point. The first-order conditions are

$$u'(x^G(p,r)) + \eta(x^G(p,r) - r)^{\alpha-1} = p \tag{48}$$

$$u'(x^L(p,r)) + (\eta+\Lambda)(r - x^L(p,r))^{\alpha-1} = p \tag{49}$$

In previous formulations, there were no behavioral responses to a marginal change in the reference point in the gain and loss domains ($x_r^G = x_r^L = 0$), but with diminishing sensitivity there are such behavioral responses. Differentiating the first-order conditions with respect to $r$, we find

$$x_r^G = \frac{-\eta(1-\alpha)(x^G-r)^{\alpha-2}}{u''(x^G)-\eta(1-\alpha)(x^G-r)^{\alpha-2}} = \frac{\nu''}{u''+\nu''} \tag{50}$$

$$x_r^L = \frac{\eta(1-\alpha)(r-x^L)^{\alpha-2}}{u''(x^G)+\eta(1-\alpha)(r-x^L)^{\alpha-2}} = \frac{\nu''}{u''+\nu''}, \tag{51}$$

Inspecting these, we find that $x_r^G > 0$ everywhere in the gain domain. However, the sign of $x_r^L$ is ambiguous in the loss domain, where $x_r^L > 0$ for $x$ close to the reference point but $x_r^L < 0$ far from to $r$. Under a single-crossing condition (which is true with an isoelastic $u$, for instance), there are four relevant cases to consider. Ordered from the those obtaining at the lowest to highest $r$, these are:

1. The Gain domain ($G$), where $x_r > 0$
2. The Reference Domain ($R$), where $x_r = 1 > 0$
3. The low-reference point portion of the loss domain ($L_+$), where $x_r > 0$
4. The high-reference point portion of the loss domain ($L_-$), where $x_r < 0$ .

Proposition 1 would hold in the first three cases, but fails due to the fourth case. To understand how this matters for welfare, we return to the direct vs. behavioral effects characterization from equation (4), whose derivation does not require diminishing sensitivity. Note that for all the formulations we consider, including with diminishing sensitivity, $v_x$ and $v_r$ are opposite-signed. Provided $x_r \geq 0$ everywhere, which is true in all formulations satisfying Assumption 1 and 2, direct and behavioral welfare effects are ensured to be (weakly) same-signed, such that the sign of the total welfare effect does not depend on $\pi$. If $x_r < 0$ somewhere, however, the sign of the behavioral welfare effect changes, and thus the sign of the welfare effect can depend on $\pi$, unlike in Proposition 1. We therefore obtain the following characterization of the sign of the welfare effect of changing $r$:

- Under $\pi = 1$, $w_r < 0$ everywhere.
- Under $\pi = 0$,
    1. In $G$, $R$, and $L_+$, $w_r < 0$.
    2. In $L_-$, $w_r > 0$.

Hence, the welfare effects of increasing the reference point are generally negative as in formulations without sensitivity, but the sign changes for the case of the high-reference point part of the loss domain under $\pi = 0$. Building on this, the individually optimal reference point is the lowest possible one under $\pi = 1$. Under $\pi = 0$, there are two individually optimal reference points: the lowest and the highest possible reference point. To see why, note that the second term in equation (48) converges to zero both as $r \to -\infty$ and as $r \to \infty$. Consequently, behavior converges to the intrinsic optimum for either of these extreme reference points:

$$\lim_{r \to -\infty} x^G(p, r) = \lim_{r \to \infty} x^L(p, r) = r^*$$

Intuitively, as $r$ grows to either extreme, the individual stops chasing gains or avoiding losses because they are so far from the reference point that a marginally larger gain or loss does not matter much to them. Behavior thus converges to the intrinsic optimum, as if individuals did not care about reference dependence, and of course the intrinsic optimum is the optimal choice under $\pi = 0$. It is important to note that the lowest possible reference point is a robust choice in the sense that it is optimal regardless of $\pi$. However, the highest possible reference point is optimal only under $\pi = 0$, while it minimizes welfare under $\pi = 1$.[41]

[41] There is an interesting analogy to the welfare effects of default options in Goldin and Reck (2022). In the context of defaults, "penalty defaults" that promote active choices maximize welfare under $\pi = 0$ but minimize welfare under $\pi = 1$.

**Simple Loss Aversion with Diminishing Sensitivity.** As an alternative formulation, we consider a variant of Simple Loss Aversion with diminishing sensitivity. This can be done by simply setting $\eta = 0$ in equation (47). The modifies the welfare effects of changing the reference point as follows:

- Under $\pi = 1$, $w_r = 0$ for $(p, r) \in G$ and $w_r < 0$ everywhere else.
  1. In $G$, $w_r = 0$.
  2. Everywhere else, $w_r < 0$.
- Under $\pi = 0$,
  1. In $G$, $w_r = 0$.
  2. In $R$, $w_r < 0$.
  3. In $L_+$, $w_r < 0$.
  4. In $L_-$, $w_r > 0$.

The welfare effect of increasing $r$ is weakly positive everywhere except in the high-reference point part of the loss domain under $\pi = 0$. Thus, the individually optimal reference point is $(-\infty, r^*]$ under $\pi = 1$. Under $\pi = 0$, any reference point in $(-\infty, r^*]$ remains optimal, but another optimum is given by $r \to \infty$. As above, this implies that welfare effects can deviate from Simple Loss Aversion far away from the reference point.

## B.6 Two-Dimensional Reference Dependence

Some of the theoretical literature on reference dependence, including Tversky and Kahneman (1991) and Kőszegi and Rabin (2006), considers that reference dependence in more than one dimension. In this section, we examine formulations of reference-dependent payoffs over both good $x$ and $y$.

### B.6.1 Two-Dimensional Loss Aversion

**Setup.** Following prior literature, we assume that payoffs are additively separable across dimensions. We also assume that the formulation of payoffs is the same in each dimension but parameter values may differ. With two-dimensional payoffs, the reference point is two-dimensional: $r = (r_x, r_y)$. We begin by considering Simple Loss Averse in each dimension; we incorporate gain utility later on. We specify reference-dependent payoffs as

$$v(x, y, r) = 1\{x < r_x\}\Lambda_x(x - r_x) + 1\{y < r_y\}\Lambda_y(y - r_y)$$

It is useful to re-express reference-dependent payoffs as a function of $x$ only. To do this, let $r'_x = (z - r_y)/p$. Using the individual's budget constraint, we can express $v$ as a function of $x$ and the two reference points $r_x$ and $r'_x$.

$$v(x, r) = 1\{x < r_x\}\Lambda_x(x - r_x) - 1\{x > r'_x\}\Lambda_y p(x - r'_x). \tag{52}$$

Viewed in this reduced form, two-dimensional loss aversion resembles a combination of loss aversion over $x$ with reference point $r_x$ and gain discounting over $x$ with reference point $r'_x$. This insight helps us characterize welfare in the two-dimensional model, and to map two-dimensional reference dependence into our Flexible Reduced-Form model in the next section.

We will consider two types of variation in the two-dimensional reference point: changing $r_x$ or $r_y$ holding the other fixed, or varying both along the individual budget constraint. The latter is our focus in the main text, and in particular in the empirical application where the Normal Retirement Age can serve as a reference point in terms of leisure and consumption that lies on the budget constraint. If $(r_x, r_y)$ is on the budget constraint, $r_x = r'_x$ and the loss domain for good $y$ coincides completely with the gain domain for good $x$.

**Behavior.** In the case where the reference point falls on the budget constraint, there are three cases like in equation (27), with the following first-order conditions describing demand in the $G$ and $L$ domains:

$$\frac{u'(x^G(p))}{1+\Gamma} = p, \tag{53}$$

$$u'(x^L(p)) + \Lambda = p \tag{54}$$

Panel (c) of Figure 2 depicts observed and intrinsic demand in this model. Note that the graph becomes identical to Simple Loss Aversion when $\Lambda_y = 0$ and very similar to Gain Discounting when $\Lambda_x = 0$.

If the reference point falls in the interior of the budget constraint (implying $r_{x'} > r_x$), the individual's optimal choice will avoid any reference-dependent losses and we have simply $u'(x_i(p, r)) = p$ at the optimum.

If the reference point instead falls outside the budget constraint (implying $r_{x'} < r_x$), we have five behavioral cases instead of the three from equation (27). There are two reference domains: one where $x = r_x$, and one where $y = r_y \iff x = r'_x$. Then there are three first-order conditions describing demand in the gain and loss domains over $x$ and $y$:

$$\begin{aligned} u'(x^{LG}) + \Lambda_x &= p \ \ (L_x G_y) \\ \frac{u'(x^{LL}) + \Lambda_x}{1+\Lambda_y} &= p \ \ (L_x L_y) \\ \frac{u'(x^{GL})}{1+\Lambda_y} &= p \ \ (G_x L_y) \end{aligned}$$

We have $x^{GL} < x^{LL} < x^{LG}$. In total, the five cases for behavior are as follows:

$$x(p, r) = \begin{cases} x^{GL}, & x^{GL} < r_{x'} \\ r'_x, & x^{GL} < r'_x < x^{LL} \\ x^{LL}, & r_x < x^{LL} < r'_x \\ r_x, & x^{LL} < r_x < x^{LG} \\ x^{LG}, & x^{LG} > r_x \end{cases} \tag{55}$$

**Welfare.** Table B2 summarizes welfare effects of two types of variation in $(r_x, r_y)$: a change in $(r_x, r_y)$ along the budget constraint, and a change in $r_x$ holding $r_y$ fixed. In the first scenario, the Single-Peaked case from Lemma 1 obtains. Proposition 1.2 then implies that the unique optimal reference point is $(r_x^*, r_y^*)$. With two-dimensional reference dependence, $(r_x^*, r_y^*)$ is defined such that $u'(r_x^*) = p$ and $r_y^* = z - pr_x^*$.

