CESifo WORKING PAPERS

9999
2022

October 2022

The Welfare Economics of Reference Dependence

Daniel Reck, Arthur Seibold

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9999*

The Welfare Economics of Reference Dependence

Abstract

Empirical evidence suggests that individuals often evaluate options relative to a reference point, especially seeking to avoid losses. In this paper, we provide the first welfare analysis under reference-dependent preferences. We decompose the welfare impact of changes in reference points and prices into direct and behavioral effects, and describe how these effects depend on whether reference dependence reflects a bias or a normative preference. In a simple model of loss aversion grounded in common empirical findings, we find that lowering reference points robustly improves welfare, while the welfare effect of a price change depends critically on normative judgments. We also derive sufficient statistics characterizations of the welfare effects of changing reference points and prices. We illustrate these theoretical findings with an empirical application to reference dependence exhibited in German workers' retirement decisions. Both simulation and sufficient statistics results suggest positive welfare effects of increasing the Normal Retirement Age, but ambiguous effects of financial incentives to postpone retirement. Finally, we study how adopting alternative models of reference dependent preferences modifies key welfare effects.

JEL-Codes: D910, D600, H550, J260.

Keywords: reference-dependent preferences, loss aversion, welfare, pension reform.

Daniel Reck
University of Maryland / College Park / USA
dreck@umd.edu

Arthur Seibold
University of Mannheim / Germany
seibold@uni-mannheim.de

September 2022

We thank Jack Fisher, Jacob Goldin, Xavier Jaravel, Camille Landais, Alex Rees-Jones, Emilie Sartre, 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.

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), and the timing of retirement (Seibold, 2021).

As policymakers take notice of mounting evidence of the fundamental 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? As in other behavioral settings, the influence of reference points on behavior creates ambiguity over welfare. One possibility is that individuals simply have strange normative preferences, so that reference dependence should be viewed as a part of revealed preferences. Alternatively, one might suppose that reference dependence distorts behavior relative to what is welfare-maximizing. Mainly because of this difficulty, the vast literature on reference dependence has entirely avoided welfare analysis.[2]

This paper undertakes the first study of the welfare economics of reference dependence. We tackle the central challenge by explicitly analyzing the normative ambiguity inherent in the theory. We view the resolution of the ambiguity over welfare as a normative judgment that must be made by a social planner, and we map this judgment to welfare quantities (Goldin and Reck, 2022). Specifically, we characterize the welfare impact of changes in the reference point, and of changes in prices (or taxes), under varying 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.

Besides normative ambiguity, a second challenge in analyzing welfare in the presence of reference dependence is the multitude of proposed models of behavior. Our approach is to start with the simplest model capable of delivering the key insights, and then develop extensions. Specifically, we begin with a simple model based on Tversky and Kahneman (1991), where the only deviation from standard preferences comes from loss aversion over the consumption of a single good, with an exogenous reference point.[3] Moreover, we consider these preferences in a static, non-stochastic setting. We argue that the simple model is a useful starting point for two reasons. First, this type of model is often used in applied work, as it is sufficient to rationalize empirically observed behavior in many contexts. Adopting it allows us to analyze welfare and

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

[2] In their review, O'Donoghue and Sprenger (2018) state that in existing literature "there is relatively little discussion of the welfare implications of reference-dependent preferences." In discussing why this is the case, they write, "When one takes a normative approach to reference-dependent preferences, a number of issues arise. Perhaps first and foremost is the question of whether gain-loss utility should be given normative weight – i.e., whether we should assume that the same preferences that rationalize behavior should also be used for welfare analysis."

[3] The determinants of reference points are the subject of much debate in the literature. We discuss how our approach relates to this debate in Section 2.3.

develop our intuition in a simple and tractable way. Second, our basic approach of tackling normative ambiguity can be used in any of the common models of reference dependence, but the more complicated models tend to entail more nuanced ambiguities and they are less straightforward to map to empirical evidence. Nevertheless, we relax each of the key simplifying assumptions behind our simple model in a number of extensions later on.

We begin by analyzing the welfare effect of a policy that influences 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 show that a change in the reference point has two potential first-order welfare effects; which of these matters depends on the planner's normative judgment. If the planner judges that reference dependence is normative, the only first-order welfare effect is the *direct effect* of a change in the reference point on gain-loss utility, holding behavior fixed. In this case, any change in behavior has no first-order welfare effect due to the envelope theorem. In contrast, if the planner judges that reference dependence is a bias, then the sole first-order welfare effect comes from the *behavioral effect* of a change in the reference point, and the direct effect is immaterial.

Characterizing these direct and behavioral effects in the simple model, we find that decreasing the reference point is a *robust Pareto improvement*: a lower reference point improves welfare regardless of the normative view one takes.[4] When reference dependence is normative, the direct effect implies that loss-averse individuals are better off when the reference point decreases because they incur smaller utility losses. When reference dependence is a bias, individuals tend to over-consume the good in order to reduce their losses - i.e. there is a *negative internality* (Mullainathan et al., 2012). Decreasing the reference point mitigates this over-consumption.

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 first-order direct welfare effect. And as before, when reference dependence is judged to be normative, the change in behavior has no first-order welfare implications. When loss aversion causes over-consumption of the good, however, the decrease in consumption caused by a price increase has a positive first-order behavioral welfare effect. Furthermore, this logic implies that when reference dependence is a bias, there is scope for corrective taxation to address this bias, and we derive the optimal corrective tax schedule in the simple model.

We develop simple sufficient statistics formulas approximating the welfare effects of policies that influence reference points or change prices. To quantify the welfare effect of a small change in the reference point, the sufficient statistics are the preference parameter governing the strength of loss aversion and the fractions of individuals whose outcomes are located above or below the reference point. For the welfare effects of price changes, the price elasticity of demand and the average consumption level are additionally required. Our sufficient statistics formulas will be straightforward to implement in many applied contexts: the loss aversion parameter and demand elasticities are commonly estimated in the literature, and the remaining objects can be easily calculated given data on individual outcomes.

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

[4]The intuition that comparing one's outcomes to a low reference point is desirable if reference dependence enters welfare is likely known among many behavioral economists, but to our knowledge this has never been formalized. We formalize this result and, crucially, we also demonstrate that it is robust to the normative view of reference dependence because both direct and behavioral welfare effects have the same sign.

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 reference points for retirement decisions. On the other hand, 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 (1) individual-level simulations of a model of retirement behavior and (2) 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. We find that in addition to the positive fiscal effects it entails, such a reform robustly increases 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 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 depend on strongly on normative judgments. 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. In fact, we can calculate optimal corrective subsidies for later retirement in this case. Due to strong estimated reference dependence in our setting, such a subsidy would have to be very large. 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.

In order to understand how alternative approaches to modeling reference-dependent preferences modify welfare effects, we consider a number of extensions, in turn relaxing the main simplifying assumptions behind our baseline model. First, we allow reference dependence to be present in more than one dimension. For example, one could suppose that there is reference dependence over labor supply/leisure or over consumption or over both in the retirement setting (Crawford and Meng, 2011; Behaghel and Blau, 2012). Building on the logic of direct and behavioral effects from the simple model, we describe welfare in a two-dimensional model of reference dependence. As before, normative judgments shape which effects matter for welfare. However, the sign and magnitude of some welfare effects can theoretically differ from the simple model, depending on the relative strength of reference dependence in the two dimensions. We further illustrate the implications of two-dimensional reference dependence in our empirical application. We argue that the shape of the empirical retirement age distribution around the NRA aligns well with reference dependence over labor supply/leisure and less well with reference dependence over consumption. We then

show that all the main welfare effects of pension reforms qualitatively persist under our preferred estimate of the strength of consumption reference dependence. However, if consumption reference dependence is very strong (and judged to be normative), the welfare effects of increasing the NRA can turn negative, because workers retiring early experience additional loss disutility relative to a higher consumption reference point.

The second extension considers an implication of some formulations of reference dependence (e.g. Tversky and Kahneman, 1991; Kőszegi and Rabin, 2006), namely that not only decisions over losses but also decisions over gains may be affected. Although reference dependence is often modeled in this way in the literature, including gain utility is not necessary to explain the empirical patterns typically attributed to reference dependence.[5] Moreover, we find that the signs of the main welfare effects do not change when introducing gain utility, and if anything the magnitude of effects would become larger. Our third extension considers the deliberate setting of reference points by individuals out of another concern, like a separate behavioral bias they wish to overcome. We study this in a model of goal-setting like Koch and Nafziger (2011) and examine how policy implications depend on the extent to which individuals are able to optimally set their own reference points. Finally, we discuss a number of considerations for future work, including "diminishing sensitivity", reference dependence under risk and uncertainty, and the question of choice bracketing.

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, including Chetty et al. (2009), Mullainathan et al. (2012), Allcott and Taubinsky (2015) and Allcott et al. (2019). 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 the main obstacle to welfare analysis. Our notion of normative ambiguity builds on foundational work on ambiguously revealed preferences by Bernheim (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 findings can be used to derive novel policy implications for the wide range of contexts where reference-dependent preferences have been shown to play a role.

Our empirical application also 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). Applying our theoretical findings to the retirement context complements recent approaches to the welfare effects of pension reforms (Kolsrud et al., 2021; Haller, 2022). In particular, we consider

[5] We formalize and prove this claim in Appendix D. See also Barseghyan et al. (2013).

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. In Section 2, we introduce the basic model and derive our main theoretical results, Section 3 presents the empirical application to retirement behavior, Section 4 discusses extensions, and Section 5 concludes.

2 A Simple Model of Welfare Under Reference Dependence

In this section, we lay out a simple model of reference dependence and characterize behavior and welfare within this model. We begin by describing the empirical patterns whose welfare implications we aim to understand. We then develop the simplest possible model that can rationalize these empirical observations: a model of loss aversion over the consumption of a single good. This model embeds some simplifying assumptions relative to the broader family of models of reference dependence. We relax these assumptions when we enrich the theory in Section 4.

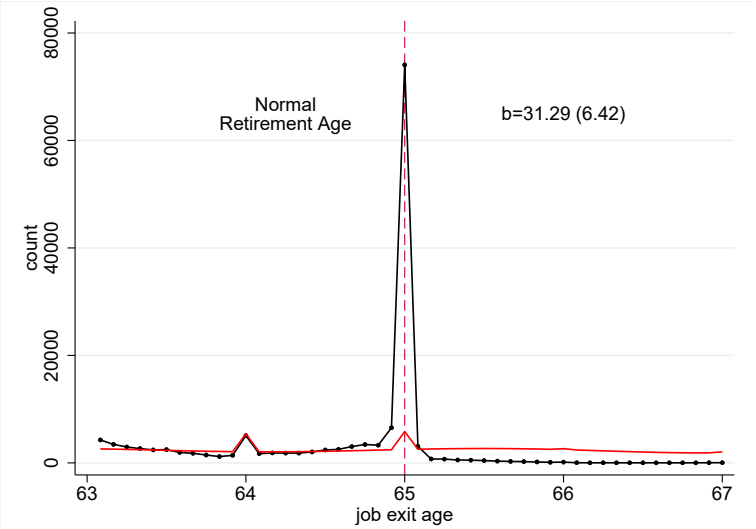
2.1 Empirical Motivation

A number of recent empirical studies document that individual behavior matches the predictions of models of reference dependence and, in particular, loss aversion. One key example is that a model along these lines can explain why many individuals retire precisely at *statutory retirement ages* even in the absence of financial incentives (Seibold, 2021). To illustrate this, Figure 1 plots the distribution of retirement ages among German workers around the Normal Retirement Age (NRA). The retirement density exhibits two striking features. First, there is sharp bunching precisely in the month of the NRA, even though individuals do not face any financial incentive to retire at this particular age. Second, there is a visible drop in the density above the NRA relative to below the NRA, suggesting that bunching is driven by individuals moving retirement forward toward the NRA. Section 3 provides a detailed description of the retirement setting and bunching estimation at the NRA, and Section 4.1 discusses the density shift in more detail.

From a revealed preference perspective, these behavioral patterns suggest that individuals have a significant willingness to pay to retire at or before the NRA. This corresponds precisely to the predictions of models of reference dependence with loss aversion. Figure 1 can be rationalized with a model in which the NRA serves as a reference point for retirement behavior, relative to which individuals experience loss aversion over lifetime leisure. Moreover, Seibold (2021) shows that pension reforms shifting statutory retirement ages like the NRA have large effects on retirement behavior, suggesting that retirement reference points can be influenced by policy.

Similar patterns can be found in other contexts, including behavioral responses to taxation. For instance, Rees-Jones (2018) finds that the distribution of tax liabilities at the time of filing a tax return in the U.S. exhibits sharp bunching at zero and a downward shift in the region of above zero. This suggests that income tax withholding creates an arbitrary reference point at zero tax due, with associated loss aversion over positive tax liabilities at the time of filing. Another example is given by Homonoff (2018), who finds that implementing a plastic bag tax as a tax (penalty) rather than a subsidy (bonus) for re-usable bag use has a large positive effect on the use of re-usable grocery bags. Again, this can be rationalized by reference dependence with loss aversion, where consumers perceive the subsidy as a gain and the tax as a loss. Besides speaking in favor of reference-dependent preferences, these empirical studies have another common feature:

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 is predicted (see Section 3). The parameter b denotes the excess mass at the NRA with its standard error shown in parantheses.

Government policy seems to be able to influence reference points by creating salient benchmarks or by framing incentives. This apparent malleability of reference points is a key motivation for us to study the welfare effects of policies that affect reference points.

2.2 Setup

Motivated by this empirical evidence, we develop a model where loss aversion over the consumption of a single good is the sole deviation from standard behavior. This is the simplest model sufficient to rationalize the empirical patterns commonly attributed to reference dependence. In Section 4, we enrich the theory in a number of ways building on the broader literature on reference dependence.

Behavior. A population of individuals of measure one, indexed by i, chooses a good $x \in \mathbb{R}$ and a background good $y \in \mathbb{R}$ subject to a standard linear budget constraint with income z_i. The exogenous price of x is p, and the price of y is normalized to 1. Each individual chooses according to a utility function $U(x, y)$. We assume $U(x, y)$ consists of quasi-linear utility over x and y plus a reference-dependent payoff from consuming x, with a reference point $r \in \mathbb{R}$.

$$\begin{aligned} \max_{x,y} U(x, y) &= u_i(x) + y + v_i(x|r) \\ \text{subject to } & px + y = z_i. \end{aligned} \tag{1}$$

In the language of Kahneman et al. (1997), $U(x, y)$ can be referred to as *decision utility* as it generates behavior, which may be distinct from *experienced utility* or welfare. We assume an interior solution, and that $u_i' > 0$

and $u_i'' < 0$ always.