Marginal internalities are positive in the gain domain and negative in the loss domain:

- If $x(p, r) > r$, $m(p, r; \pi) = (1 - \pi)\Lambda_y p$

- If $x(p,r) < r$, $m(p,r;\pi) = -(1-\pi)\Lambda_x$
- If $x(p,r) = r$,
  - $m(p,r;\pi)$ is undefined when $\pi = 1$.
  - $m(p,r;pi) = u'(r) - p$ when $\pi = 0$, with $-\Lambda_x \leq m \leq \Lambda_y p$

Intuitively, when $\pi = 0$, the individual under-consumes $x$ when $x > r$ in order to reduce reference-dependent losses over $y$, and over-consumes $x$ when $x < r$ to reduce losses over $x$.

Figure B5 illustrates the individual welfare effects of variation in the reference point, and Figure B6 illustrates price changes. Increasing the reference point generates direct positive welfare effects in the gain domain and direct negative impacts in the loss domain under $\pi = 1$. Behavioral welfare effects under $\pi = 0$ are concentrated in the reference domain and the sign of these effects turns on the location of the reference point relative to the individual's intrinsic optimum (similar to Figure 3 in the main text). The welfare effect of price changes combines the standard direct effect with behavioral effects depending on the sign of the marginal internality in each domain.

In the second scenario where $r_x$ changes *ceteris paribus*, the characterization of welfare is similar to Simple Loss Aversion. However, Assumption 1.2 fails: if $r < r_y$, it is not the case that $v = 0$ when $x = r_x$. Consequently, we cannot apply Proposition 1.1. In most cases, we find that lowering reference points is weakly welfare improving, but when $\pi = 0$, there is one case where $w_{r_x} > 0$ because increasing $r_x$ mitigates over-consumption of $x$ out of loss aversion over good $y$. This issue occurs when $\pi = 0$ in the fourth case from equation (55) (i.e. $x = r_x$) and the reference point lies outside the budget constraint in the subdomain where $u'(x) > p$. If $\Lambda_x \leq \Lambda_y p$, the condition $u'(x) > p$ is met whenever $x = r_x$; otherwise, the condition is met for sufficiently low prices. In this case, there is a positive internality from consuming more $x$ due to loss aversion over good $y$, so decreasing the reference point for $x$ does not improve welfare. Note that in whenever $w_{r_x} > 0$, it is alternatively possible to increase welfare by decreasing $r_y$ because the individual is incurring losses over good $y$. The individually optimal reference points are then the ones at which the individual avoids all losses: any $(r_x, r_y) \leq (r_x^*, r_y^*)$ is individually optimal.

### B.6.2 Two-Dimensionsal Loss Aversion with Gain Utility

Next, we incorporate gain utility into two-dimensional reference dependence. We consider the following payoff formulation:

$$v(x,y,r) = \begin{cases} \eta_x(x - r_x) + (\eta_y + \Lambda_y)(y - r_y), & x \geq r \\ (\eta_x + \Lambda_x)(x - r_x) + \eta_y(y - r_y) & x < r. \end{cases} \tag{56}$$

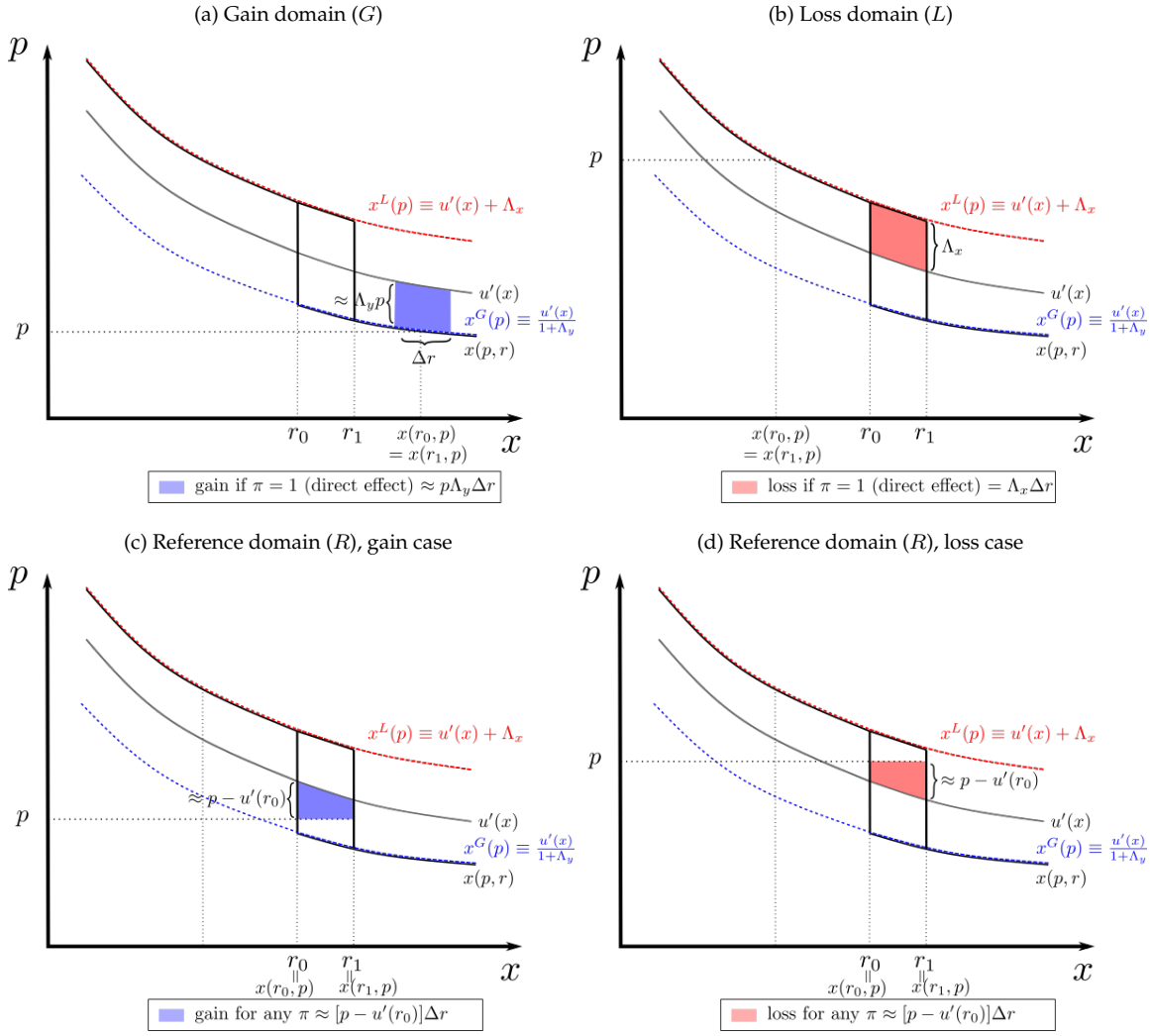
We focus on the first scenario from above, where the reference point is changed along the budget constraint ($r_{x'} = z - pr_y = r_x$). We can express $v$ as a function of $x$ and the reference point for $x$:

$$v(x,r) = \begin{cases} \eta_x(x - r_x) + p(\eta_y + \Lambda_y)(r_x - x), & x \geq r \\ (\eta_x + \Lambda_x)(x - r_x) + p\eta_y(r_x - x) & x < r. \end{cases} \tag{57}$$

Note that this formulation satisfies Assumptions 1 and 2, so we can apply Proposition 1. However, which case from Lemma 1 applies depends on parameter values. Differentiating $v$ yields