Empirical evidence from various contexts, including the examples cited above, suggests that individuals behave as if there is a discontinuous change in marginal utility over good x at some reference point r, whereby losses incur a penalty relative to gains.[6] In our simple model, the only deviation from classical preferences we adopt is the modification to decision utility necessary to rationalize this empirical observation. We specify the deviation from the classical model in the $v_i(x|r)$ term as follows:

$$v_i(x|r) = \begin{cases} 0 & x > r \\ \Lambda_i(x-r) & x \leq r. \end{cases} \tag{2}$$

The parameter $\Lambda_i > 0$ governs the extent of *loss aversion*, the size of the penalty that marginal losses incur relative to marginal gains. The domain where $x > r$ is called the *gain domain*, and the domain where $x < r$ is called the *loss domain*.

Simplifying Assumptions. The model described by equations (1) and (2) embeds three key simplifying assumptions. We relax all three in Section 4.

First, we assume that the individual receives no reference-dependent payoff in the gain domain. The model of Tversky and Kahneman (1991) contains gain domain payoffs, but in practice it is difficult to distinguish between these payoffs and utility over consumption of good x.[7] In fact, assuming away gain domain payoffs is immaterial for explaining behavior and, as we show in Section 4.2, incorporating these payoffs does not qualitatively change any of our welfare results. In revealed preference terms, a key implication of this first simplifying assumption is that choices in the gain domain are deemed "welfare relevant" (Bernheim and Rangel, 2009), while choices in the loss domain are more suspect.[8] Note that by ruling out gain domain payoffs, we also rule out another alternative to the specification in equation (2), which would be to keep loss-domain payoffs at zero in $v(.)$ and instead include a term in the gain domain with a negative payoff proportional to $(x - r)$. We argue that this specification is appropriate because a sizable literature in psychology and neuroeconomics suggests that loss aversion is driven by a negative affective response to the incursion of perceived losses (see the discussion of mechanisms below).

Our second simplifying assumption is that reference dependence affects payoffs over a single dimension, ruling out additional reference-dependent payoffs over good y. In Section 4.1, we relax this assumption and analyze how multi-dimensional reference dependence would alter our results both in theory and in the empirical application. Third, we assume that the reference point relative to which options are evaluated is exogenous, ruling out that individuals exert influence over it. In Section 4.3, we consider an extension of our model where individuals choose their own reference points.

[6] A possible alternative formulation of reference dependence could be a utility notch (a discontinuity in the level of utility) rather than a utility kink (a discontinuity in marginal utility) (see e.g. Allen et al., 2017). We focus on the kink formulation for two reasons. First, it is by far the most commonly adopted functional form in the literature. Second, the kink formulation 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)).

[7] In Tversky and Kahneman (1991), gain-loss utility is posited as the sole component of preferences. Later applications incorporate non-reference dependent concerns as we do in the first part of equation (1) (see e.g. Kőszegi and Rabin (2006); O'Donoghue and Sprenger (2018)). We also disregard *diminishing sensitivity*, which would require that $v_i'' > 0$ in the loss domain and $v_i'' < 0$ in the gain domain. We discuss this further in Section 4.4.

[8] See Appendix E for further discussion of the relationship between our work and the general revealed preference framework of Bernheim and Rangel (2009).

Demand. We now describe demand $x_i(p, r)$ in the simple model. Following our assumption of quasi-linear preferences, we suppress z_i as an input into demand and other functions. We first characterize potentially optimal choices in the gain domain (x_i^G) and in the loss domain (x_i^L) as follows:

$$u'(x_i^G(p)) = p, \tag{3}$$

$$u'(x_i^L(p)) + \Lambda_i = p. \tag{4}$$

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 for a given individual is

$$x_i(p, r) = \begin{cases} x_i^G(p), \text{ if } x_i^G(p) > r & (G) \\ x_i^L(p), \text{ if } x_i^L(p) < r & (L) \\ r, \qquad \text{otherwise.} & (R) \end{cases} \tag{5}$$

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):

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

We usually suppress inputs into group labels to economize on notation.

Individual Welfare. The planner must judge whether reference-dependent decision utility should be given normative weight, i.e. whether to respect loss aversion or regard it as a bias. We parametrize this decision by $\pi \in \{0, 1\}$, where $\pi = 1$ if the planner respects loss aversion.[9] We express normative preferences as

$$U_i^*(x, y) = u_i(x) + y + \pi v_i(x|r), \tag{6}$$

We denote indirect utility, or welfare at a given price, income and reference point, by

$$w_i(p, r) \equiv U_i^*(x_i(p, r), z_i - p x_i(p, r)). \tag{7}$$

Given the judgment encoded by π, U_i^* is a money-metric measure of welfare.

Mechanisms and the Correct Value of π. We interpret the choice of π as a normative judgment the social planner must make. The psychological mechanisms behind loss aversion have some bearing on this, but understanding the psychological mechanism will arguably not resolve the need for a normative judgment. For example, one strain of the psychological literature suggests that loss aversion has emotional origins (see Rick (2011) for a review). The findings of an influential study by Kermer et al. (2006) suggest that loss aversion derives from an *affective forecasting error*: people wrongly project that they will experience emotional pain if they incur a loss, so they try to avoid losses. In this case, it may be appropriate to set $\pi = 0$. However, more recent evidence indicates that the emotional pain of incurring losses is real rather than a forecasting error and that emotional regulation strategies mitigate loss aversion (Sokol-Hessner et al., 2009). This idea is

[9] We focus on the cases where $\pi = 0$ and $\pi = 1$ for clarity, but our analytic expressions could also be evaluated for values of $\pi \in (0, 1)$.

further borne out by neurological evidence associating activity in the amygdala with loss aversion and the incursion of perceived losses in a number of ways (De Martino et al., 2010; Sokol-Hessner et al., 2013; Sokol-Hessner and Rutledge, 2019). Under this premise, the choice of π becomes a deeper question about whether individuals should let negative emotions like fear or regret influence their choices, or whether individuals should make decisions dispassionately (see e.g. Loewenstein and O'Donoghue, 2006). Thus, understanding the mechanisms at play is interesting, but it does not entirely resolve normative ambiguity. Henceforth, we remain agnostic about the value of π.

Social Welfare. Some of our results can be derived from individual welfare alone, but other policy changes will create winners and losers. We require a notion of social welfare to evaluate such policies. We adopt a simple utilitarian social welfare function:

$$W(p,r) = \int_i w_i(p,r)di. \tag{8}$$

Due to our assumption of quasi-linear preferences, maximizing this social welfare function is equivalent to maximizing the sum of compensating or equivalent variation, relative to any arbitrary benchmark. One could relax the assumption of utilitarian social preferences and quasi-linearity, for instance with an application of Saez and Stantcheva (2016). However, addressing distributional concerns related to reference dependence is beyond the scope of this paper, and we defer this to future work.

2.3 Results

Next, we lay out the main theoretical results arising from the simple model. We begin with an intuition-building characterization of welfare, and then derive the welfare effects of references points and prices.

The Marginal Internality. A key statistic for behavioral welfare analysis is the *marginal internality* (Mullainathan et al., 2012; Allcott and Taubinsky, 2015; Allcott et al., 2019). In our context, this is the (money metric) welfare effect of a marginal change in x along the budget constraint, evaluated at observed demand. Using the first-order conditions in equations (3) and (4) and the behavioral characterization in (5), it is straightforward to derive the following:

Lemma 1. ***The Marginal Internality.*** *Define* $m_i(p,r;\pi) \equiv \frac{dU_i^*(x,z_i-px)}{dx}\Big|_{x=x_i(p,r)}$.

L1.1. *If* $x_i(p,r) > r$, $m_i(p,r;pi) = 0$.

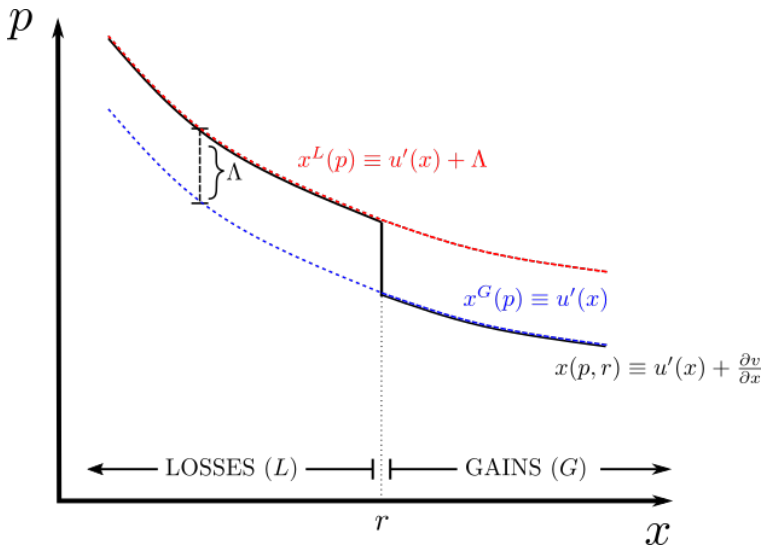
L1.2. *If* $x_i(p,r) < r$, $m_i(p,r;pi) = -(1-\pi)\Lambda_i$

L1.3. *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* $-\Lambda_i \leq m_i \leq 0$

We interpret the marginal internality as the welfare effect of paternalistically inducing the consumer to choose a little bit more of good x, starting from observed demand. 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 induced change in behavior would improve welfare. When $\pi = 0$, in

FIGURE 2: OBSERVED DEMAND, WELFARE-MAXIMIZING DEMAND, AND MARGINAL INTERNALITIES



Notes: The figure depicts observed demand $x(p, r)$ at a given reference point r in the black line. We also plot demand in the gain and loss domains, $x^G(p)$ (in blue) and $x^L(p)$ (in red). The vertical distance between $x^G(p)$ and $x^L(p)$ in the loss domain equals the loss aversion parameter Λ. When $\pi = 1$, observed demand is welfare maximizing. When $\pi = 0$, $x^G(p)$ is welfare maximizing, and by Lemma 1, the marginal internality is $-\Lambda$ in the loss domain.

contrast, individuals with $x_i \leq r$ are over-consuming good x out of loss aversion, so the marginal internality is negative.

Building on Lemma 1 and our characterization of behavior, Figure 2 plots individual demand for good x. We also plot demand according to $x^G(p)$ and $x^L(p)$, as defined in equations (3) and (4), for illustration. Individual demand coincides with welfare-maximizing demand when $\pi = 1$. Under $\pi = 0$, on the other hand, the marginal internality drives a wedge between observed demand and welfare-maximizing demand in the loss domain. The size of the marginal internality in the loss domain is $-\lambda$, corresponding to the vertical difference between the two demand curves in the figure.

2.3.1 Welfare Effects of Changing Reference Points

Can Policy Change Reference Points? The origin of reference points is the subject of much discussion in the literature. In the design of their early experiments, Kahneman and Tversky apparently viewed reference points as exogenous features of the environment which can be manipulated.[10] The evidence from field settings we discussed in Section 2.1 is consistent with this view, as policy design appears to exert strong influence over individual reference points in Homonoff (2018), Rees-Jones (2018) and Seibold (2021). Similarly, the psychological literature suggests that reference points can be influenced by external factors, including salient options (Rosch, 1975) and externally set goals (Heath et al., 1999).

Given the evidence that policy can influence reference points in important real-world contexts, we argue that characterizing the welfare effects of changing reference points is valuable regardless of the exact origins of individual reference points. For any policy P that affects a reference point $r(P)$, we can write the welfare

[10]For example, in the famous "Asian disease experiment" from Tversky and Kahneman (1981), in one experimental condition the potential options are described in terms of gains (lives saved) and in the other condition the same options are described in terms losses (lives lost). The predominant expressed preference – a greater appetite for risk under the loss framing – implies that the change in framing shifts the reference point against which the options are compared.

effect of this policy as $\frac{dW}{dP} = \frac{\partial W}{\partial P} + \frac{\partial W}{\partial r}\frac{\partial r}{\partial P}$. The first term captures any direct effect of the policy on welfare, while the second term captures the welfare effect that the policy has because it changes the reference point. Our analysis can be thought of as characterizing the $\frac{\partial W}{\partial r}$ term, abstracting from other welfare effects such policy changes might have. For simplicity, we consider policies for which $\frac{\partial r}{\partial P} = 1$, i.e. policies that directly shift reference points. While the evidence cited above indicates that policy can affect reference points, less is known about by "how much" reference points can be changed. For instance, one may conjecture that there are limits to more extreme attempts to influence behavior via reference points. In principle, our analysis could easily accommodate other values of $\frac{\partial r}{\partial P}$.

One strand of the literature considers endogenous reference points in terms of beliefs or expectations, typically focusing on the case with uncertainty (e.g. Kőszegi and Rabin, 2006, 2007).[11] There is experimental evidence that changing expectations influences reference points (Abeler et al., 2011; Ericson and Fuster, 2011), while other experiments suggest that simple models of expectations-based reference points have limited explanatory power (Gneezy et al., 2017; Goette et al., 2021). Relatedly, DellaVigna et al. (2017) and Thakral and Tô (2021) find that past experiences can influence reference points. A policy that changes expectations or another determinant of the reference point might have effects on welfare that are not driven by reference dependence itself, as in the $\frac{\partial W}{\partial P}$ term above, but in order to characterize the full welfare effect we need to consider the effect on welfare through the reference dependence channel. While we do not explicitly consider models of expectations- or experience-based reference points, our analysis thus relates to these models by characterizing $\frac{\partial W}{\partial r}$.