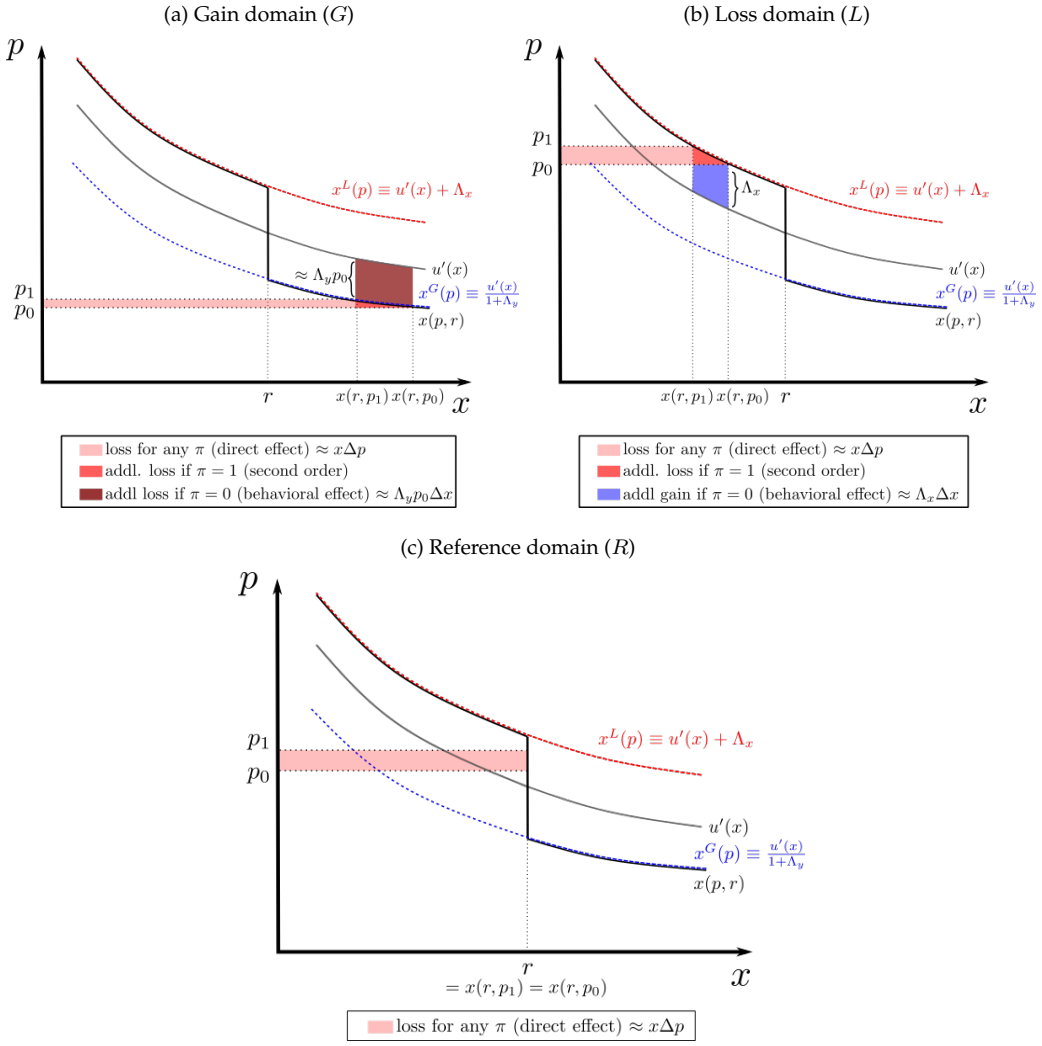
$$v_x = \begin{cases} \eta_x - p\eta_y - p\Lambda_y & x \geq r \\ \eta_x - p\eta_y + \Lambda_x & x < r. \end{cases} \tag{58}$$

**FIGURE B5: WELFARE EFFECTS OF CHANGING THE REFERENCE POINT UNDER TWO-DIMENSIONAL REFERENCE DEPENDENCE**




*Notes*: The figure illustrates the welfare effects of changing the reference point along the budget constraint under two-dimensional reference dependence. Effects are shown in gain domain (Panel a), the loss domain (Panel b) and the reference domain (Panels c and d). For the latter, we show effects separately for individuals experiencing a marginal gain and loss. We denote observed demand in black, intrinsic demand in grey, and gain and loss domain demand in blue and red, respectively, as in Panel (c) of Figure 2. Red shaded areas denote welfare losses and blue shaded areas denote welfare gains. The legend of each panel provides further interpretation of the main welfare effects. Note that because the size of the direct effect on the $G$ group depends on the price, so it is depicted slightly differently from Figure 3.

**FIGURE B6: WELFARE EFFECTS OF PRICE CHANGES UNDER TWO-DIMENSIONAL REFERENCE DEPENDENCE**




*Notes*: The figure illustrates the welfare effects of changing prices along the budget constraint under two-dimensional reference dependence. Effects are shown in gain domain (Panel a), the loss domain (Panel b) and the reference domain (Panel c). We denote observed demand in black, intrinsic demand in grey, and gain and loss domain demand in blue and red, respectively, as in Panel (c) of Figure 2. Red shaded areas denote welfare losses and blue shaded areas denote welfare gains. The legend of each panel provides further interpretation of the main welfare effects.

We can sign the derivative in order to find which case of Lemma 1 obtains.[42]

1. **Loss aversion dominates:** If $\eta_x < p\eta_y + p\Lambda_y$ and $\eta_x + \Lambda_x > p\eta_y$, we are in the Single-Peaked case.
2. $\eta_x$ **dominates:** If $\eta_x > p\eta_y + p\Lambda_y$, we are in the Everywhere Increasing case.
3. $\eta_y$ **dominates:** If $\eta_x + \Lambda_x < p\eta_y$, we are in the Everywhere Decreasing case.

These three cases are intuitive. When the value of a marginal gain is similar in both dimensions, $\eta_x \approx p\eta_y$, the model becomes equivalent to Simple Loss Aversion in two dimensions, which is in the Singled-Peaked case. With this restriction, equation (56) reduces to equation (52). In fact, we note that the restriction on the magnitude of payoff parameters across different dimensions proposed by Kőszegi and Rabin (2006) effectively imposes $\eta_x = p\eta_y$. As the inequalities above are strict, we find that the unique optimal reference point is the intrinsic optimum in this case, just as under Simple Loss Aversion in two dimensions.

The other two cases are those where the value of a marginal gain in one of the dimensions, $\eta_x$ or $p\eta_y$, is very large. When $\eta_x$ dominates, the individual consumes more $x$ in the gain domain over $x$ than intrinsic utility would imply because they chase gains over $x$. This lifts the gain domain demand curve above intrinsic demand $u'(x)$, different from Figures B5 and B6. The individual optimum tends toward extremes in this case. When $\pi = 1$, decreasing the reference point strictly improves welfare and the individual optimum is the lowest possible reference point. When $\pi = 0$, the individual optimum is any reference point that puts the individual in the gain domain for good $x$. Denoting the reference point at the boundary between the gain and the reference domain $\tilde{r}$ as before, lowering the reference point beyond $\tilde{r}$ has no effects on behavior or welfare.

Analogously, when $\eta_y$ is very strong, the individual chases gains over good $y$ in the loss domain for good $x$. In this case the individually optimal reference point tends toward the opposite extreme, namely high reference points. Under $\pi = 1$, increasing the reference point always improves welfare, while under $\pi = 0$, increasing the reference point improves welfare up to the boundary between the loss and reference domains.

## B.7 Our Flexible Reduced-Form Formulation as an Approximation

In this section, we formalize how our Flexible Reduced-Form specification is an approximation of any payoff formulation satisfying Assumptions 1 and 2. This also clarifies what can be empirically identified in situations where the true formulation is one of of those approximated by the Flexible Reduced Form.

The Flexible Reduced Form from equation (10) is given by

$$v(x,r) = \begin{cases} (1-\beta)\Lambda(x-r) & x \leq r \\ -\beta\Lambda(x-r) & x > r. \end{cases}$$

where $\Lambda > 0$ and $\beta \in [0,1]$

In the following, we show that (i) any formulation satisfying Assumptions 1 and 2 admits a first-order approximation via equation (10) with $\Lambda > 0$ equal to the size of the kink in preferences and $\beta \in \mathbb{R}$, and (ii) if in addition the formulation falls in the Singe-Peaked case from Lemma 1, then $\beta \in [0,1]$.

[42] That the weight of reference dependence parameters in these expressions depends on the price is due the fact that we specify reference dependence over the amount of the good. If instead we use a utils formulation or scale parameters by the price, this issue does not arise.

Suppose $v(x,r) = \nu(\mu(x) - \mu(r))$ satisfies Assumptions 1 and 2. The result we aim for follows from a first-order Taylor approximation of $v(x,r)$ about some point $(r_0, r_0)$. However, the non-differentiability in $v$ at points where $x = r$ necessitates using different Taylor approximations above and below $x = r$. Using Assumption 1.1, 1.2 and 1.3, we can approximate reference-dependent payoffs in both domains:

$$v(x,r) \approx \begin{cases} v(r_0,r_0) + \nu'_-(0)\mu'(r_0)(x-r_0) - \nu'_-(0)\mu'(r_0)(r-r_0) & x < r_0 \\ v(r_0,r_0) + \nu'_+(0)\mu'(r_0)(x-r_0) - \nu'_+(0)\mu'(r_0)(r-r_0) & x > r_0 \\ 0 & x = r_0 \end{cases} \equiv \hat{v}(x,r)$$

By Assumption 1.2, $v(r_0, r_0) = 0$. Simplifying, we have

$$\hat{v}(x,r) = \begin{cases} \nu'_-(0)\mu'(r_0)(x-r) & x \leq r_0 \\ \nu'_+(0)\mu'(r_0)(x-r) & x > r_0 \end{cases} \tag{59}$$

Let $\Lambda = \nu'_-(0)\mu'(r_0) - \nu'_+(0)\mu'(r_0)$. Note that this is the implied size of the kink in preferences around $x = r_0$ above. And let $\beta = -\nu'_+(0)\mu'(r_0)/\Lambda$. Note that $1 - \beta = \nu'_-(0)\mu'(r_0)/\Lambda$. Then the approximate formulation (59) becomes equation (10). All that remains to check are the parametric restrictions. We have $\Lambda > 0$ by Assumption 1.3. The parameter $\beta$ then turns on which of the cases from Lemma 1 obtains:

- $v(x,r)$ is Single-Peaked if and if only $\beta \in [0,1]$.
- $v(x,r)$ is Everywhere Increasing ($v_x > 0$ everywhere) if and if only $\beta < 0$.
- $v(x,r)$ is Everywhere Decreasing ($v_x < 0$ everywhere) if and if only $\beta > 1$.

Note that we have not relied on ruling out diminishing sensitivity here (Assumption 2), although we use sub-Assumption 2.1 in order to invoke the cases from Lemma 1. Because it is a first-order approximation, equation (59) has a second derivative of zero and thereby satisfies Assumption 2 automatically.

When we estimate $\Lambda$ and $\beta$ empirically, we should think of $r_0$ as the status quo reference point. For instance, this would be the pre-reform Normal Retirement Age in our empirical application. Welfare effects of changes in prices and reference points for individuals choosing options near the status quo are then well-approximated according to the logic of a first-order Taylor approximation, while welfare effects for those further away may be subject to larger approximation errors. This has some noteworthy implications for quantitative evaluations of changes in $r$. Namely, because behavioral welfare effects of variation in $r$ are concentrated near the reference point, these are insensitive to potential approximation errors. The same cannot be said for direct welfare effects of changing $r$, as these occur further away as well.

# C Proofs

This section presents proofs of all propositions and a few notes on the theory.

**Lemma 1.** *Under Assumptions 1 and 2.1, at least one of the following must be true:*

- *(Everywhere Increasing)* $v_x \geq 0$ *for all* $x \neq r$.
- *(Everywhere Decreasing)* $v_x \leq 0$ *for all* $x \neq r$.
- *(Single-Peaked)* $v_x \geq 0$ *for all* $x < r$, *and* $v_x \leq 0$ *for all* $x > r$.

*Proof.* Under our domain-specific monotonicity assumption, Ass. 2.1, there are four possibilities: $v_x$ may be positive or negative for all $x > r$, and it may be positive or negative for all $x < r$. $v_x$ being positive over gains and negative over losses would violate the direction of the kink in preferences under loss aversion (Assumption 1.3), as we approach the point where $x = r$ from the right or left. At least one of the other three cases must therefore obtain. □

**Proposition 1.** ***Signing the Welfare Effects of Reference Point Variation.*** *Maintain Assumptions 1 and 2 and consider any* $(p, r)$ *that is not on the boundary of* $R$.

**P1.1.** *If* $v$ *is Everywhere Increasing, then* $w_r(p, r) \leq 0$. *If* $v$ *is Everywhere Decreasing, then* $w_r(p, r) \geq 0$.

**P1.2.** *Let* $r^*$ *be the reference point such that* $u'(r^*) = p$. *If* $v$ *is Single-Peaked, then* $w_r(p, r) \geq 0$ *when* $r \leq r^*$, *and* $w_r(p, r) \leq 0$ *when* $r \geq r^*$. *Consequently,* $r^*$ *is an individually optimal reference point.*

*Proof.* Most of the key steps in Proposition 1 are covered in the main text.

The derivative we wish to characterize is expressed in equation (4) for the $G$ and $L$ domains, while equation (6) covers the $R$ domain.

Generically there are two candidates for optima in the interior of the $G$ and $L$ domain:

$$\begin{aligned} u'(x^L) + \nu'\mu'(x^L) &= p;\ x^L < r \\ u'(x^G) + \nu'\mu'(x^G) &= p;\ x^G > r \end{aligned} \tag{60}$$

We can derive (4) using these first-order conditions. Outside the $R$ domain,

$$w_r = (u'(x) - p + \pi v_x)x_r + \pi v_r.$$

The first order condition implies $u'(x) - p = -v_x$, so from the envelope condition we have

$$w_r = -(1 - \pi)v_x x_r + \pi v_r.$$

Substituting for $v_x$ and $v_r$ using equation (2) (as in equation (60)), we have

$$w_r = -\nu'\mu'(1 - \pi)x_r - \pi\nu'\mu'(r).$$

Note that we have not relied on Assumption 2.2 yet. As noted in the main text, equation (4) can be derived without this assumption. If we differentiate the first-order condition above and apply Assumption 2.2, we find $x_r = 0$ in the $G$ and $L$ domains. The above expression simplifies to $w_r = -\pi\nu'\mu'(r)$. Note that $\mu' > 0$ everywhere by Assumption 1.1. Thus if $v_x \geq 0 \iff \nu' \geq 0$ everywhere (the Everywhere Increasing case), $w_r \leq 0$ everywhere. If $v_x \leq 0$, $\nu' \leq 0$ and $w_r \geq 0$. In the Single-Peaked case, $\nu' \geq 0$ and $w_r \leq 0$ in the

loss domain, while in the gain domain $v_x \leq 0$ and $w_r \geq 0$. This establishes the result for the gain and loss domains.

Finally we establish the result for the $R$ domain. Recall that under Assumption 3, there is a range of values for $r$ where $x^G < r < x^L$ and $x(p,r) = r$, which defines the $R$ domain. From Assumption 1.3, the fact that $u'' < 0$, and the first-order conditions (60) we have

$$x^G < r < x^L \implies -\nu'(\mu(x^L) - r)\mu'(x^L) < u'(r) - p < -\nu'(\mu(x^G) - r)\mu'(x^G).$$

A version of this expression appears as equation (7) in the main text. In the everywhere increasing case, $u'(r) - p$ is bounded by two (weakly) negative quantities so it must be (weakly) negative. Likewise in the everywhere decreasing case, $u'(r) - p$ must be weakly positive. This completes the proof of Proposition 1.1.

In the Single-Peaked case, $u'(r) - p$ is bounded between a weakly positive and a weakly negative quantity, so by $u'' < 0$ there must be some $r^*$ with $(r^*, p) \in R$ and $u'(r) - p = 0$. We call this the intrinsic optimum in the main text. Obviously, $u'(r) - p > 0$ for $r < r^*$ and the opposite is true for $r > r^*$. This completes the proof for the $R$ case. □

**Remark on the Derivation of Equation (8).** We note that the expression for the welfare effects of price changes in equation (8) can be derived following identical steps to the derivation of the generic expressions for $w_r$ in the previous proof.

**Proposition 2.** ***Sufficient Statistics Characterizations***

**P2.1.** *Up to a first-order approximation, the social welfare effect of a change in the reference point $\Delta r$ is*

$$\begin{aligned}\Delta W \approx \Delta r \pi \left\{E[\beta_i \Lambda_i \mid i \in G] P[i \in G] - E[(1-\beta_i)\Lambda_i \mid i \in L] P[i \in L]\right\} \\ - \Delta r E\left[u_i'(r) - p \;\middle|\; i \in R\right] P[i \in R].\end{aligned}$$

**P2.2.** *If the distribution of $u_i'(r) - p$ is independent of $(\beta_i, \Lambda_i)$ and locally uniform for $i \in R(p,r)$, the social welfare effect of $\Delta r$ can be further approximated as*

$$\begin{aligned}\Delta W \approx \Delta r \pi \left\{E[\beta_i \Lambda_i \mid i \in G] P[i \in G] - E[(1-\beta_i)\Lambda_i \mid i \in L] P[i \in L]\right\} \\ + \Delta r E\left[\Lambda_i \left(\beta_i - \frac{1}{2}\right) \;\middle|\; i \in R\right] P[i \in R].\end{aligned}$$

**P2.3.** *Up to a first-order approximation, the social welfare effect of a change in price $\Delta p$ is*

$$\begin{aligned}\Delta W &\approx \Delta p (1-\pi) \left\{E\left[\beta_i \Lambda_i x_{p,i} \;\middle|\; i \in G\right] P[i \in G] - E\left[(1-\beta_i)\Lambda_i x_{p,i} \;\middle|\; i \in L\right] P[i \in L]\right\} \\ &\quad - \Delta p E[x_i(p,r)] \\ &= \Delta p (1-\pi) \left\{E\left[\beta_i \Lambda_i \varepsilon_i \frac{x_i}{p} \;\middle|\; i \in G\right] P[i \in G] - E\left[(1-\beta_i)\Lambda_i \varepsilon_i \frac{x_i}{p} \;\middle|\; i \in L\right] P[i \in L]\right\} \\ &\quad - \Delta p E[x_i],\end{aligned}$$

*where $\varepsilon_i$ is the price elasticity of demand for good $x$.*

*Proof. Proof of 2.1.* There are two main steps in this proof. The first step is to derive the individual welfare effect under the Flexible Reduced-Form specification (10) in each of the three domains. The second step is to show that welfare effects at the boundary of these cases are irrelevant for the first-order welfare effect, so that the social welfare effect is a simple aggregation of the welfare effects in the three domains.