Changes in the Reference Point. Let r_1 and r_0 denote two arbitrary reference points, where $r_1 > r_0$. Based on (5), we know that there are three cases to consider under a given reference point. It is straightforward to show that $x(p, r_1) \geq x(p, r_0)$, so we have six potential cases for behavior under these two reference points. Define GR as the group of individuals located in the gain domain under the old reference point and at the reference point under the new reference point, $GR = \{i | x_i(p, r_0) = x_i^G(p) \ \& \ x_i(p, r_1) = r_1\}$, and define the other groups analogously. The change in individual welfare induced by a change in the reference point is:

$$w_i(p, r_1) - w_i(p, r_0) = \begin{cases} 0, \ i \in GG \\ u_i(r_1) - u_i(x_i^G) - p(r_1 - x_i^G), \ i \in GR \\ u_i(x_i^L) - u_i(x_i^G) - p(x_i^L - x_i^G) + \pi\Lambda_i(x^L - r_1), \ i \in GL \\ u_i(r_1) - u_i(r_0) - p(r_1 - r_0), \ i \in RR \\ u_i(x^L) - u_i(r_0) - p(x^L - r_0) + \pi\Lambda(x_i^L - r_0), \ i \in RL \\ -\pi\Lambda_i(r_1 - r_0), \ i \in LL. \end{cases} \tag{9}$$

Marginal Changes and Sufficient Statistics. With a parameterized and estimated structural model, one could directly use equation (9) to calculate the social welfare impact of changes in the reference point, taking the sum of the effects across individuals in all six groups. Considering a marginal change in the reference point yields a simpler *sufficient statistics* characterization of the first-order components of this welfare effect.

Let $\Delta r = r_1 - r_0$, and let $\Delta x_i = x_i(p, r_1) - x_i(p, r_0)$. Differentiating indirect utility $w_i(p, r)$ from equation (7) with respect to r and applying the definition of $m_i(p, r, \pi)$, we find that where $w_i(p, r)$ is differentiable, the marginal effect of a change in the reference point is approximately

[11] Salient options vs. expectations are often thought of as two alternative ways to model the origin of reference points, but these possibilities are not entirely mutually exclusive. For instance, points made salient by policy could influence expectations when individuals have weak priors.

$$\Delta w_i \approx \underbrace{\pi \left.\frac{\partial v_i}{\partial r}\right|_{x=x_i(p,r)}}_{\text{Direct Effect}} \Delta r + \underbrace{m_i(\pi)\Delta x}_{\text{Behavioral Effect}} . \tag{10}$$

Equation (10) shows that a change in the reference point has two potential first-order effects on welfare. Which of these two effects matters depends on the normative judgment π. First, holding behavior fixed, there is a direct effect because $v_i(x|r)$ is decreasing in r. However, this direct effect only materializes when the planner judges that reference-dependent payoffs carry normative weight, i.e. $\pi = 1$. Second, there is a behavioral effect, whereby changes in the reference point affect consumption of good x, which can also affect welfare. From Lemma 1, we know that internalities are only present when $\pi = 0$ and thus a first-order behavioral welfare effect only occurs in this case. On the other hand, the envelope theorem implies the behavioral effect is second-order when $\pi = 1$.

We note that the characterization of welfare in equation (10) is general. The same basic characterization of first-order welfare effects obtains for virtually any formulation of reference-dependent payoffs. So we can build intuition about how positive assumptions about reference-dependent payoffs shape welfare by considering the differences in m_i, $\frac{\partial v_i}{\partial r}$, and Δx_i implied by different formulations. We follow this approach with respect to each of our three simplifying assumptions in Section 4.

Now we implement equation (10) in the simple model. Increasing r increases consumption of x, so $\Delta x > 0$. Lemma 1 shows that marginal internalities are negative where they are present. This means that behavioral welfare effects of increasing the reference point will be negative; individuals consume too much of good x out of loss aversion. Meanwhile, reference-dependent payoffs are decreasing in r, $\frac{\partial v}{\partial r} < 0$, so the direct welfare effect will also be negative. Intuitively, individuals feel worse when comparing their outcomes to a higher reference point. Because both effects are negative, the sign of the overall welfare effect of a change in r is unambiguously negative: increasing the reference point decreases welfare. In fact, decreasing the reference point is a Pareto improvement. We refer to this result as *robust* to the planner's normative judgment of reference dependence, as it arises regardless of the chosen value of π.[12]

Proposition 1. ***First-Order Welfare Effects of a Change in the Reference Point.***

P1.1. ***Individual Welfare Effects.*** *Consider a change in the reference point small enough that the GL group is negligible. The effect of this change in the reference point on individual welfare is approximately*

$$\Delta w_i \approx \begin{cases} 0, \ i \in GG, GR \\ -(p - u_i'(r_0))\Delta r, \ i \in RR \\ -(1-\pi)\Lambda_i \Delta x_i - \pi \Lambda_i \Delta r, \ i \in RL, \\ -\pi \Lambda_i \Delta r, \ i \in LL. \end{cases} \tag{11}$$

P1.2. ***Sufficient Statistics Formula for Social Welfare.*** *If the distribution of $u_i'(r_0)$ is approximately uniform over $[r_0, r_1]$,*[13] *then the effect of a small change in the reference point of Δr on social welfare is approximately*

$$\begin{aligned} \Delta W \approx & -\Delta r \pi E[\Lambda_i \mid i \in L(p, r_0)] P[i \in L(p, r_0)] \\ & - \Delta r E\left[\frac{\Lambda_i}{2} \mid i \in R(p, r_0)\right] P[i \in R(p, r_0)]. \end{aligned} \tag{12}$$

[12] Equivalently, in the simple model, all individuals prefer lower reference points according to the robust revealed preference criterion of Bernheim (2009). See Appendix E.

[13] Formally, the required assumption is that for any value Λ, the distribution of $u_i'(r_0)$ conditional on $\Lambda_i = \Lambda$ is approximately uniform.

Figure 3 illustrates the welfare effect of a change in r for the cases in Proposition 1.1. In each case, the individual welfare effect can be approximated as the sum of a behavioral and a direct effect as in equation (10); some of these effects are zero in some cases. For instance, an inframarginal individual in the loss domain (LL) experiences no behavioral effect and a direct welfare loss – increasing r makes losses relative to the reference point larger. In the RR case, the potential behavioral and direct effects are approximately equal to one another, but which of these effects actually causes a welfare loss depends on π. Intuitively, in the $\pi = 1$ case, the individual is at a corner solution and increasing r exacerbates their loss from being stuck on a corner, which is a direct welfare effect. In the $\pi = 0$ case, the individual is over-consuming good x out of loss aversion, and increasing the reference point worsens that over-consumption: a behavioral welfare effect. In the marginal loss case (RL) we observe both direct and behavioral effects: the increase in the reference point causes the individual to start incurring losses when $\pi = 1$, while it worsens their over-consumption of good x when $\pi = 0$. There are no direct or behavioral effects in the inframarginal gain case (GG), while in the marginal gain case (GR) there are no first-order effects but only a second order loss due to a behavioral effect.

We turn to social welfare in Proposition 1.2. Social welfare is everywhere continuous, so the marginal gain and marginal loss cases are of second-order importance for social welfare – these are marginal welfare effects for marginal groups. The first term of equation (12) represents the direct effect on the L group, which only materializes if $\pi = 1$. The second term represents the robust welfare effect on the R group. Thus, the overall social welfare loss from an increase in r is larger if $\pi = 1$, because in this case the L group's utility losses enter social welfare.

Proposition 1.2 simplifies the welfare effect further by using a trapezoidal approximation of the underlying integral for the effect on the R group. For a given value of loss aversion Λ_i, losses in the R group range from 0 to Λ_i, so the total loss averages out to $\Lambda_i/2$ under an approximately uniformly distributed willingness to pay for x at the reference point. Equation (12) thereby provides an easily implementable sufficient statistics characterization of the social welfare effect of changing r. Besides the normative judgment π, the welfare effect of changing the reference point depends only on the mean strength of loss aversion (Λ_i) in the R and L groups, and the size of these two groups. Because the loss aversion parameter Λ is a key estimation target in empirical studies of reference dependence and the relative size of R and L groups can be measured directly given data on individual outcomes, this sufficient statistics formula will be straightforward to apply in many contexts.

In summary, the social planner should prefer to lower reference points, all else equal. We also derived a simple sufficient statistics characterization, allowing for a straightforward quantitative approximation of the welfare effects of changing reference points.

2.3.2 Welfare Effects of Changing Prices

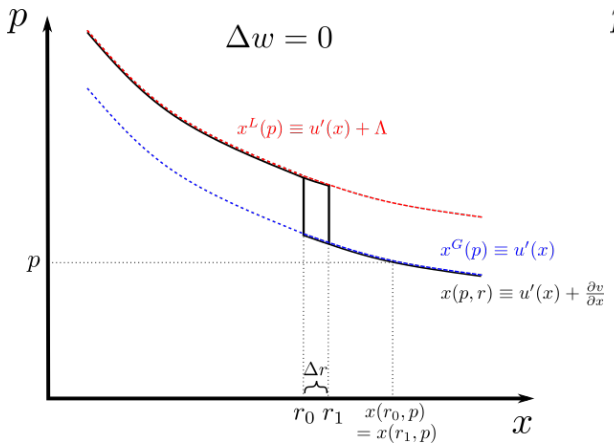
Price changes also have direct and behavioral effects. The analogue to equation (10) for the case of a price change is:

$$\Delta w_i \approx \underbrace{x\Delta p}_{\text{Direct Effect}} + \underbrace{m_i(\pi)\Delta x}_{\text{Behavioral Effect}}, \tag{13}$$

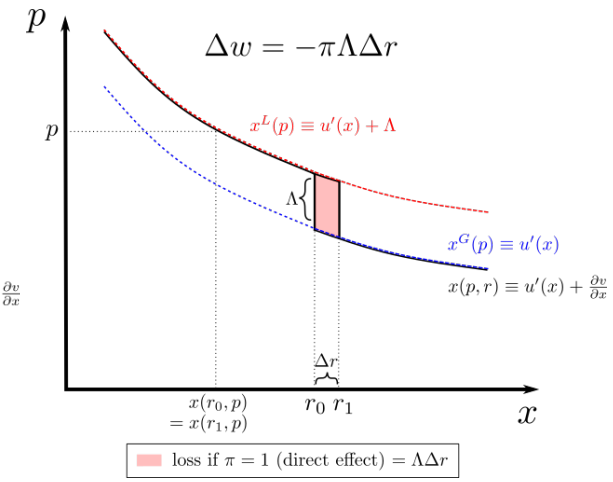
where $\Delta p = p_1 - p_0$, $\Delta x_i = x_i(p_1, r) - x_i(p_0, r)$, and Δw_i is defined similarly. The first term is the standard welfare loss from a price change, which can be offset by a change in firm revenues (or government revenues for a tax change). The second term reflects a first-order welfare effect when there are internalities, i.e. when $\pi = 0$. In this case, individuals over-consume good x out of loss aversion, $m_i < 0$ and $\Delta x < 0$, so the be-

FIGURE 3: WELFARE EFFECTS OF CHANGING THE REFERENCE POINT

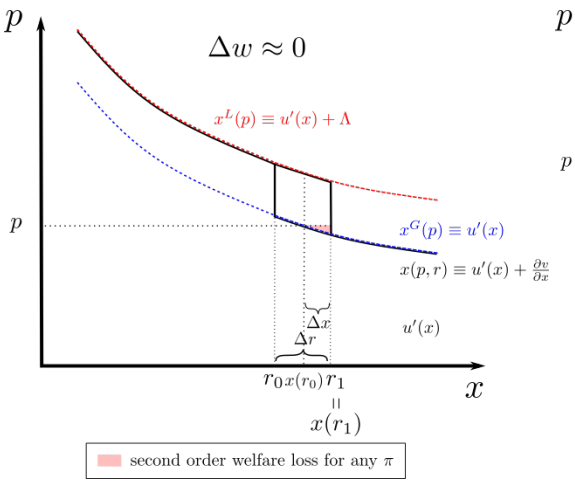
(a) Gain domain (GG)



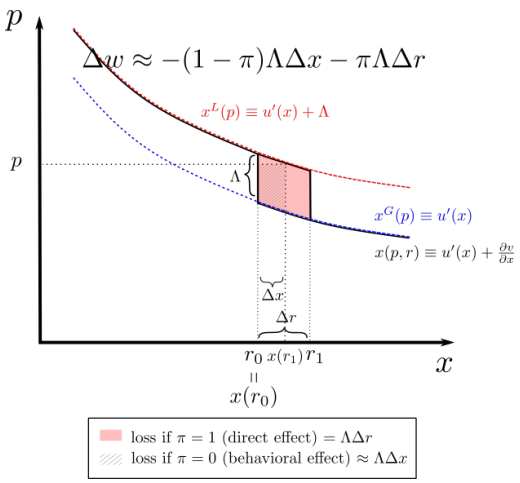
(b) Loss domain (LL)



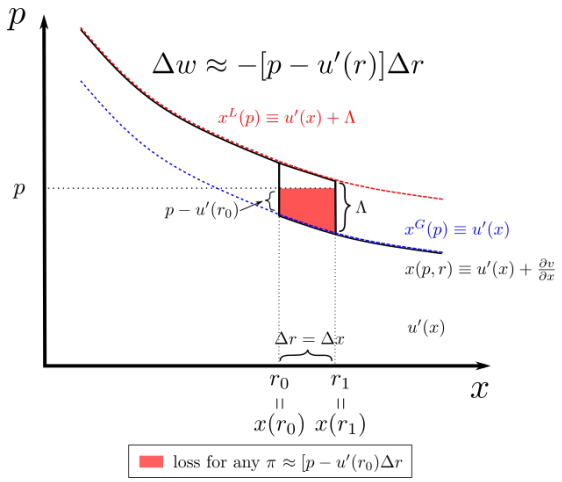
(c) Marginal Gain (GR)



(d) Marginal Loss (RL)



(e) Reference point (RR)



Notes: The figure illustrates the welfare effects of changing the reference point for individuals 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 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 π, but whether the depicted welfare loss represents a behavioral welfare effect or a direct welfare effect depends on π, so we use dark red shading.

havioral effect is positive: a price increase mitigates over-consumption of x. Using the same group notation as above, and applying equation (13) in the simple model, we obtain the following characterization of the first-order welfare effects of price changes.

Proposition 2. ***The First-Order Welfare Effects of a Change in Price.***

P2.1. ***Individual Welfare Effects.*** *Consider a change in price that is small enough that the GL group is negligible. The effect of this price change on individual welfare is approximately*

$$\Delta w_i \approx \begin{cases} -x(p_0, r)\Delta p, & i \in GG, GR, RR \\ -x(p_1, r)\Delta p - (1-\pi)\Lambda_i \frac{\partial x^L}{\partial p}\Delta p, & i \in LL, LR \end{cases} \tag{14}$$

where $\frac{\partial x^L}{\partial p}$ is evaluated at p_1 for $i \in, LL, LR$.[14]

P2.2. ***Sufficient Statistics Formula for Social Welfare.*** *The effect of a small price change Δp on social welfare is approximately*

$$\Delta W \approx \left(-(1-\pi) E\left[\Lambda_i \frac{\partial x_i^L}{\partial p} \;\middle|\; i \in L \right] P[i \in L] - E[x_i(p_0, r)] \right) \Delta p, \tag{15}$$

$$= \left(-(1-\pi) E\left[\Lambda_i \varepsilon_i^L \frac{x_i(p_0, r)}{p_0} \;\middle|\; i \in L \right] P[i \in L] - E[x_i(p_0, r)] \right) \Delta p, \tag{16}$$

where $\frac{\partial x_i^L}{\partial p}$ and $\varepsilon_i^L \equiv \frac{\partial x_i^L}{\partial p} \frac{p}{x_i^L}$ are evaluated at (p_0, r).

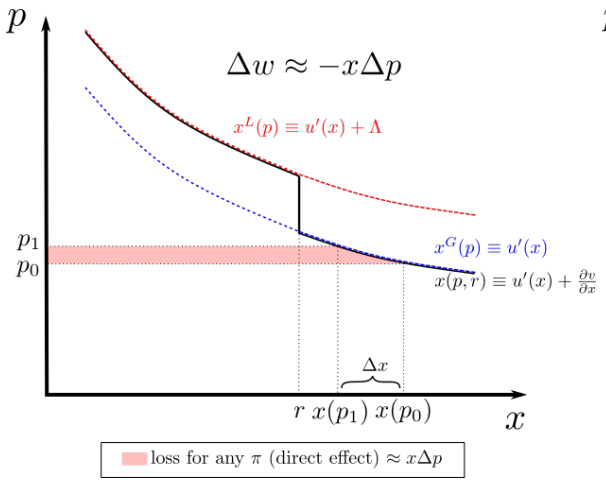
Figure 4 illustrates the welfare effects of a change in p for each case in Proposition 2.1. Again, individual welfare effects can be decomposed into direct and behavioral effects from equation (13), along with any second-order effects. In the gain domain, there are no internalities, so the direct effect of a price change is the sole first-order welfare effect (and there is a second-order welfare loss due to the behavioral response). At the reference point, internalities are present but a small change in price has no effect on behavior, so there is only a direct effect and no behavioral effect. In the loss domain, the reduction in demand combined with a negative internality under $\pi = 0$ implies a positive behavioral welfare effect.

Proposition 2.2 aggregates these effects for social welfare, where the marginal gain domain and marginal loss domain cases can again be disregarded. We find that the first-order social welfare effect of a price change consists of the standard direct effect for all individuals ($-E[x_i]\Delta p$), and a behavioral effect for the L group, which equals the marginal internality times the demand effect of the price change. Toward empirical implementation, we can express the demand effect from equation (15) in terms an elasticity in equation (16). The resulting expression provides a simple sufficient statistics formula. To approximate the welfare effect of a price change, one needs an estimate of the loss aversion parameter Λ and of the price elasticity of demand ε (in the loss domain), as well as information on the level of demand and prices and the size of the L group. Again, the sufficient statistics formula is easy to apply in many contexts, as the loss aversion parameter and demand elasticities are commonly estimated, and prices and group sizes can be directly observed. We note that in applying the formula in practice, one will have to assume that individual demand elasticities and the strength of loss aversion are independent, or estimate the covariance of the two.

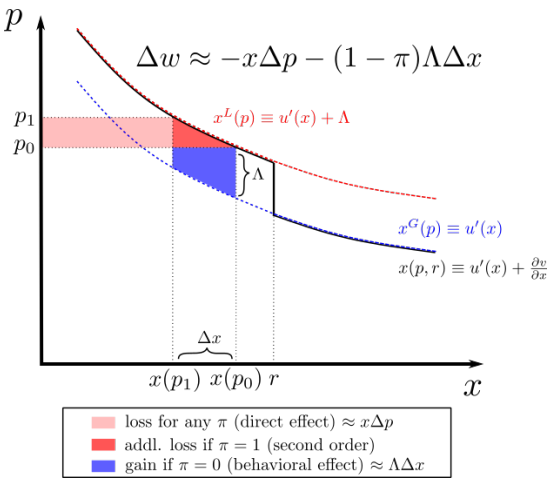
[14] Evaluating this derivative at the new price p_1 allows us to avoid the non-differentiability at the reference point for the LR group while maintaining a valid approximation (see also Figure 4). Note that one could evaluate the approximation at either p_0 or p_1 in the GG, RR, and LL cases.

FIGURE 4: WELFARE EFFECTS OF CHANGING PRICES

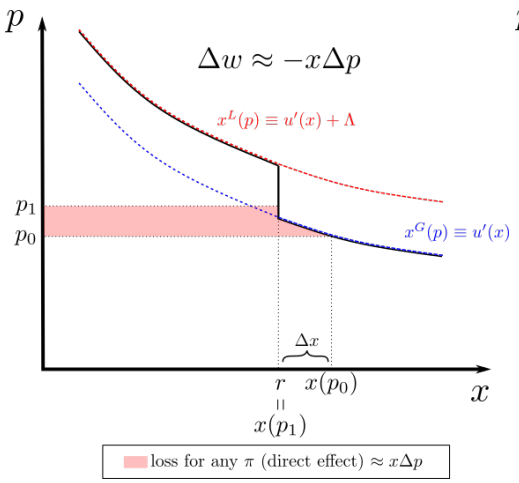
(a) Gain domain (GG)



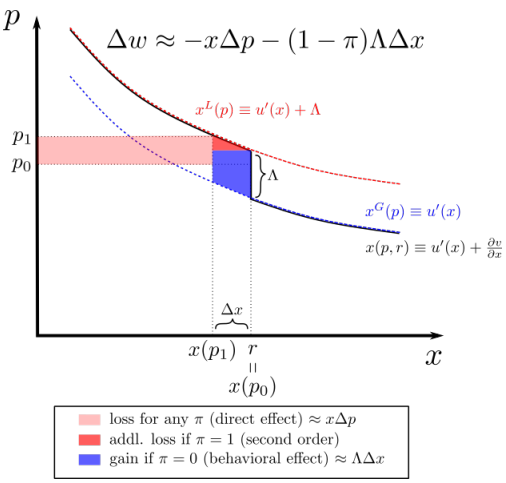
(b) Loss domain (LL)



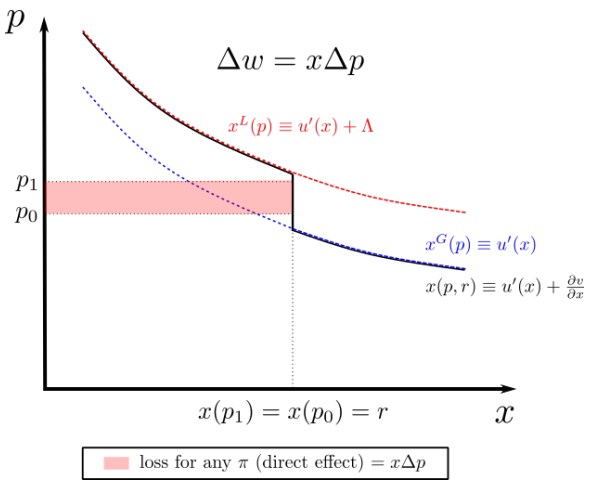
(c) Marginal Gain domain (GR)



(d) Marginal Loss domain (RL)



(e) Reference point (RR)



Notes: The figure illustrates the welfare effects changing prices for individuals 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 Figure 2. The negative direct welfare effect of a price change is depicted in red shaded regions. The positive behavioral welfare effect attributable to internalities, the normative significance of which depends on π, is depicted in blue shaded regions.

Proposition 3. ***Corrective Taxes for Reference Dependence.*** *The corrective tax schedule for good x that maximizes social welfare for a given a reference point, r, is characterized by*

$$T(x,p,r) = \begin{cases} 0 & x \geq r \\ t^*(p,r)(x-r) & x < r; \end{cases} \tag{17}$$

$$t^*(p,r) = (1-\pi)\frac{E\left[\Lambda_i \frac{\partial x_i^L}{\partial p} \middle| i \in L(p+t^*(p,r),r)\right]}{E\left[\frac{\partial x_i^L}{\partial p} \middle| i \in L(p+t^*(p,r),r)\right]}. \tag{18}$$

Depending on the planner's normative judgment, reference dependence can be viewed as leading individuals to suboptimal consumption choices. Thus, it may be natural to ask whether tax incentives should be used to correct for such mistakes. We investigate this in Proposition 3, extending work on corrective taxes with internalities (Mullainathan et al., 2012; Allcott and Taubinsky, 2015) to a situation in which whether an internality exists depends on a normative judgment. Unsurprisingly, when reference dependence carries full normative weight ($\pi = 1$), the answer is no, 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. Equation (18) 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 dependence and the demand response to a price change are independent, the optimal corrective tax simplifies to the average value of Λ_i among individuals in the loss domain. Otherwise, the covariance between Λ_i and the demand response has to be taken into account.[15]

2.3.3 Externalities

In some settings, including our empirical application, a fiscal or other externality needs to be incorporated into welfare calculations. For a linear externality equal to $\alpha E[x_i]$, the welfare effect of either a change in the reference point or in price is simply the effect from equations (10) and (13) plus $\alpha E[\Delta x_i]$, i.e. the marginal externality times the change in aggregate demand for good x.

Consider, for example, the plastic bag incentives studied by Homonoff (2018). Framing the incentive as a tax rather than a bonus effectively raises the reference point, so that the cost of using a plastic bag is evaluated in the loss domain. Without an externality, our results would suggest that this intervention decreases welfare. However, with a negative externality for plastic bag use, which of course was the motivation for introducing the incentive to begin with, the policy implication could go in the opposite direction, in favor of the loss framing. We can further infer from equation (12) that the size of the externality needed to justify the loss framing is larger when $\pi = 1$ than when $\pi = 0$, since in the $\pi = 1$ case changing the reference point has a negative direct welfare effect on those individuals always using plactic bags (the LL group).

[15] Equation (18) can be re-written as

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

3 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 their retirement decision. In this context, our goal is to characterize the welfare effects of changes to the Normal Retirement Age, and of financial incentives for delayed retirement. These types of pension reforms often debated in practice and they are closely related to our theoretical results. After describing the setting and mapping it into the theory from the previous section, we present one set of welfare results based on model simulations, and a complementary set of results based on the sufficient statistics characterizations of welfare effects from Propositions 1.2 and 2.2.

3.1 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 just over 50% (OECD, 2019), and public pensions are the main source of income for most recipients.

The first key policy dimension for the purpose of this paper are 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 the birth cohorts considered in our analysis, 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 among the sample we use for our simulations later.

The second key policy dimension is given by 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 in addition to 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 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. This situation is similar to U.S. Social Security, which features approximately actuarially fair benefit adjustment below the NRA, but higher marginal pension adjustment via the DRC above the NRA.

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).[16] 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.

3.2 Model and Parameter Estimation

From our theoretical results, the most important factors for welfare analysis are the strength of reference dependence, the number of individuals in the three groups (G, R, L), and the behavioral response to price changes. We capture these three components parsimoniously in a static model of retirement behavior with reference dependence, as in Seibold (2021). Preferences of a reference-dependent agent are[17]

$$U_i(C, R) = C - \frac{n_i}{1+\frac{1}{\varepsilon}}\left(\frac{R}{n_i}\right)^{1+\frac{1}{\varepsilon}} - \begin{cases} 0 & R < \hat{R} \\ \widetilde{\Lambda}(R - \hat{R}) & R \geq \hat{R} \end{cases} \tag{19}$$

where C is lifetime consumption and R is the worker's retirement age relative to a career starting age normalized to 0. The parameter $\widetilde{\Lambda}$ captures the strength of reference dependence. 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 ε 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. Loosely speaking, $\widetilde{\Lambda}$ can be identified by the behavioral response to statutory retirement ages holding financial incentives fixed, ε can be identified by the behavioral response to financial incentives, and the distribution of n_i can be specified to fit the observed variation in retirement ages.

Importantly, equation (19) assumes *loss aversion in lifetime leisure*. The last term in the equation captures reference dependence as in the simple model from Section 2. 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 in lifetime leisure, where

[16] Due to the extended closure of the research data center, the results currently shown in this paper had to be obtained from a random 1% sample.

[17] 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 theory from Section 2.

workers perceive postponing retirement as a loss relative to a normal time to retire.

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

$$C(R) = \sum_{t=0}^{R-1} \delta^t w_t (1 - \tilde{\tau}_t) + \sum_{t=R}^{T} \delta^t B(R) \tag{20}$$

where w is the gross wage per period, $\tilde{\tau}$ is the payroll tax/pension contribution rate, T is the time of death, and δ is the discount factor.[18] 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 the model.[19] As Seibold (2021) shows, the model predicts bunching at the NRA when it is perceived as a reference point by workers. 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

$$\frac{b}{\hat{R}} = \left(\frac{1 - \tau}{1 - \tau - \Delta\tau - \Lambda} \right)^{\varepsilon} - 1, \tag{21}$$

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.[20]

We use the identification strategy of Seibold (2021) in order to estimate Λ and ε. 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 i, bunching can be written as

$$\frac{b_i}{\hat{R}_i} = \left(\frac{1 - \tau_i}{1 - \tau_i - \Delta\tau_i - \Lambda \cdot D_i} \right)^{\varepsilon} - 1 + \xi_i \tag{22}$$

where D_i is an indicator for the NRA and ξ_i is an error term.[21]

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

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

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

[20] We assume that Λ is homogeneous across individuals for simplicity.

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

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 (22), 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.

3.3 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; Manoli and Weber, 2018; Cribb et al., 2016), which is largely driven by shifting individuals' reference points to a higher retirement age (Behaghel and Blau, 2012; Seibold, 2021).

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 3.2. Thus, we should conceptually think of a reform that increases the NRA as one that *lowers the reference point* in the sense of Proposition 1. Second, while our theory 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. If this feature is maintained, increasing the NRA would also move the non-convex kink in the budget constraint to the new NRA. Moreover, if the NRA coincides with the FRA, the age from which the "full" pension is available may move upwards with the NRA, such that increasing the NRA effectively implies a benefit cut across the board.