To do this, we need only evaluate the derivatives in equation (4). As noted in the previous proof, we have $x_{i,r} = 0$ for $i \in G, L$ because equation (10) satisfies Assumption 2.2. For these cases therefore we have

$$\begin{aligned} i \in G(p,r) &\implies w_{r,i} = \pi\beta_i\Lambda_i \\ i \in L(p,r) &\implies w_{r,i} = -\pi(1-\beta_i)\Lambda_i \end{aligned}$$

while for the $R$ case, we already have the welfare effect from equation (6), which uses that $\nu(0) = 0$ under Assumption 1.2:

$$i \in R(p,r) \implies w_{r,i} = u_i'(r) - p.$$

Now we prove that the social welfare effect of a change in $r$ is simply the aggregation of the welfare effects for individuals in the three cases. As the three groups cover the full population, we can decompose social welfare as follows:

$$W(p,r) = \int_{i \in G(p,r)} w_i(p,r)dF(i) + \int_{i \in R(p,r)} w_i(p,r)dF(i) + \int_{i \in L(p,r)} w_i(p,r)dF(i).$$

Because $w_i(p,r)$ and behavior are everywhere continuous and differentiable almost everywhere, the result then follows immediately from applying the generalization of Leibniz integral rule for measure spaces, which implies

$$W_r(p,r) = \int_{i \in G(p,r)} w_{r,i}(p,r)dF(i) + \int_{i \in R(p,r)} w_{r,i}(p,r)dF(i) + \int_{i \in L(p,r)} w_{r,i}(p,r)dF(i).$$

Intuitively, at the two boundaries between $G$ and $R$ and between $R$ and $L$, we have a marginal change in welfare for a marginal group, so the effect is of second order. Using the expressions derived above for $w_{r,i}$ and restating the integrals in terms of conditional expectations, we arrive at the desired result.

Because some readers may be unfamiliar with more general versions of Leibniz integral rules, we also provide a less abstract argument under the assumption that there is a single dimension of heterogeneity $\theta_i$, and $\Lambda_i$ and $\beta_i$ are both homogeneous. By definition $u_i'(r) = u_r(r, \theta_i)$, so supposing without loss of further generality (beyond the one-dimensional types assumption) that a higher $\theta$ corresponds to a higher level of marginal utility, for any $(p,r)$ there are cutoffs $\theta_L$ and $\theta_G$ such that $u_r(r, \theta_L(r)) + (1-\beta)\Lambda = p$, and $u_r(r, \theta_L) - \beta\Lambda = p$. We find $\theta_L < \theta_G$ due to diminishing marginal utility. These cutoffs depend on $(p,r)$ so we express them as $\theta^L(p,r)$ and $\theta^G(p,r)$.

With this restriction on individual heterogeneity we can re-write the previous expression as

$$W(p,r) = \int_{\theta^G(p,r)}^{\infty} w(p,r,\theta)dF_\theta(\theta) + \int_{\theta^L(p,r)}^{\theta^G(p,r)} w(p,r,\theta)dF_\theta(\theta) + \int_{-\infty}^{\theta^L(p,r)} w(p,r,\theta)dF_\theta(\theta)$$

Differentiating with respect to $r$ and applying the one-dimensional Leibniz rule for integrals, we find

$$\begin{aligned} W(p,r) = &\int_{\theta^G(p,r)}^{\infty} w_r(p,r,\theta)dF_\theta(\theta) - \theta_r^G w(p,r,\theta^G)f_\theta(\theta^G) \\ &+ \int_{\theta^L(p,r)}^{\theta^G(p,r)} w_r(p,r,\theta)dF_\theta(\theta) + \theta_r^G w(p,r,\theta^G)f_\theta(\theta^G) - \theta_r^L w(p,r,\theta^L)f_\theta(\theta^L) \\ &+ \int_{-\infty}^{\theta^L(p,r)} w_r(p,r,\theta)dF_\theta(\theta) + \theta_r^L w(p,r,\theta^L)f_\theta(\theta^L) \end{aligned}$$

The boundary terms all cancel and the resulting expression is the desired result. This illustrates more concretely that the boundary cases are second order when evaluating welfare effects if welfare and behavior

evolve continuously at the boundary. For the proofs of the remaining parts of proposition 2, we take for granted that the boundary cases are second order.

*Proof of 2.2* Once we have established the previous result, all we need to do is show how the term for the $R$ group simplifies under the assumption that $\Delta_i = u_i'(r) - p$ is uniform conditional on $i \in R$. Without loss of generality we can write the welfare effect in the $R$ group as:

$$\int_{i \in R(p,r)} w_{r,i}(p,r) dF(i) = \int_\Lambda \int_\beta \int_{\Delta=-(1-\beta)\Lambda}^{\beta\Lambda} \Delta f(\Delta|\beta,\Lambda) d\Delta dF_{\beta,\Lambda}(\beta,\Lambda)$$

Now we apply the uniformity assumption, that is $f(\Delta|\beta,\Lambda)$ is constant in the $R$ domain. Under our conditional independence assumption the constant does not depend on $\beta, \Lambda$. Denoting the constant $C$, the above expression becomes

$$\int_{i \in R(p,r)} w_{r,i}(p,r) dF(i) = \int_\Lambda \int_\beta C \frac{-(1-\beta)\Lambda + \beta\Lambda}{2} dF_{\beta,\Lambda}(\beta,\Lambda)$$

Noting that $\frac{-(1-\beta)\Lambda+\beta\Lambda}{2} = \Lambda(\beta - \frac{1}{2})$ and that this expression represents a conditional expectation for $i \in R$ multiplied by $P[i \in R]$, we arrive at the desired result.

*Proof of 2.3* By the same argument as above, we have:

$$W_p(p,r) = \int_{i \in G(p,r)} w_{p,i} dF(i) + \int_{i \in R(p,r)} w_{p,i} dF(i) + \int_{i \in L(p,r)} w_{p,i} dF(i)$$

Evaluating the derivatives for the individual welfare effect of a price change in equation (8) under formulation (10), we find:

$$\begin{aligned}
i \in G(p,r) &\implies w_{r,i} = -x_i(p,r) - \pi\beta_i\Lambda_i x_{p,i} \\
i \in L(p,r) &\implies w_{r,i} = -x_i(p,r) + \pi(1-\beta_i)\Lambda_i x_{p,i} \\
i \in R(p,r) &\implies w_{r,i} = -x_i(p,r)
\end{aligned}$$

For the $R$ case, we are using the fact that welfare equals $w_i(p,r) = u_i(r) - pr$ locally to arrive at the above (note also $x_{p,i} = 0$ locally). Substituting in these expressions and simplifying yields the desired result. For the re-statement in terms of elasticities, we use the definition of the price elasticity: $\varepsilon_{p,i} = x_{p,i}\frac{p}{x_i(p,r)}$. □

**Proposition 3.** ***Identification from Bunching.*** *Define a random variable $\Delta_i = u_i'(r) - p$ and denote its density and cumulative distribution by $f_\Delta$ and $F_\Delta$. Assume that $\Delta$, $\Lambda$ and $\beta$ are mutually independent.*

**P3.1.** *Excess bunching at $x_i = r$ is characterized approximately by*

$$\frac{P[i \in R]}{f_\Delta(0)} \approx E[\Lambda_i]$$

**P 3.2.** *The share of bunching that comes from the right – defined as the share of individuals who would choose to consume more than $r$ in the absence of reference-dependent payoffs – is approximately*

$$P[r_i^* > r | i \in R] \approx E[\beta_i]$$

*These approximations are based on a first-order Taylor approximation of $F_\Delta$ around $\Delta = 0$; they are exact when $f_\Delta$ is locally uniform for $i \in R(p,r)$.*

*Proof.* We begin by building on the characterization of the composition of the three groups in terms of $\beta$, $\Lambda$, and $\Delta$.

$$P[i \in G] = Pr(\Delta > \beta\Lambda) = \int_{(\beta,\Lambda)} 1 - F_\Delta(\beta\Lambda|\beta,\Lambda) dF_{\beta,\Lambda}(\beta,\Lambda)$$
$$P[i \in L] = Pr(\Delta < -(1-\beta)\Lambda) = \int_{(\beta,\Lambda)} F_\Delta(-(1-\beta)\Lambda|\beta,\Lambda) dF_{\beta,\Lambda}(\beta,\Lambda)$$
$$P[i \in R] = Pr(-(1-\beta)\Lambda < \delta < \beta\Lambda) = \int_{(\beta,\Lambda)} F_\Delta(\beta\Lambda|\beta,\Lambda) - F_\Delta(-(1-\beta)\Lambda|\beta,\Lambda)$$

By the independence assumption, $F_\Delta(\Delta|\beta,\Lambda) = F_\Delta(\Delta)$. Then, using a first-order Taylor Approximation of $F(\Delta)$ around $\Delta = 0$, we have

$$Pr[i \in G] \approx \int_{\beta,\Lambda} \left[1 - F_\Delta(0) - f_\Delta(0)\beta\Lambda\right] dF_{\beta,\Lambda}(\beta,\Lambda)$$
$$Pr[i \in L] \approx \int_{\beta,\Lambda} \left[F_\Delta(0) + f_\Delta(0)(-(1-\beta)\Lambda)\right] dF_{\beta,\Lambda}(\beta,\Lambda)$$
$$Pr[i \in R] \approx \int_{\beta,\Lambda} \left[f_\Delta(0)(\beta\Lambda - -(1-\beta)\Lambda)\right] dF_{\beta,\Lambda}(\beta,\Lambda) = \int_{\beta,\Lambda} \left[f_\Delta(0)\Lambda\right] dF_{\beta,\Lambda}(\beta,\Lambda)$$