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., 2021). 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" in the sense of Proposition 3, which incentivizes individuals to move away from the reference point of the NRA by increasing their retirement age.

3.4 Welfare Effects of Pension Reforms: Simulation Approach

As a first empirical approach, we can use the model from Section 3.2 to simulate individual retirement behavior under different policy scenarios and to calculate the welfare effects of pension reforms.

3.4.1 Simulation Methods

We simulate the welfare effects of two pension reforms of the types discussed above, 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. The reform shifts individuals' retirement reference points and entails a (relatively small) change in the budget constraint. In order to maintain the feature of a budget constraint kink at the NRA, the DRC only applies above the new NRA in the simulation. However, the counterfactual scenario does not feature a benefit cut across the board below the NRA in order to avoid confounding the effects of influencing reference points with large mechanical fiscal and consumption effects. The second reform is an increase in the DRC. In order to anchor the second reform, 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 obtain this counterfactual distribution by fitting 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.[22] 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 the two counterfactual policy scenarios. For this, we simulate individual lifetime budget constraints from equation (20) as in Seibold (2021), based on observed individual earnings and contribution histories, and choose the retirement age that maximizes utility from equation (19) 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 each of the following outcomes: contributions to the pension system, benefits paid to workers, workers' lifetime consumption. Moreover, we calculate the effects on disutility from working and reference dependence payoffs given the preferences in equation (19). 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

3.4.2 Main Simulation Results

Table 1 summarizes the effects of the two simulated pension reforms. Appendix Figure A2 provides further illustration of how the different components sum up to welfare effects.

[22] 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.

Increasing the Normal Retirement Age. Column (1) shows the effects of the NRA increase. Shifting the NRA by one year increases average actual retirement ages by 7.3 months. As shown in Seibold (2021), such a reform improves the fiscal balance of the pension system. 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 +€10.4k per worker. Next, the reform affects individual welfare. Lifetime consumption increases by around +€6.9k 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 Proposition 1. In words, 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 around +€12.5k per worker.

In addition, if the planner places normative weight on reference dependence ($\pi = 1$), we should also account for changes in reference dependence 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 about -€11.2k due to additional disutility from work, and a positive component +€13.0k from the decrease in the reference point itself.[23] When $\pi = 1$, the first of these modifies the behavioral 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 -€9.1k. We observe that this behavioral welfare effect and the net fiscal effect (+€10.4k) 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 $\pi = 1$. Panel (b) of Figure A2 provides a visual illustration of this offsetting.[24]

With the behavioral effect largely eliminated under $\pi = 1$, the direct welfare effect, i.e. the effect on reference dependence 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 around +€14.3k, even larger than under $\pi = 0$. Building on the logic above, most of this welfare effect (+€13.0k) is attributable to the direct effect of the change in the reference point. Consistent with Proposition 1.2, the total welfare effect is larger under $\pi = 1$ than under $\pi = 0$.[25]

Appendix Table A3, Panel A, shows average welfare effects of the NRA increase by groups, where each group is defined by their retirement age relative to the NRA before and after the reform analogously to Section 2.3.1. For instance, the LR group consists of workers who retire in the loss domain above the old NRA before the reform, but retire at the new NRA after the reform. Most importantly, Table A3 shows that the key theoretical intuition about behavioral and direct effects carries through to each sub-group. When $\pi = 0$, workers retire sub-optimally early, so the groups whose behavior is affected by the NRA change –

[23] See Appendix F.1 for details of this decomposition of reference dependence payoffs.

[24] 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.

[25] This occurs because the R group in Proposition 1.2 experiences the same welfare effect regardless of π, while the L group experiences a positive direct welfare effect of a lower reference point when $\pi = 1$, and no welfare effect when $\pi = 0$. 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 theory. 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.

all but the LL group – experience a positive total welfare effect (net of fiscal effects). But when $\pi = 1$, the additional behavioral effect via reference-dependent disutility from work largely eliminates this effect. The main determinant of welfare when $\pi = 1$ is the direct welfare effect via reference-dependent utility from the reference point itself. As in theory, this effect is not present in the gain domain.[26]

TABLE 1: WELFARE EFFECTS OF PENSION REFORMS

	(1) Policy 1: Normal Retirement Age to 66	(2) Policy 2: Delayed Retirement Credit to 10.44%
Contributions collected	+3,865	+3,658
Benefits paid	+6,553	–6,448
Net fiscal effect	+10,419	–2,790
Worker consumption	+6,932	+19,375
Disutility from work	–4,834	–3,360
Worker welfare ($\pi = 0$)	+2,098	+16,015
Ref. dep. disutility from work	–11,200	–13,930
Ref. dep. utility from ref. point	+13,021	0
Worker welfare ($\pi = 1$)	+3,919	+2,085
Total welfare ($\pi = 0$)	+12,517	+13,225
Total welfare ($\pi = 1$)	+14,337	–705

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 (+7.3 months). 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. 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 (2) 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 the NRA increase. However, a first important difference to the NRA reform is the fiscal effect. The net fiscal effect is negative at –€2.8k per worker. Workers also contribute for longer in this scenario, but the positive effect on contributions is more than offset by the large increase in benefit payments.[27] 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 just how much later to retire. Thus, there is a large positive effect of +€16.0k 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 –€13.9k 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

[26] Relatedly, we observe that for the RR and RG groups, the total welfare effect of a change in the NRA does not depend on π, as shown in Proposition 1.1.

[27] 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.

effect into account, individual welfare increases only by +€2.1k under $\pi = 1$. Finally, the total welfare effect is positive at +€13.2k 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 the theoretical results in Proposition 2. 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 welfare effect by (partially) correcting the internality. 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 show in the extended simulations below.

3.4.3 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 may be warranted in interpreting the findings as we are extrapolating further from observed data than before.

While Table 1 considers a specific change to the Normal Retirement Age, Appendix Figure A3 shows results for a range of simulated counterfactual NRAs between 65 and 67 (in monthly increments). Overall, the figure confirms the robust positive welfare effects of increasing the NRA. To begin with, Panel (a) shows that the fiscal balance of the pension system increases with the NRA. In Panel (b), individual consumption increases with the NRA, but workers also experience higher disutility from working longer. Adding up those two components of standard preferences and the net fiscal effect, total welfare under $\pi = 0$ increases with the NRA. In Panel (c), we add reference dependence payoffs in order to obtain welfare effects under $\pi = 1$. As in Table 1, incorporating reference-dependent disutility from work eliminates most of the behavioral welfare effect when $\pi = 1$, but the simultaneous introduction of a positive, direct welfare effect of increasing the reference point leads to an overall increase in welfare. As Panel (d) shows, total welfare increases monotonically with the NRA both under $\pi = 0$ and $\pi = 1$, where the welfare increase is stronger under $\pi = 1$.

Similarly, Appendix Figure A4 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. In Panel (a), 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 to actuarial fairness. Panel (b) shows that under $\pi = 0$ the total welfare effect of increasing the credit is positive throughout the large range we consider, because consumption increases by more than disutility from work and the negative fiscal effects, 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. While increasing the NRA always increases total welfare, both under $\pi = 0$ and $\pi = 1$, Panels (b) to (d) of Figure A4 show that it is possible to find an optimal level of the DRC. Importantly, 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 $\pi = 0$, as theoretically shown in Proposition 3. Such a large marginal financial return to working longer, or implicit price of leisure, can induce workers to retire later and move towards their optimal retirement age. In Panels (c) and (d), 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 the main ideas from our theoretical results. Increasing the NRA, which corresponds to lowering reference points in terms of lifetime leisure, yields robust increases in total welfare. Increasing the DRC, which corresponds to an implicit tax on leisure in the loss domain, increases total welfare 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. Two further policy implications of our simulation results are worth noting. First, if policymakers are mainly concerned about the fiscal sustainability of pension systems, increasing the NRA may be attractive in its own right as it yields sizable positive fiscal effects. Higher late retirement subsidies tend to yield negative fiscal effects, on the other hand. Second, if policymakers are uncertain about the appropriate welfare judgment of reference dependence, increasing the NRA appears even more attractive, as the positive welfare effect of this reform is robust to the choice of π. The sign and magnitude of the welfare effects of the DRC, on the other hand, strongly depend on this normative judgment.

3.5 Using Sufficient Statistics to Calculate Welfare Effects of Pension Reforms

As a second empirical approach, we can use the sufficient statistics formulas from Propositions 1.2 and 2.2 to approximate the welfare effects of pension reforms. Compared to the full individual-level simulations, this approach is substantially easier to implement. Adapting equation (12) to the retirement context, we can express the first-order welfare effect of a small change in the Normal Retirement Age as

$$\begin{aligned} \Delta W \approx\ & \Delta \hat{R}\ \pi\ E[\Lambda w_i |\ i \in L]\ P[i \in L] \\ & + \Delta \hat{R}\ E\left[\left(\frac{\Lambda}{2} + \tau_i \right) w_i \ \middle|\ i \in R \right] P[i \in R] \end{aligned} \tag{23}$$

where, as in Section 3.2, $\hat{R}$ is the retirement age reference point (given by the NRA), w is the gross wage per period, τ is the implicit tax rate on working for an additional period, and Λ 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

$$\Delta W \approx \left(E\left[\{-\tau_i - (1-\pi)\Lambda\} w_i \frac{\partial l_i^L}{\partial [w_i(1-\tau_i)]} \Delta\tau_i w_i \ \middle|\ i \in L \right] P[i \in L] \right) \tag{24}$$

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 2.3 to the retirement setting, several 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 τw, 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.[28] 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 (24) is the responsiveness of lifetime leisure to the implicit net wage. As we have estimated the retirement age elasticity ε 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 τ, 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 (23) and (24) empirically. We require values of the reference dependence parameter Λ and the price elasticity ε, which we have estimated, as well as information on average wages and implicit tax rates, which we directly calculate from the data. Appendix Table A4 summarizes all parameter values used as inputs into the sufficient statistics formulas.

Table 2 shows results based on the sufficient statistics approach. For comparison, welfare effects based on the simulation approach are also 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 +€10.9k under $\pi = 0$ and +€14.2k under $\pi = 1$, which is very similar to the effects of +€12.5k and +€14.3k, respectively, from the simulation approach. In Column (2), the sufficient statistics formulas yield a welfare effect of +€4.8k under $\pi = 0$ and +€1.8k 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. As Panel B of Appendix Table A3 shows, more than half of the total welfare effect of increasing the DRC under $\pi = 0$ is driven by the RL group, i.e. by individuals de-bunching away from the NRA towards older retirement ages. The sufficient statistics characterization disregards the effect on the RL group because this 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 3.4.3). The local approximation of the sufficient statistics approach captures the initial increase in welfare, but cannot account for a large DRC increase 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 +€0.5k vs. +€0.9k in the case of $\pi = 0$. Under $\pi = 1$, the sufficient statistics approach now correctly yields a small positive effect of +€0.2k, compared to a simulated effect of

[28] The impact of fiscal externalities can be captured by direct and behavioral effects on government revenue. 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, since the direct revenue effect offsets the direct effect of the tax change on individual welfare.

around +€0.1k.[29]

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 ($\pi = 0$)	+10,853	+4,809	+520
Total welfare ($\pi = 1$)	+14,204	+1,791	+194
Simulation Approach			
Total welfare ($\pi = 0$)	+12,517	+13,225	+912
Total welfare ($\pi = 1$)	+14,337	-705	+74

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 (23) and (24), and the last two rows show simulated welfare effects described in 3.4. 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 65 and above, and are in Euros per worker, in terms of net present value at age 65.

Compared to the simulation approach, using the sufficient statistics formulas has advantages and disadvantages. Perhaps its most important advantage is that the sufficient statistics approach is substantially easier to implement than full-fledged microsimulations. Estimates of the reference dependence parameter and a price elasticity are required, which can be obtained using reduced-form methods in many settings. Moreover, the sufficient statistics approach is based on the less restrictive assumptions about preferences from Section 2.2, while the simulations require assuming a specific utility function. 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 Extensions

In this section, we consider three extensions to the simple model laid out in Section 2, building on the rich theoretical literature on reference dependence. We focus here on how incorporating these extensions changes our intuition and key welfare results. A more formal treatment of the extensions is provided in Appendix B.

[29] Remaining discrepancies between sufficient statistics and simulated effects occur due to a non-negligible RL group existing even for small reforms, and simulated effects exhibiting some non-linearity even locally around the status quo.

4.1 Multi-Dimensional Reference Dependence

So far, we considered reference dependence along a single dimension of the menu space, motivated by the empirical evidence in Section 2.1. In some contexts, however, reference-dependent payoffs may be present in more than one dimension. In Appendix B.1, we analyze welfare in a two-dimensional model of reference dependence.[30] In our empirical application, we can interpret the two dimensions as representing reference dependence over leisure and consumption, similarly to Crawford and Meng (2011), who argue that a two-dimensional model explains well the daily labor supply behavior of taxi drivers.

4.1.1 Insights from the Two-Dimensional Model

Relative to the simple model in Section 2, the main new element we add reference dependence over good y, using an otherwise similar formulation to equation (2). We denote the loss aversion parameter for good y – the analogue of Λ_i for good x – by Γ_i. Decision utility under two-dimensional reference dependence is

$$U_i(x,y) = u_i(x) + y + \mathbb{1}\{x \leq r_x\}\Lambda_i(x - r_y) + \mathbb{1}\{y \leq r_y\}\Gamma_i(y - r_y) \tag{25}$$

For many applications, it seems natural to assume that the reference point for goods x and y falls on the budget constraint, i.e. $pr_x + r_y = z_i$. This would imply that the gain domain for good x coincides with the loss domain for good y. In our application, for example, if the NRA pins down the reference point for consumption and leisure, retiring before the NRA will entail a gain in leisure and a loss in consumption, while retiring after the NRA will entail a loss in leisure and a gain in consumption. With this assumption, using policy to increase the reference point for good x also decreases the reference point for good y. Figure 5 depicts demand for good x in the gain and loss domains in the two-dimensional model. Because of loss aversion over good y, demand in the absence of reference-dependent payoffs ($p = u'(x)$), falls *between* gain- and loss-domain demand for x.