Note that these approximations are accurate when $f'_\Delta(\Delta) = 0$, i.e. when the distribution of $\Delta$ is uniform over the relevant domain. The above expressions simplify to

$$Pr[i \in G] \approx 1 - F_\Delta(0) - f_\Delta(0)E[\beta_i\Lambda_i]$$
$$Pr[i \in L] \approx F_\Delta(0) - f_\Delta(0)E[(1-\beta_i)\Lambda_i]$$
$$Pr[i \in R] \approx f_\Delta(0)E[\Lambda_i]$$

Now excess bunching at the reference point, defined as the probability $i \in R$ scaled by the probability that $r_i^* = r \iff \Delta = 0$, is given by

$$\frac{Pr[i \in R]}{f_\Delta(0)} \approx E[\Lambda_i].$$

Similarly to the characterization of $Pr[i \in R]$ above, the fraction of the full population who bunch from the right, i.e. those $i \in R(p,r)$ for whom $r_i^* > r \iff \Delta > 0$, is

$$P[r_i^* > r \;\&\; i \in R] = \int_{\beta,\Lambda} \left[f_\Delta(0)(\beta\Lambda - 0)\right] dF_{\beta,\Lambda}(\beta,\Lambda) \approx f_\Delta(0)E[\beta_i\Lambda_i].$$

Using the assumption that $\Lambda$ and $\beta$ are independent, we combine the previous two expressions to obtain the probability of bunching from the right conditional on $i \in R$, $P[r_i^* > r | i \in R] = P[\Delta_i > 0 | i \in R]$:

$$P[r_i^* > r | i \in R] = \frac{P[r_i^* > r \;\&\; i \in R]}{P[i \in R]} \approx E[\beta_i].$$

□

**Welfare Effects with Fiscal Externalities.**

Here we derive our main welfare effects in the presence of fiscal externalities. Doing so helps us understand the fiscal externality component of welfare effects in our empirical application. Our aim is to understand how incorporating fiscal externalities modifies equations (13) and (15) from the main text. Here we focus on the case of Simple Loss Aversion, i.e. we set $\beta = 0$. Relaxing this restriction would be straightforward, but we rely on this simplification because we only use the restricted version of the sufficient statistics formulas

in the empirical application.

With a fiscal externality, we can characterize efficiency using

$$\Delta W = \Delta W^{ind} + \Delta G$$

where $\Delta W^{ind}$ is the change in utilitarian social welfare approximated by the above results, $\Delta G$ is the change in government revenues. Note that because we focus on efficiency, we implicitly set the marginal cost of public funds equal to 1 here.

Suppose that good $x$ is taxed at some linear rate $t$. Then $\Delta G = \Delta E[t \cdot x_i]$. For a change that leaves tax incentives fixed, such as a ceteris paribus change in the reference point, $\Delta G = E[t\Delta x_i]$, and if the tax rate is fixed across individuals, we can express this as $\Delta G = tE[\Delta x_i]$.

Assuming such a uniform tax rate and considering a cetris paribus change in the tax rate, we have $E[\Delta x_i] \approx \Delta r P(i \in R)$, because individuals in the the $G$ and $L$ groups do not change behavior, the marginal gain and marginal loss cases are second order, and $\Delta x_i = \Delta r$ for $i \in RR$.

$$\begin{aligned}\Delta W \approx &- \Delta r \pi E[\Lambda_i \mid i \in L(p, r_0)] P[i \in L(p, r_0)] \\ &- \Delta r E\left[\frac{\Lambda_i}{2} \mid i \in R(p, r_0)\right] P[i \in R(p, r_0)] + t\Delta r P[i \in R(p, r_0)]\end{aligned}$$

Simplifying

$$\begin{aligned}\Delta W \approx &- \Delta r \pi E[\Lambda_i \mid i \in L(p, r_0)] P[i \in L(p, r_0)] \\ &+ \Delta r E\left[-\frac{\Lambda_i}{2} + t \mid i \in R(p, r_0)\right] P[i \in R(p, r_0)]\end{aligned}$$

With individual-specific marginal tax rates on good $x$, we would rather have

$$\begin{aligned}\Delta W \approx &- \Delta r \pi E[\Lambda_i \mid i \in L(p, r_0)] P[i \in L(p, r_0)] \\ &+ \Delta r E\left[-\frac{\Lambda_i}{2} + t_i \mid i \in R(p, r_0)\right] P[i \in R(p, r_0)]\end{aligned}$$

$$\begin{aligned}\Delta W \approx &- \Delta r \pi E[\Lambda_i \mid i \in L(p, r_0)] P[i \in L(p, r_0)] \\ &+ \Delta r \left\{-E\left[\frac{\Lambda_i}{2} \mid i \in R(p, r_0)\right] + E[t_i \mid i \in R(p, r_0)]\right\} P[i \in R(p, r_0)]\end{aligned}$$

For a reform that changes tax rates ceteris paribus (i.e. keeping $r$ and other components of prices fixed), we have direct and behavioral revenue effects:

$$\begin{aligned}\Delta G &\approx E[t\Delta x_i + x_i \Delta t] = tE[\Delta x_i] + E[x_i]\Delta t \\ &= tE\left[\frac{\partial x_i}{\partial p}\right]\Delta t + E[x_i]\Delta t \\ &= tE\left[\varepsilon \frac{x_i}{p+t}\right]\Delta t + E[x_i]\Delta t\end{aligned}$$

For a change in prices operating through a change in tax rates, the individual welfare effect $\Delta W^{ind}$ is given by equation (14) with $\Delta p = \Delta t$.

Putting these together:

$$\Delta W \approx \left( -(1-\pi) E \left[ \Lambda_i \frac{\partial x_i^L}{\partial p} \;\middle|\; i \in L \right] P[i \in L] - E[x_i(p_0, r)] \right) \Delta t + tE \left[ \frac{\partial x_i}{\partial p} \right] \Delta t + E[x_i] \Delta t,$$

Noting that the direct revenue effect and the direct individual welfare effect offset one another perfectly, this simplifies to

$$\Delta W \approx \left( -(1-\pi) E \left[ \Lambda_i \frac{\partial x_i^L}{\partial p} \;\middle|\; i \in L \right] P[i \in L] \right) \Delta t + tE \left[ \frac{\partial x_i}{\partial p} \right] \Delta t,$$

To characterize the new term further, note that obviously,

$$\frac{\partial x_i}{\partial p} = \begin{cases} \frac{\partial x_i^G}{\partial p}, & i \in G \\ \frac{\partial x_i^L}{\partial p}, & i \in L \\ 0, & i \in R. \end{cases}$$

We could also express these terms as elasticities, as in equation (15).

# D Relationship to Bernheim and Rangel (2009)

Bernheim and Rangel (2009) propose a general framework for decision-theoretic behavioral welfare economics. This appendix describes in detail the relationship between our analysis and this framework. We focus on mapping our Flexible Reduced-Form specification from equation (10) into the Bernheim-Rangel framework; a similar line of reasoning can be applied to other formulations.

The first step in applying this framework is to conceive of an observed choice in terms of a menu and an ancillary condition, or *frame* (denoted by $f$) – see also Bernheim and Taubinsky (2018). In describing this process, Bernheim and Taubinsky (2018) describe frames as those aspects of the choice situation that "have no direct bearing on well-being, but that instead impact biases."

What are the frames in our context? An initial guess might be that the reference point itself is a frame, but based on the definition above, this is not appropriate. We show that a change in the reference point can have a direct welfare effect, e.g. by modifying the incurred losses of individuals in the loss domain. Whether this direct effect should carry normative weight is a question of central importance, but this question belongs to a later step of the analysis, not the definition of a frame. Similarly, our theory implies that individuals should have a willingness to pay to change the reference point, suggesting that it has a direct bearing on well-being. Thus, we do not conceive of the reference point as a frame in the. A similar justification is used by Bernheim et al. (2015) in their application of this framework to the welfare economics of default options, to justify the treatment of the default as a component of the menu rather than a frame.

Nevertheless, there is a formal sense in which our results can be interpreted within the Bernheim-Rangel framework, which we now describe. First, we suppose that what we call observed demand in our analysis comes from choices under a single frame $f_1$. This frame is analogous to what Bernheim et al. (2015) refer to as a "naturally occurring frame." Under the frame $f_1$, the individual reveals preferences consistent with the utility function in equation (1), which we re-write here:

$$U(x, y, r, f_1) = u(x) + y + v(x, r), \tag{61}$$

where $v$ takes the form described in equation (10), or some other formulation from Appendix B.