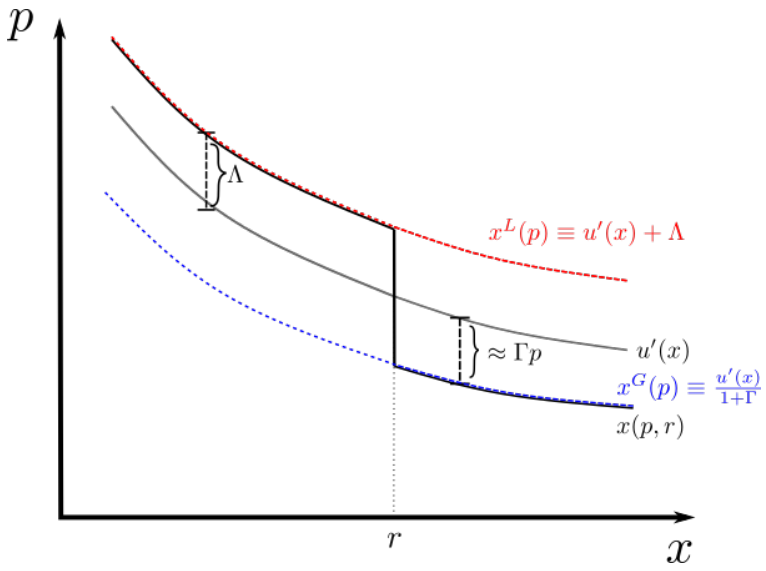
We can understand how considering multiple dimensions of reference dependence changes our welfare effects by adapting the direct and behavioral effects from equations (10) and (13). In the loss domain for good x, the direct and behavioral effects are similar to the simple model because being in the gain domain for good y has no additional implications for welfare. For example, when $\pi = 0$ these individuals continue to experience a negative internality from consuming good x, which is proportional to loss aversion over x (Λ_i). However, new effects appear for individuals consuming in the gain domain for x due to loss aversion over good y. These welfare effects are proportional to loss aversion over good y (Γ_i), and opposite-signed to the analogous effects due to loss aversion over good x. When $\pi = 0$, individuals experience a *positive* internality from consuming good x in the gain domain for good x. Similarly, potential direct effects due to loss aversion over y appear for individuals consuming in the gain domain for good x. Whether direct and behavioral effects matter hinges on π in the same fashion as before.

We illustrate the individual welfare effects of changing reference points and prices in Appendix Figures B1 and B2. Decreasing the reference point for good x no longer yields a robust welfare improvement if we incorporate loss aversion over good y, because increasing the reference point for good x will decrease the reference point for good y by assumption.[31] For an individual in the gain domain for good x and the loss

[30] Extending this analysis to arbitrary dimensionality of the goods space is straightforward, but it will be difficult to identify a many-dimensional model empirically. As we illustrate in Section 4.1.2, identifying reference dependence parameters in two dimensions is already challenging. Alternatively, one could impose further theoretical restrictions to make identification more tractable. For instance, Kőszegi and Rabin (2006) propose a functional form restriction under which there just two parameters governing reference-dependent payoffs in N dimensions. However, such theoretical restrictions may not be empirically valid in all contexts.

[31] If it were feasible for a planner to decrease the reference point for good x or good y in isolation, without affecting the reference

FIGURE 5: OBSERVED DEMAND, WELFARE-MAXIMIZING DEMAND, AND MARGINAL INTERNALITIES UNDER TWO-DIMENSIONAL REFERENCE DEPENDENCE



Notes: The figure depicts observed demand $x(p, r)$ at a given reference point r in the black line. We also plot demand in the gain and loss domains, $x^G(p)$ (in blue) and $x^L(p)$ (in red), as well as $u'(x)$ (in grey). The vertical distance between observed demand and $u'(x)$ is Λ in the loss domain and Γp in the gain domain. When $\pi = 1$, observed demand is welfare maximizing. When $\pi = 0$, welfare-maximizing demand coincides with $u'(x)$. The marginal internality under $\pi = 0$ is $-\Lambda$ in the loss domain and Γp in the gain domain.

domain for good y, moving the reference point along the budget constraint in this way shrinks individuals losses over good y (the direct effect, relevant when $\pi = 1$), and it mitigates potential over-consumption of good y out of loss aversion over y (the behavioral effect when $\pi = 0$). Likewise for a change in prices, the sign of the behavioral welfare effect is now ambiguous. When the price of good x increases, individuals in the loss domain for good x continue to experience a positive behavioral welfare effect from mitigating over-consumption of x. However, individuals in the loss domain for good y instead experience a negative behavioral welfare effect from exacerbating over-consumption of good y.

How much do these new effects matter for overall social welfare? The answer to this question turns on (1) the number of individuals in each domain, and (2) the relative magnitude of loss aversion over goods x and y. We next turn to our empirical application to illustrate this point.

4.1.2 Disciplining Dimensionality Empirically

We can infer from the similarities in demand curves between Figures 2 and 5 that separately identifying reference dependence over different goods from within-individual demand is challenging. However, one possibility to quantify the relative importance of reference dependence in given dimension is to examine the character of bunching around the reference point more closely.

In our empirical application, besides reference dependence in leisure, there could also be reference dependence in the consumption dimension, for instance because "full" pension benefits become available at the NRA and individuals perceive the associated consumption level as a reference point (Behaghel and Blau 2012). We specify a model with both reference dependence over leisure and over consumption in Appendix

point for the other good, doing so would generate a robust Pareto improvement exactly like in the simple model.

F.2. Preferences are identical to the initial specification in equation (19) except that (1) we add a component of utility to capture reference-dependent payoffs over consumption, and (2) we denote the loss aversion parameter in the leisure dimension by Λ_l and in the consumption dimension by Λ_c.

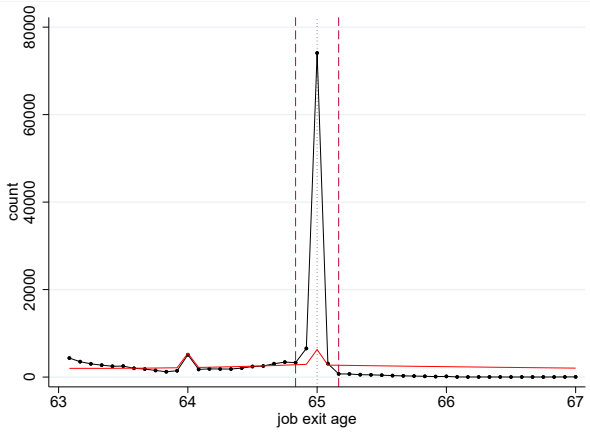
Figure 6 compares the empirical retirement age distribution around the NRA (Panel a) to stylized predicted distributions in three cases: when reference dependence is present over leisure and not consumption (Panel b), over consumption and not leisure (Panel c), and over both consumption and leisure (Panel d). Under reference dependence over leisure, the model from Section 3.2 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 F.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.

As we highlighted in Figure 1, the empirical density in Panel (a) of Figure 6 exhibits a visible downward shift above the NRA, suggesting that the dominant driver of bunching is reference dependence over leisure. However, this does not necessarily exclude *any* degree of reference dependence over consumption. The arising empirical challenge is that reference dependence parameters in both dimensions cannot be separately identified based on observed bunching at the NRA alone. A given amount of bunching could be rationalized by reference dependence over leisure, reference dependence over consumption or a combination of the two. We propose two approaches to make further progress on two-dimensional reference dependence in our empirical application. First, we can calculate a range of combinations of Λ_l and Λ_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 A5 shows estimated parameter combinations consistent with the observed amount of excess mass. The negative slope of the relationship illustrates that the two types of reference dependence are substitutes in terms of rationalizing observed excess mass. The higher the assumed share of bunching from the left, the larger the implied Λ_c, but the smaller the implied Λ_l. The labeled dots mark parameter combinations corresponding to selected left bunching shares.

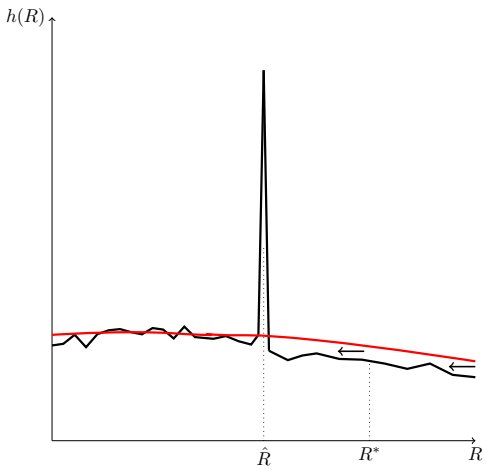
The possible range of Λ_c shown in Panel (a) of the figure is still wide, which might create ambiguity in welfare. Thus, a second approach is to obtain a preferred estimate of Λ_c using the information contained in the observed retirement age distribution around the NRA. Appendix F.2.2 provides more details of this procedure. Intuitively, the counterfactual density – which would prevail in the absence of any reference dependence – 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. Panel (b) of Appendix Figure A5 illustrates the procedure and confirms that the implied density shift is much more substantial above than below the NRA, implying a point estimate of the left bunching share of 13.3%. This magnitude of relative bunching implies a consumption reference dependence parameter of $\Lambda_c \approx 0.67$ and a leisure reference dependence parameter of $\Lambda_l \approx 0.46$.

FIGURE 6: BUNCHING AND THE DIMENSIONS OF REFERENCE DEPENDENCE

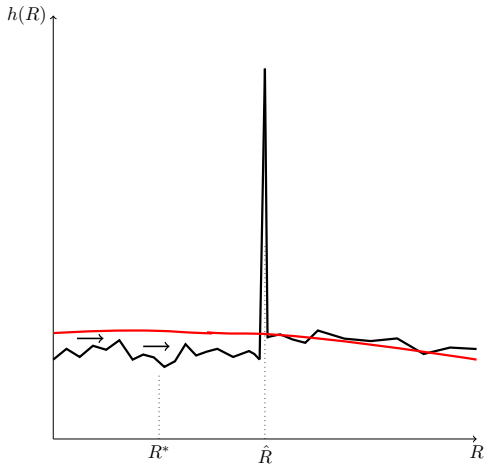
(a) Empirical Retirement Density



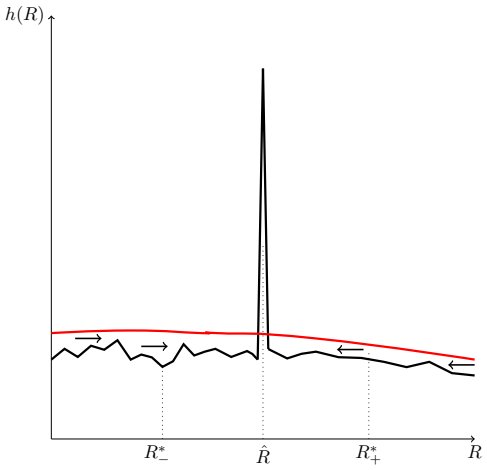
(b) Reference Dependence over Leisure



(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 3.2, Panel (c) corresponds to reference dependence over consumption as described in Appendix F.2, and Panel (d) corresponds to reference dependence in both dimensions.

4.1.3 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 same policies considered in Table 1 under our preferred two-dimensional reference dependence parameter estimates. Note that 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.

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 rational ($\pi = 1$). In the $\pi = 0$ case, some of the behavioral effect now comes from workers who are retiring *too late* out of loss aversion over consumpion, and increasing the NRA exacerbates this internality (while still mitigating the internality for those above the NRA who retire too early). As a result, the behavioral effect under $\pi = 0$, i.e. the net effect of increased worker consumption and larger disutility form work, is smaller than in Table 1, and the total welfare effect under $\pi = 0$ is somewhat reduced. 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.

We also calculate welfare effects of the NRA reform for a wider range of values of Λ_c in order to account for some uncertainty which remains in estimating the left-bunching share in Figure A5. Figure 7 shows welfare effects under $\pi = 0$ and $\pi = 1$ for Λ_c between zero and around 2. The dashed vertical lines mark the cases where the left bunching share is zero (corresponding to the baseline simulation), 13% (our preferred estimate) and 27% (double our preferred estimate). The welfare effects of the NRA reform generally decrease with Λ_c. Under $\pi = 0$, the effect remains positive over the range in the graph, but under $\pi = 1$, the effect turns negative around $\Lambda_c = 1$, corresponding to a left bunching share of around 18%. The faster decline of the welfare effect with Λ_c under $\pi = 1$ is due to growing negative direct effects on pre-NRA retirees.[32]

Overall, these results highlight a potential caveat to our finding from Section 3 that increasing the NRA robustly increases welfare. Namely, the welfare effects of a change in the NRA under $\pi = 1$ can turn negative

[32]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.

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	+4,801	–4,105
Net fiscal effect	+7,686	–1,778
Worker consumption	+5,336	+12,308
Disutility from work	–5,392	–2,258
Worker welfare ($\pi = 0$)	–56	+10,050
Ref dep disutility from work	–9,015	-8,780
Utility from retirement ref point	+10,198	0
Ref dep utility from consumption	+721	0
Disutility from consumption ref point	–6,821	0
Worker welfare ($\pi = 1$)	–4,973	+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 and an increase in the Delayed Retirement Credit to 10.44% as in 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.

under strong consumption reference dependence. In our empirical setting, the welfare effects of increasing the NRA remain positive under our preferred estimates of two-dimensional reference dependence parameters regardless of the value of π, and the positive welfare effects under $\pi = 0$ are robust to a wide range of reference dependence parameters. Nevertheless, the relative strength of reference dependence over leisure vs. consumption could depend on how retirement and benefit rules are communicated to workers in a pension system, and thus some caution may be warranted when extrapolating our results to other settings.

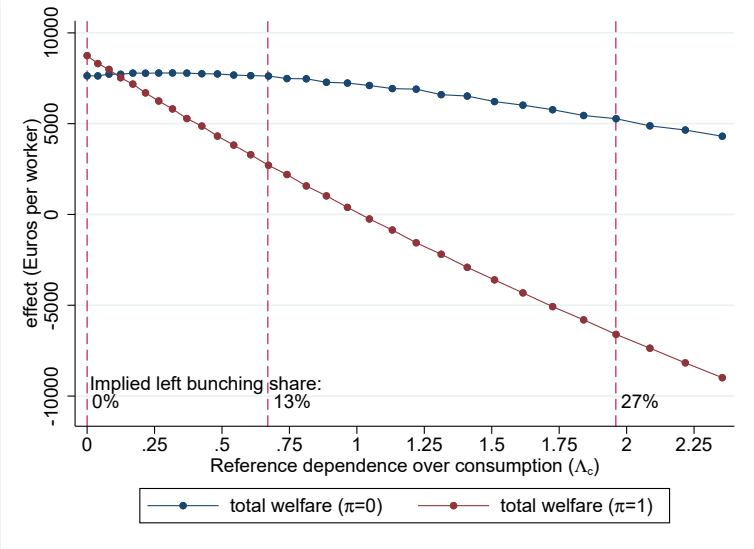
4.2 Reference-Dependent Gain Domain Payoffs

We next consider how our main welfare results change when individuals experience a reference-dependent payoffs in the gain domain rather than payoffs only due to loss aversion. Returning to the case of one-dimensional reference dependence, we incorporate these payoffs using a form based on Tversky and Kahneman (1991):[33]

$$U_i(x, y) = u_i(x) + y + \begin{cases} \eta_i(x - r), & x > r \\ (\eta_i + \Lambda_i)(x - r), & x < r, \end{cases} \tag{26}$$

[33]This is slightly different from the the formulation of Tversky and Kahneman (1991) with canonical parameters η and λ. We use this formulation in order to maintain the interpretation of the Λ parameter as governing the strength of loss aversion as in the simple model, while also adding gain domain payoffs. See Appendix B.2 for more detail.

FIGURE 7: WELFARE EFFECTS OF INCREASING THE NRA BY STRENGTH OF CONSUMPTION REFERENCE DEPENDENCE



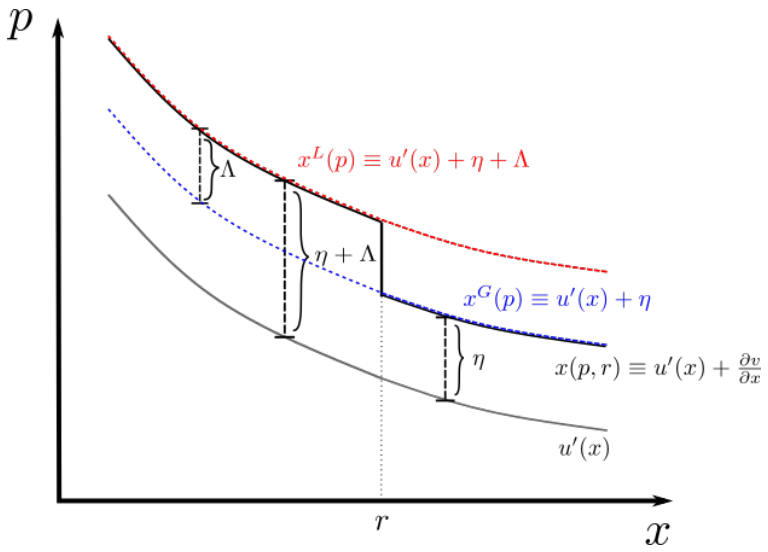
Notes: The figure shows the total welfare effects of increasing the NRA from 65 to 66 by the strength of consumption reference dependence Λ_c. Simulations are conducted for birth cohort 1946. The effects are calculated among workers retiring at age 63 and above, and are in Euros per worker, in terms of net present value at age 65. The dashed vertical lines denote selected values of Λ_c, corresponding to implied left bunching shares of 0 (no consumption reference dependence), 13% (our preferred estimate, on which the results in Table 3 are based), and 27% (twice our preferred estimate).

The parameter η now governs the strength of reference-dependent payoffs apart from loss aversion. Empirically, η is typically unidentified in choice data because it cannot be separately identified from marginal utility from non-reference-dependent payoffs $u_i'(x)$. In other words, the same observed behavior could be generated by reference-dependent preferences with or without gain domain payoffs. We demonstrate this formally in Appendix D (see also Barseghyan et al., 2013). Accordingly, there is little existing evidence on the empirical magnitude of η.

Figure 8 illustrates how behavior in this model relates to the primitives. Marginal utility from non-reference-dependent payoffs $u_i'(x)$ now falls *below* both gain and loss domain demand. The gap between gain domain demand and $u_i'(x)$ is η_i. As before, loss aversion further boosts demand by Λ_i in the loss domain.

Again, the direct and behavioral effects decomposition from equations (10) and (13) helps us understand how the introduction of η_i modifies welfare effects. First, we note that in the simple model, internalities under $\pi = 0$ are negative and only present in the loss domain. With gain domain payoffs, there is a negative internality in the gain domain governed by η_i, and an even larger negative internality in the loss domain governed by $\eta_i + \Lambda_i$. Under $\pi = 0$, a decrease in the reference point mitigates over-consumption of x for individuals in both gain and loss domains rather than just those in the loss domain. In the $\pi = 1$ case, the internalities vanish but decreasing the reference point confers positive direct effects on individuals in the gain domain, and larger positive direct effects on those in the loss domain. Together, larger negative internalities and larger negative direct effects imply that a decrease in the reference point is a robust Pareto improvement as in the simple model, and indeed the magnitude of welfare effects becomes even larger. Meanwhile, for a price increase, the additional payoff due to η amplifies negative internalities when $\pi = 0$,

FIGURE 8: OBSERVED DEMAND, WELFARE-MAXIMIZING DEMAND, AND MARGINAL INTERNALITIES WITH GAIN-DOMAIN PAYOFFS



Notes: The figure depicts observed demand $x(p, r)$ at a given reference point r in the black line. We also plot demand in the gain and loss domains, $x^G(p)$ (in blue) and $x^L(p)$ (in red), as well as $u'(x)$ (in grey). The vertical distance between observed demand and $u'(x)$ is $\eta + \Lambda$ in the loss domain and ηp in the gain domain. When $\pi = 1$, observed demand is welfare maximizing. When $\pi = 0$, welfare-maximizing demand coincides with $u'(x)$. The marginal internality under $\pi = 0$ is $-(\eta + \Lambda)$ in the loss domain and $-\eta$ in the gain domain.

creating even larger positive behavioral effects. Moreover, there is now scope for corrective taxation in both gain and loss domains, whereby the optimal corrective tax schedule features a higher rate in the loss domain.

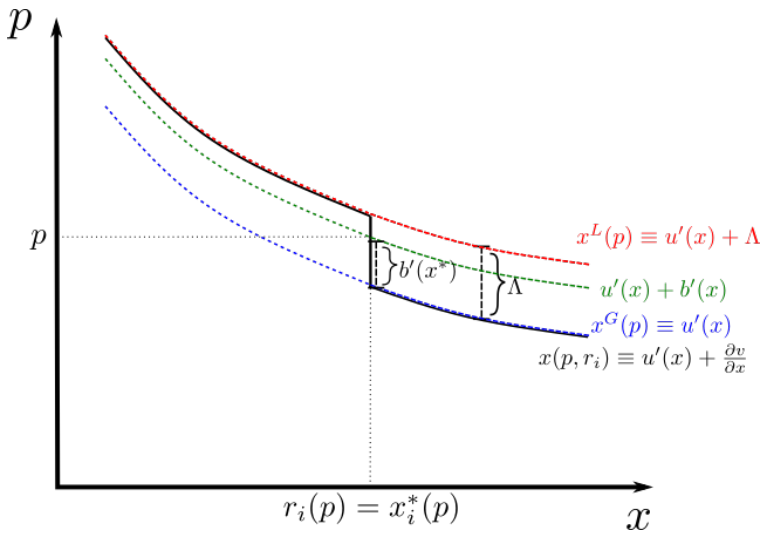
In summary, the signs of the main welfare effects of interest do not change when introducing gain domain payoffs, but in fact the magnitude of effects increases and the effects materialize for all individuals rather than only in the loss domain. As η_i cannot be easily identified empirically, we do not quantify how results from our empirical application would be modified. Qualitatively, we know that the welfare gains from increasing the NRA, and when $\pi = 0$, from incentivizing later retirement via the DRC, would be strengthened.

4.3 Reference Points as Goals

In our empirical application and the other settings discussed in Section 2.1, reference points are determined by policy. Thus, we have ignored the possibility that individuals may influence their own reference points so far. However, part of the literature on reference dependence examines situations where reference points serve as goals in order to over-come another bias, such as present bias (Koch and Nafziger, 2011, 2016; Loewenstein and O'Donoghue, 2006). This type of framework contains two new elements compared to our simple model: an additional behavioral bias, which provides a reason why the individual does not always prefer the lowest possible reference point, and the possibility that individuals set their own reference points with some degree of sophistication about their bias. We can use a similar model to consider the case where individuals choose their own reference point, possibly optimally, out of another concern.

To extend our theory along these lines, we first suppose that decision utility governing the choice of

FIGURE 9: THE OPTIMAL REFERENCE POINT UNDER GOAL-SETTING



Notes: The figure illustrates the optimal choice of reference point given an additional bias $b(x)$ and reference-dependent preferences over x, as in the goal-setting model from Section 4.3. Observed demand $x(p, r)$ is shown in the black line, and demand in the gain and loss domains is shown in blue and red, respectively. The green line depicts $u'(x) + b'(x)$, corresponding to welfare-maximizing demand under $\pi = 0$. The marginal bias $b'(x)$ drives a wedge between welfare-maximizing demand and demand in the gain domain. Under the assumption that $b'(x^*) < \Lambda$, the individual sets a reference point r_i to completely correct the bias and then chooses x exactly at the reference point.

x is the same as in our original model, i.e. equations (1) and (2). Because reference dependence induces individuals to consume more in order to avoid losses, using reference points to overcome biases is generally useful when biases lead to under-consumption of some good. As such, the new component entering welfare is an additional positive internality from consuming good x, which we model in a reduced-form fashion as follows:

$$U_i^*(x, y) = u_i(x) + y + b_i(x) + \pi v_i(x|r). \tag{27}$$

The new bias term $b_i(x)$ captures the extent to which the individual under-values x in their decision. We assume $b_i'(x) > 0$, i.e. a positive marginal internality from $b_i(x)$, and $b_i''(x) \leq 0$, which ensures an interior solution for the welfare-maximizing choice of x. As a benchmark, we assume individuals are *fully sophisticated* about their biases. That is, the individual is perfectly aware of the bias $b(x)$ when they set their reference point. We also make a simplifying assumption, namely that the bias $b_i(x)$ is small enough that it can be overcome by setting a reference point. Formally, $b_i'(x_i^*) \leq \Lambda_i$, where $x_i^* = \arg\max u_i(x) + b_i(x) + z - px$. [34]

Allowing the individual to choose r changes the welfare economics of the model significantly. A fully sophisticated individual will choose a reference point of $r_i = x_i^*$. The individual subsequently chooses x exactly at the reference point. In other words, fully sophisticated individuals set a goal r to perfectly offset their biases and then meet this goal. Figure 9 illustrates the choice of r in this case, relying on the simplifying assumption mentioned above. Interestingly, the choice of r does not depend on whether individuals treat their reference-dependent payoffs $v(x|r)$ as normative, what we might call their "inner π." A perfectly sophisticated goal-setter never incurs a loss, so whether loss aversion is normative becomes irrelevant for them.

How should we view welfare in a model of chosen reference points? If we regard the reference point as

[34] Without this simplifying assumption, the individual would choose a reference point in order to induce the highest consumption of x possible without incurring a loss. In other words, they would set r such that $u_i'(r) + \Lambda_i = p$, so that $r = x_i^L(p)$.

an option chosen optimally by the individual, policies that aim to reduce r will no longer improve welfare. In this case, due to the envelope theorem, inducing a marginal increase or decrease in r has no first-order welfare effects and a second-order loss. For the effects of price changes, note that the individual always ends up in the R case, where a marginal change in price has no effect on behavior, only a first-order negative direct effect on welfare.

Building on this logic, we can infer what happens if we relax the assumption of full sophistication about biases. An individual who underestimates their bias will set a goal r that is sub-optimally low. In this case, inducing the individual to set a higher reference point would have a first-order, positive impact on welfare. Similarly, an individual who over-estimates their future bias will set an over-optimistic goal r, and could be made better off by setting a lower reference point, especially when $\pi = 1$ and the losses incurred by failing to meet one's goal generate a negative payoff. Whether policymakers would have enough information to correct these types of mistakes in the choice of goal reference points is not clear (Glaeser, 2006).

In other words, when individuals choose their own reference points, optimal policy questions turn on the optimality of individuals' choice of r, rather than depending solely on the optimality of the choice of x. If individuals choose their reference points fully optimally, the envelope theorem bites and changing reference points via policy has no first-order welfare effects. However, if individuals are not fully sophisticated about their biases, welfare effects depend on the direction of their mistakes in setting reference points. Finally, we note that one would obtain a similar characterization of welfare in models where individuals exert control over reference points for other reasons, such as models of anticipatory utility (Sarver, 2012).

4.4 Further Theoretical Extensions

In this section, we briefly discuss three further complications that could be added to our model and which may be useful in some applications. We defer fully characterizing welfare under these extensions to future work. They might be best explored in applied settings where empirical evidence and features of the environment can help discipline the structure imposed in extending the model.

Diminishing Sensitivity. As discussed in the context of our simplifying assumptions, we have so far ignored diminishing sensitivity. Compared to equation (2) or (26), diminishing sensitivity would require that $v'' < 0$ for $x > r$ and $v'' > 0$ for $x < r$. Adding it to the model would be straightforward, but evidence in favor of diminishing sensitivity, especially in decision-making under certainty, is limited. Moreover, with diminishing sensitivity, it becomes difficult to empirically distinguish curvature of intrinsic utility over x, which we denoted $u''(x)$, and curvature over reference-dependent utility $v''(x)$ in the gain domain.

Allowing for diminishing sensitivity is unlikely to change much of the intuition about the direct and behavioral effects of a change in the reference point or in prices. For instance, the presence and sign of the marginal internality will be unaffected, and the result that lower reference points improve welfare will obtain under diminishing sensitivity. However, with curvature in v'', the various demand curves in Figures 2 through 4 are no longer parallel, which implies that the size of welfare effects might be different. Insofar as we only consider choices nearby the reference point, such differences may be negligible as in this region the piece-wise linear formulation we use would remain approximately accurate.

Risk and Uncertainty. In this paper, we consider the case of reference dependence under certainty. Reference dependence in choice under uncertainty is the subject of a rich theoretical and experimental literature.