In order to map our analysis into the Bernheim-Rangel framework, we need to consider a hypothetical choice situation in which reference dependence is eliminated in order to capture normative choices in the $\pi = 0$ case. If we wish to consider the possibility that reference dependence may be a bias, what preferences would be revealed by choices in an unbiased state? We represent choices made in a state without reference dependence via a frame $f_0$. Choices under $f_0$ maximize

$$U(x, y, r, f_0) = u(x) + y. \tag{62}$$

Choices under $f_0$ would of course be difficult to directly observe in empirical data, but the application of the Bernheim-Rangel framework does not require that all relevant parts of the choice correspondence are empirically observable. Choices under $f_0$ could potentially be observed by eliminating the effect of the reference point through some experimental intervention; we infer information about choices under $f_0$ implicitly using bunching methods in Proposition 3. Alternatively, under the restriction $\beta = 0$, $U$ under the two frames coincide in the gain domain, so we could identify $U$ under $f_0$ by observing choices under an extremely low reference point in $f_1$. Similarly under $\beta = 1$ we could identify $U$ under $f_0$ by observing demand in the loss domain.

Note that setting $f_1 = 1$ and $f_0 = 0$, we can represent choices in either frame $f \in \{0, 1\}$ by:

$$u(x, y, r, f) = u(x) + y + f \cdot v(x, r), \tag{63}$$

In this notation, the frame $f$ plays a similar role to $\pi$, but here we conceive of the two different frames purely in terms of choices in different situations, without a normative judgment.

The second step in applying the framework is to designate a subset of choice situations as the *welfare-relevant domain*, i.e. situations from which we wish to take normative inference. There are three intuitive possibilities for the welfare relevant domain, each of which reflects a normative judgment:

- (J1) include only choices under the naturally occurring frame ($f = 1$),
- (J2) include only choices under the no-reference-dependence frame ($f = 0$), or
- (J3) include choices under both frames.

The third step of the analysis is then to consider what revealed preferences are consistently expressed for choices within the welfare-relevant domain. If $a$ is chosen when $b$ is available for some situation in the welfare-relevant domain, and $b$ is never chosen when $a$ is available for other such situations, then we conclude that $a$ is preferred to $b$.

If we interpret our results within the Bernheim-Rangel framework, our goal and contribution is mainly to show how the these alternative judgments about the welfare-relevant domain influence welfare and optimal policy considerations. Under (J1) or (J2), there is a single utility function (either equation (61) or equation (62)) that ranks all options in the menu space (i.e. all combinations of $(x, y, r)$). Under (J3), however, we obtain only an incomplete ranking. Our results map into the Bernheim-Rangel framework as follows:

- (J1) Restricting the welfare-relevant domain to choices under $f = 1$ is equivalent to judging $\pi = 1$
- (J2) Restricting the welfare-relevant domain to choices under $f = 0$ is equivalent to judging $\pi = 0$.
- (J3) Including both $f = 0$ and $f = 1$ in the welfare relevant domain is equivalent to only taking welfare inference from welfare comparisons where some option $(x_0, y_0, r_0)$ is preferred to some other option $(x_1, y_1, r_1)$ for any $\pi \in \{0, 1\}$.

Proposition 1 shows that we can characterize the sign of the welfare effects of changes in $r$ without reference to $\pi$. This means that under the restriction on payoff formulations in Assumption 1 and 2, we always obtain robust over variation in $r$ that are independent of $\pi$. Through the lens of the Bernheim-Rangel framework, this suggests that even if we include choices under both $f_1$ and $f_0$ in the welfare relevant domain (J3) and use the revealed preference criterion proposed by Bernheim and Rangel, we would conclude that individuals prefer either higher or lower reference points according to the conditions laid out in Proposition 1.

**Alternative Approach Under $\pi = 0$.** Suppose, contrary to our preferred line of reasoning above, that we wish to conceive of the reference point as a frame. In this case, we could actually think of demand in the gain domain and demand in the loss domain as demand under two different frames. Note that under $\pi = 0$, with equation (10) we nest the case where demand in the gain domain is normative by $\beta = 0$, as in this case we have $v(p, r) = 0$ for $(p, r) \in G$. Similarly, demand in the loss domain is normative when $\beta = 1$. Thus, we could think of the parameter $\beta$ as capturing normative ambiguity over whether gain or loss domain demand are normative, provided we are willing to also assume that reference-dependent payoffs generally are not

normative ($\pi = 0$). Note also that this approach requires that we rule out diminishing sensitivity; otherwise, decisions under every possible reference point leads to distinct revealed preferences, so we would need a distinct frame for each. There are some similarities of this approach to the anchoring model of default effects in Bernheim et al. (2015).

# E Empirical Application

## E.1 Decomposing Reference Dependence Payoffs

Besides fiscal effects and effects on standard utility components, we calculate the effects of policies on reference dependence payoffs in the simulations. In the model from from equation (18), an individual's total reference dependence payoffs are given by

$$v(R|\hat{R}) = -\begin{cases} 0 & R < \hat{R} \\ \widetilde{\Lambda}(R-\hat{R}) & R \geq \hat{R}, \end{cases}$$

where $R$ is the individual's retirement age and $\hat{R}$ is the reference point given by the Normal Retirement Age. We further decompose reference dependence payoffs into additional disutility from work due to reference dependence and direct utility from the reference point. The first component, reference dependence disutility from work, is

$$v_b(R|\hat{R}) = -\begin{cases} \widetilde{\Lambda}\hat{R}_0 & R < \hat{R} \\ \widetilde{\Lambda}R & R \geq \hat{R}, \end{cases}$$

The second component, reference dependence utility from the reference point itself, is

$$v_d(R|\hat{R}) = -\begin{cases} \widetilde{\Lambda}(-\hat{R}_0) & R < \hat{R} \\ \widetilde{\Lambda}(-\hat{R}) & R \geq \hat{R}, \end{cases}$$

Note that we introduce a "base age" $\hat{R}_0$ given by the pre-reform NRA in the case $R < \hat{R}$. This choice is inconsequential for overall welfare effects, because $v_b + v_d = v$ for any base age. However, anchoring $v_b$ and $v_d$ at the initial reference point $\hat{R}_0$ allows to avoid introducing a jump discontinuity in $v_b$ and $v_b$ at $R = \hat{R}$, which would complicate the calculation of direct versus behavioral welfare effects for individuals moving between gain and loss domains relative to $\hat{R}$.

## E.2 Two-Dimensional Reference Dependence in the Empirical Application

### E.2.1 Two-Dimensional Model

In our empirical application, besides reference dependence over leisure, there could also be reference dependence in the consumption dimension. We can modify the preferences from equation (18) to include consumption reference dependence:

$$U = C - \frac{n}{1+\frac{1}{\epsilon}}\left(\frac{R}{n}\right)^{1+\frac{1}{\epsilon}} - \begin{cases} 0 & R < \hat{R} \\ \widetilde{\Lambda}_l(R-\hat{R}) & R \geq \hat{R}, \end{cases} - \begin{cases} \Lambda_c(\hat{C}-C) & C < \hat{C} \\ 0 & C \geq \hat{C}, \end{cases} \tag{64}$$

where $\hat{C} = C(\hat{R})$ is the consumption reference point, which is assumed to correspond to the consumption level at the NRA. Thus, the two-dimensional reference point lies on the budget constraint. The parameter $\Lambda_l$ captures the strength of reference dependence over leisure and $\Lambda_c$ captures the strength of reference dependence in the consumption dimension.[43] Such loss aversion in consumption may arise for instance because "full" pension benefits become available at the NRA, and individuals perceive the associated consumption

[43] $\Lambda_c$ implies additional marginal utility from consumption in the loss domain below $\hat{C}$. For instance, $\Lambda_c = 0.5$ corresponds to 50% higher marginal utility from consumption in the loss domain than in the gain domain.

level as a reference point (Behaghel and Blau 2012).[44]

As in the one-dimensional case, the two-dimensional model predicts bunching at the NRA. However, a crucial difference between the two models lies in the direction of predicted bunching. While reference dependence over leisure induces workers to retire earlier in order to enjoy more leisure, reference dependence over consumption induces individuals to postpone retirement and increase consumption. This occurs because the consumption loss domain is the range of consumption levels and associated retirement ages below the NRA, whereas the loss domain over leisure is above the NRA. Thus, reference dependence over leisure leads to *bunching from above*, but reference dependence over consumption leads to *bunching from below*. Figure 5 illustrates the predicted effect of the two dimensions of reference dependence on the retirement age distribution. Reference dependence over leisure implies a shift in the distribution toward the NRA from above, while reference dependence over consumption leads to a shift in the distribution toward the NRA from below. A combination of the two would imply a shift towards the reference points from both sides. As we argue in Section 4.6.1, the empirically observed retirement age distribution around the NRA suggests that reference dependence over leisure dominates reference dependence over consumption.

The marginal bunching individual from above can be characterized as in Section 4.2. The upper marginal buncher's indifference curve would be tangent to the budget line at some retirement age $R_+^*$ without reference dependence, and another indifference curve is tangent exactly at $\hat{R}$ with reference dependence. All workers initially located between $\hat{R}$ and $R_+^*$ bunch at the reference point from above, while all individuals initially to the right of $R_+^*$ decrease their retirement age but stay above the reference point. The two tangency conditions for the upper marginal buncher imply $R_+^* = n_+^*[w(1-\tau)]^\varepsilon$ and $\hat{R} = n_+^*[w(1-\tau-\Delta\tau-\Lambda_l)]^\varepsilon$, where $n_+^*$ denotes her ability level and $\Lambda_l = \widetilde{\Lambda}_l/w$ is the reference dependence parameter normalized by the wage per period. Hence,

$$\frac{R_+^*}{\hat{R}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau-\Lambda_l}\right)^\varepsilon$$

Similarly, a marginal bunching individual from below can be identified. The lower marginal buncher's indifference curve would be tangent to the budget line at $R_-^*$ without reference dependence, and tangency occurs exactly at $\hat{R}$ with reference dependence. All workers initially located between $R_-^*$ and $\hat{R}$ bunch at the reference point from below, while all individuals initially to the left $R_-^*$ retire later but stay below the reference point. The two tangency conditions of the lower marginal buncher are $R^* = n_-^*[w(1-\tau)]^\varepsilon$ and $\hat{R} = n_-^*[(1+\Lambda_c)w(1-\tau)]^\varepsilon$, where $n_-^*$ denotes her ability level. Hence,

$$\frac{R_-^*}{\hat{R}} = \left(\frac{1}{1+\Lambda_c}\right)^\varepsilon$$

The total excess mass $b = B/h_0(\hat{R})$ is

$$\frac{b}{\hat{R}} = \left[\left(\frac{1-\tau}{1-\tau-\Delta\tau-\Lambda_l}\right)^\varepsilon - 1\right] + \left[1 - \left(\frac{1}{1+\Lambda_c}\right)^\varepsilon\right] \tag{65}$$

Hence, bunching has two components. The first term in equation (65) captures bunching from the right (from above) due to the retirement age/leisure reference point in combination with a potential budget set kink present at the threshold. The second term in the equation captures bunching from the left (from below) due to the consumption reference point.

[44] Whether "full" pension benefits become available at the NRA depends on the specifics of the pension system. In the German setting, full benefits become available at the Full Retirement Age, which is in principle distinct from the NRA. However, for most workers among birth cohort 1946 on whom we focus in the simulations, the NRA and FRA coincide and thus full benefits become available at the NRA.

Equation (65) yields the exact amount of bunching under the utility funciton we assume. Taking a first-order Taylor approximation about the point $(\Lambda_l, \Lambda_c) = (0, 0)$ under $\Delta_t = 0$, we obtain the following approximation of the excess mass at a two-dimensional reference point without a local financial incentive kink:

$$\frac{b}{\hat{R}} \approx \varepsilon(\Lambda_l + \Lambda_c), \tag{66}$$

This expression is closely related to our first bunching identification result from Proposition 3.1. Observed bunching at the reference point identifies the combined strength of loss aversion over leisure and consumption, $(\Lambda_l + \Lambda_c)$, given an elasticity estimate. Separately identifying ($\Lambda_l$ and $\Lambda_c$) will require information about whether bunching comes from the left or from the right, as we show in general in Proposition 3.2 and specifically for the retirement model below.

### E.2.2 Parameter Estimation and Simulations

Analogously to equation (20), bunching observed at a threshold $j$, which may be the Normal Retirement Age or a pure financial incentive discontinuity, can be written as

$$\frac{b_j}{\hat{R}_j} = \left[\left(\frac{1-\tau_j}{1-\tau_j-\Delta\tau_j-\Lambda_l \cdot D_j}\right)^{\varepsilon} - 1\right] + \left[1 - \left(\frac{1}{1+\Lambda_c \cdot D_j}\right)^{\varepsilon}\right] + \xi_j \tag{67}$$

where $D_j$ is an indicator for the Normal Retirement Age and $\xi_j$ is an error term. As discussed above, a key issue with the estimation is that $\Lambda_l$ and $\Lambda_c$ cannot be separately identified based solely on equation (67). Intuitively, both retirement age and consumption reference points lead to sharp bunching at the threshold $\hat{R}$ such that a given amount of excess mass could be rationalized by a range of combinations of $\Lambda_l$ and $\Lambda_c$.

In order to make progress, it is useful to write the two components of excess mass separately. Bunching from the right is

$$\frac{b_j^+}{\hat{R}_j} = \left[\left(\frac{1-\tau_j}{1-\tau_j-\Delta\tau_j-\Lambda_l \cdot D_j}\right)^{\varepsilon} - 1\right] + \xi_j^+ \tag{68}$$

and bunching from the left is

$$\frac{b_j^-}{\hat{R}_j} = \left[1 - \left(\frac{1}{1+\Lambda_c \cdot D_j}\right)^{\varepsilon}\right] + \xi_j^- \tag{69}$$

where $b_j = b_j^+ + b_j^-$. Denoting $\beta_j = b_j^- / b_j$ the share of excess mass originating from the left, this share ranges between zero and maximum of $\hat{\beta}_j$. The maximum left bunching share $\hat{\beta}_j$ is given by one minus the fraction of bunching that would persist if workers only bunch due to the budget constraint kink.

We follow two approaches in order to obtain joint estimates of $\Lambda_l$ and $\Lambda_c$. First, we can simulate the full range of possible combinations of the two parameters by gradually moving the share of left bunching at the NRA from zero to its maximum and estimating equations (68) and (69) using the implied values of $b_j^+$ and $b_j^-$. Panel (a) of Appendix Figure A3 shows resulting parameter combinations. The negative slope of the relationship illustrates the intuition that the two types of reference dependence are substitutes in terms of rationalizing observed excess mass. The labeled dots in in the figure mark a range of implied left bunching shares between 0 and 50%. These results allow us to simulate the welfare effects of pension reforms as a function of the relative strength of consumption reference dependence, which are shown in Figure 6.

As a second approach, we aim at obtaining a set of preferred "point" estimates of $\Lambda_l$ and $\Lambda_c$. For this, an empirical estimate of $\beta$ is needed. We argue that the empirical retirement age distribution around the NRA is informative of the relative magnitude of bunching from the two sides, and can be used for this purpose under some additional assumptions. In particular, bunching shares from both sides can be computed based on estimates of the corresponding density shifts. Intuitively, we assume the counterfactual density to be

continuous around the NRA, and infer the relative number of bunchers from the left and from the right from the vertical difference between the counterfactual density and the actually observed density on both sides of the threshold. This estimation requires a stronger assumption about the true relative density shifts being reasonably well approximated by locally observed relative shifts.

We begin with the observation that bunching at the threshold must equal the total missing density from both sides:

$$B = \int_{R_{min}}^{\hat{R}} \left(h_0(R) - h(R)\right) dR + \int_{\hat{R}}^{R_{max}} \left(h_0(R) - h(R)\right) dR$$

where $R_{min}$ and $R_{max}$ are the minimum and maximum counterfactual retirement ages from which individuals bunch at the NRA.

Measuring the true density shift over the full support is impossible in practice for two reasons. First, the shift $h_0(R) - h(R)$ may vary across $R$ in an unknown way so that $h_0(R)$ cannot be measured for all $R$ based on the observed density. Second, the full support of the counterfactual density may not be observed. Even if the full support of the actual density could be observed, this does not necessarily correspond to the counterfactual support because some counterfactual density is predicted to "disappear" at the bounds because all individuals shift out a certain range.[45]

One solution to this problem is to approximate the true density shift by a constant shift over a certain range on each side. Denote by $h_+$ and $h_-$ the observed density immediately to the right and left, respectively, of the threshold $\hat{R}$. Furthermore, denote by $h_+^0$ and $h_-^0$ the corresponding counterfactual density in the absence of the threshold. The approximation is

$$B \approx \left(h_-^0 - h_-\right)\left(\hat{R} - R^-\right) + \left(h_+^0 - h_+\right)\left(R^+ - \hat{R}\right)$$

where a constant density shift observed immediately to the left of the threshold over a range $[R^-, \hat{R}]$ approximates for the true shift on the left and a constant shift observed immediately to the right of $\hat{R}$ over $[\hat{R}, R^+]$ approximates for the shift on the right.

Assume also that the counterfactual density is continuous at $\hat{R}$ such that $h_+^0 = h_-^0 = h_0$. Then $h_0$ can be recovered as

$$h_0 \approx \frac{B + (\hat{R} - R^-)h_- + (R^+ - \hat{R})h_+}{R^+ - R^-}$$

From this, the implied bunching shares from both sides can be computed as $B^- = (h_0 - h_-)(\hat{R} - R^-)$ and $B^+ = (h_0 - h_+)(R^+ - \hat{R})$ because bunching from either side must be equal to the total density shift on that side.

Panel (b) of Appendix Figure A3 illustrates this procedure. The solid red line shows the average empirical retirement density on both sides in a window of +/-2 years around the NRA, $h_+$ and $h_-$. The dashed red line shows the implied counterfactual density $h_0$ calculated as described above. The figure shows that the difference between the observed density and the counterfactual density is much larger on the right, indicating that most "missing density" is on this side, and thus most bunching appears to originate from above. We obtain an estimate of $\beta = 0.133$. Thus, the estimated share of bunching from the left due to reference dependence over consumption is 13.3% and the share of bunching from the right due to reference dependence over leisure is 86.7% . Finally, the parameters $\Lambda_c$ and $\Lambda_l$ can be estimated by plugging the bunching shares into equations (68) and (69). We obtain estimates of $\Lambda_c = 0.672$ and $\Lambda_l = 0.457$. The simulations shown in Table 3 are conducted based on these parameter estimates.

[45] Besides, although theory predicts individuals responding to the threshold along the entire density in principle, it is unclear in practice whether those far from the threshold respond in the same way as those closer.