There are two challenges in adapting our welfare analysis to a model with uncertainty. First, the question of how to measure welfare in a model with uncertainty can be tricky. Many applications of welfare economics under uncertainty use certainty equivalent welfare metrics (Einav et al., 2010). With reference dependence, at least under $\pi = 1$, certainty equivalence becomes a poor welfare metric due to state dependence in the utility function, so a generalization of equivalent variation that allows for uncertainty and state dependence would be required.

Second, as discussed in Section 2.3.1, reference point formation is more controversially debated for the stochastic case. The mixed empirical evidence on the origins of reference points and the wide variety of proposed models makes it more difficult to choose a formulation of reference-dependent preferences and it may not be clear which policy changes induce a shift in reference points. A prominent line of research proposes that reference points are based on beliefs or expectations (e.g. Kőszegi and Rabin, 2006, 2007). In this case, changing expectations would change the reference point, but this can also influence behavior and welfare in other ways.

Narrow Bracketing. An important component of prospect theory as laid out by Kahneman and Tversky (1979) is the bracketing of payoffs, which is closely related to the concept of mental accounting (Thaler, 1985). What we consider is essentially "broad bracketing", where agents evaluate total consumption of good x relative to a reference point over total x. Correspondingly, there is a single reference point for lifetime leisure in our empirical application.

In other contexts, individuals seem to adopt *narrow bracketing*, where options are partitioned into component parts and evaluated separately, with a reference point for each. For example, narrow bracketing appears to be an important driver of the "disposition effect", where individuals receive a payoff based on whether a specific stock has gained or lost value (relative to a zero reference point) instead of receiving payoffs based only on the value of their entire portfolio (Barberis and Xiong, 2012; Imas, 2016). Modelling welfare in this type of situation would require a normative judgment over not only whether reference-dependent payoffs deserve normative weight, but also whether narrow bracketing is a mistake (see e.g. Koch and Nafziger, 2016).

5 Conclusion

In this paper, we provide a first attempt at studying the welfare economics of reference dependence. Our most important theoretical result is the characterization of the direct and behavioral welfare effects of changes in reference points and prices, and how normative judgments shape these effects. For a change in the reference point, the judgment about whether reference dependence reflects a preference or a bias governs whether direct or behavioral effects matter for welfare. For a price change, behavioral effects are only present when reference dependence is judged a bias. Examining a simple model of loss aversion over a single good, we can sign these welfare effects. In particular, we find that decreasing a reference point generates a Pareto Improvement regardless of normative judgments. We also obtain sufficient statistics representations of the welfare effects of reforms that can be applied across many contexts.

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 show that increasing the Normal Retirement Age is welfare-improving when it serves as a reference point in the labor supply/leisure

dimension. The welfare effects of subsidies for later retirement, on the other hand, are ambiguous and depend on normative judgments about reference dependence. The magnitude of the estimated welfare effects highlights that setting the Normal Retirement Age or the Delayed Retirement Credit is a high-stakes policy decision, much more so than a standard model without reference dependence would suggest. Finally, we use the empirical application to illustrate that our sufficient statistics formulas provide good approximations of welfare effects, in particular for small reforms.

Meanwhile, our analysis of extensions of the simple model raises a number of policy-relevant caveats that we hope will motivate future work. The magnitude of key welfare effects can depend on the exact form of reference-dependent payoffs and the extent to which individuals control their reference points. Both of these questions are relatively unsettled in the positive theory of reference dependence. For example, our analysis of multi-dimensional reference dependence reveals that strong reference dependence over consumption could overturn the baseline result on the positive welfare effects of increasing the Normal Retirement Age. Thus, it would be very valuable to gather more evidence on the dimensions in which reference dependence appears and to empirically distinguish between forms of reference dependence. Furthermore, taking our baseline results at face value would suggest that lowering reference points to extreme degrees, or, in the empirical application, raising the Normal Retirement Age to an extremely high level, is welfare-improving. One way to discipline this extreme result would be to suppose that attempting to use policy to shift reference points to extreme levels will cause individuals to stop using the policy as their reference point. Empirically studying the limits to influencing reference points via policy may be a promising avenue for future research.

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 many domains. Nevertheless, policy interest in behavioral economics has grown rapidly in recent years, and quantifying welfare effects is important in informing policy debates. Embracing normative ambiguity can be a productive approach because it allows us to separate questions that can be empirically analyzed, such as the influence of a change in reference point or prices on behavior, from normative judgments.

References

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference Points and Effort Provision. *American Economic Review*, 101(2):470–92.

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.

Barberis, N. and Xiong, W. (2012). Realization Utility. *Journal of Financial Economics*, 104(2):251–71.

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.

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

Clark, A. E., Senik, C., and Yamada, K. (2017). When Experienced and Decision Utility Concur: The Case of Income Comparisons. *Journal of Behavioral and Experimental Economics*, 70:1–9.

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. (2021). The Effect of Changes in Social Security's Delayed Retirement Credit: Evidence from Administrative Data. NBER working paper no. 28919.

Einav, L., Finkelstein, A., and Cullen, M. R. (2010). Estimating Welfare in Insurance Markets Using Variation in Prices. *Quarterly Journal of Economics*, 125(3):877–921.

Ericson, K. M. and Fuster, A. (2011). Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments. *Quarterly Journal of Economics*, 126(4):1879–1907.

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.

Glaeser, E. L. (2006). Paternalism and Psychology. *University of Chicago Law Review*, 73(1):133–56.

Gneezy, U., Goette, L., Sprenger, C., and Zimmermann, F. (2017). The Limits of Expectations-Based Reference Dependence. *Journal of the European Economic Association*, 15(4):861–76.

Goette, L., Harms, A., and Sprenger, C. (2021). Randomizing Endowments: An Experimental Study of Rational Expectations and Reference-Dependent Preferences. *American Economic Journal: Microeconomics*, forthcoming.

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.

Hardie, B. G., Johnson, E. J., and Fader, P. S. (1993). Modeling loss aversion and reference dependence effects on brand choice. *Marketing Science*, 12(4):378–94.

Heath, C., Larrick, R. P., and Wu, G. (1999). Goals as Reference Points. *Cognitive psychology*, 38(1):79–109.

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.

Imas, A. (2016). The Realization Effect: Risk-taking after Realized Versus Paper Losses. *American Economic Review*, 106(8):2086–2109.

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.

Koch, A. K. and Nafziger, J. (2011). Self-Regulation through Goal Setting. *Scandinavian Journal of Economics*, 113(1):212–227.

Koch, A. K. and Nafziger, J. (2016). Goals and Bracketing under Mental Accounting. *Journal of Economic Theory*, 162:305–351.

Kolsrud, J., Landais, C., Reck, D., and Spinnewijn, J. (2021). Retirement Consumption and Pension Design. Working paper.

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

Kőszegi, B. and Rabin, M. (2007). Reference-Dependent Risk Attitudes. *American Economic Review*, 97(4):1047–73.

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

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.

Manoli, D. S. and Weber, A. (2018). The Effects of the Early Retirement Age on Retirement Decisions. Working paper.

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.

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 (2019). Pensions at a Glance 2019. 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.

Rosch, E. (1975). Cognitive Reference Points. *Cognitive Psychology*, 7(4):532–47.

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.

Sarver, T. (2012). Optimal Reference Points and Anticipation. Working paper.

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.

Thaler, R. (1985). Mental Accounting and Consumer Choice. *Marketing Science*, 4(3):199–214.

Tversky, A. and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, 211(4481):453–58.

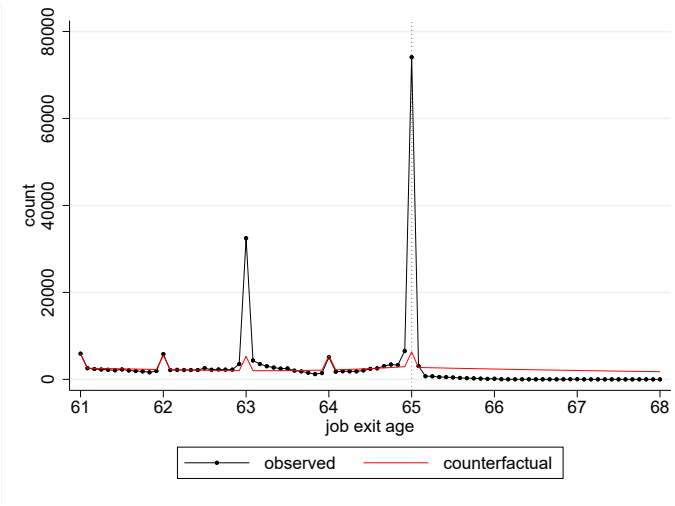
Tversky, A. and Kahneman, D. (1991). Loss Aversion in Riskless Choice: A Reference-Dependent Model. *Quarterly Journal of Economics*, 106(4):1039–61.

Appendix (For Online Publication)

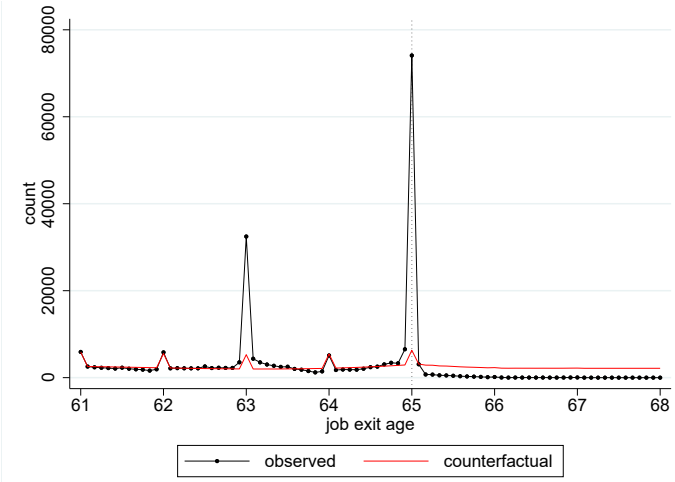
A Additional Figures and Tables

FIGURE A1: COUNTERFACTUAL RETIREMENT AGE DISTRIBUTION

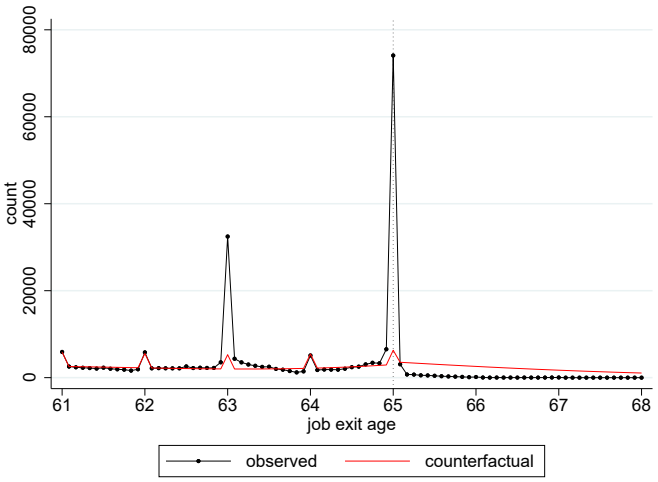
(a) Baseline: Pareto Tail



(b) Alternative 1: Uniform Tail



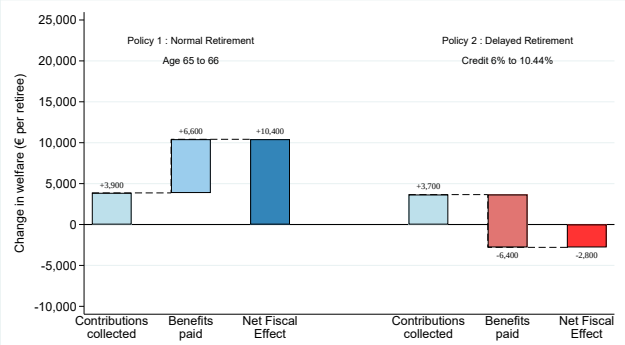
(c) Alternative 2: Lognormal Tail



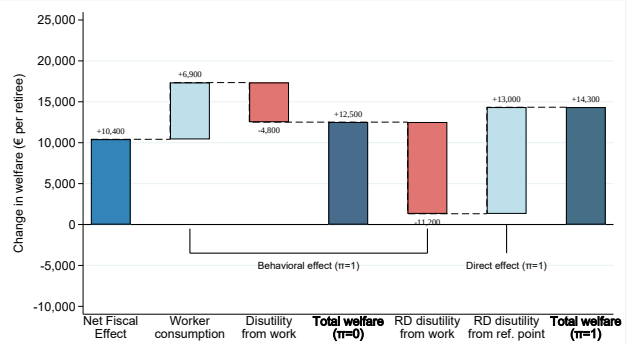
Notes: The figure shows counterfactual retirement distributions under different assumptions about the shape of the upper tail of the distribution. In all panels, the counterfactual distribution up until the NRA (age 65) is obtained by fitting a seventh-order polynomial to the observed retirement age 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: WELFARE EFFECTS OF PENSION REFORMS

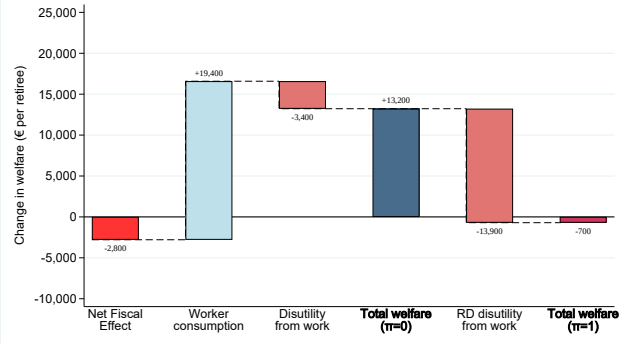
(a) Fiscal Effects



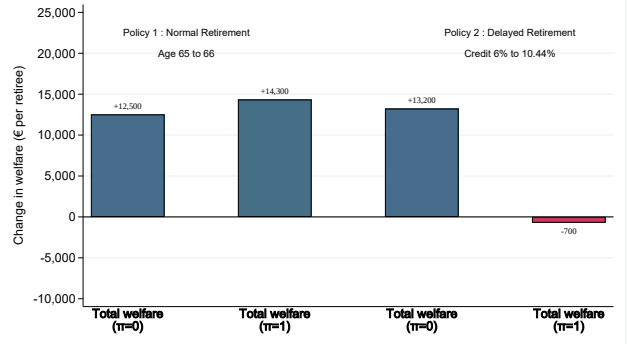
(b) Welfare Effects of NRA Increase



(c) Welfare Effects of DRC Increase

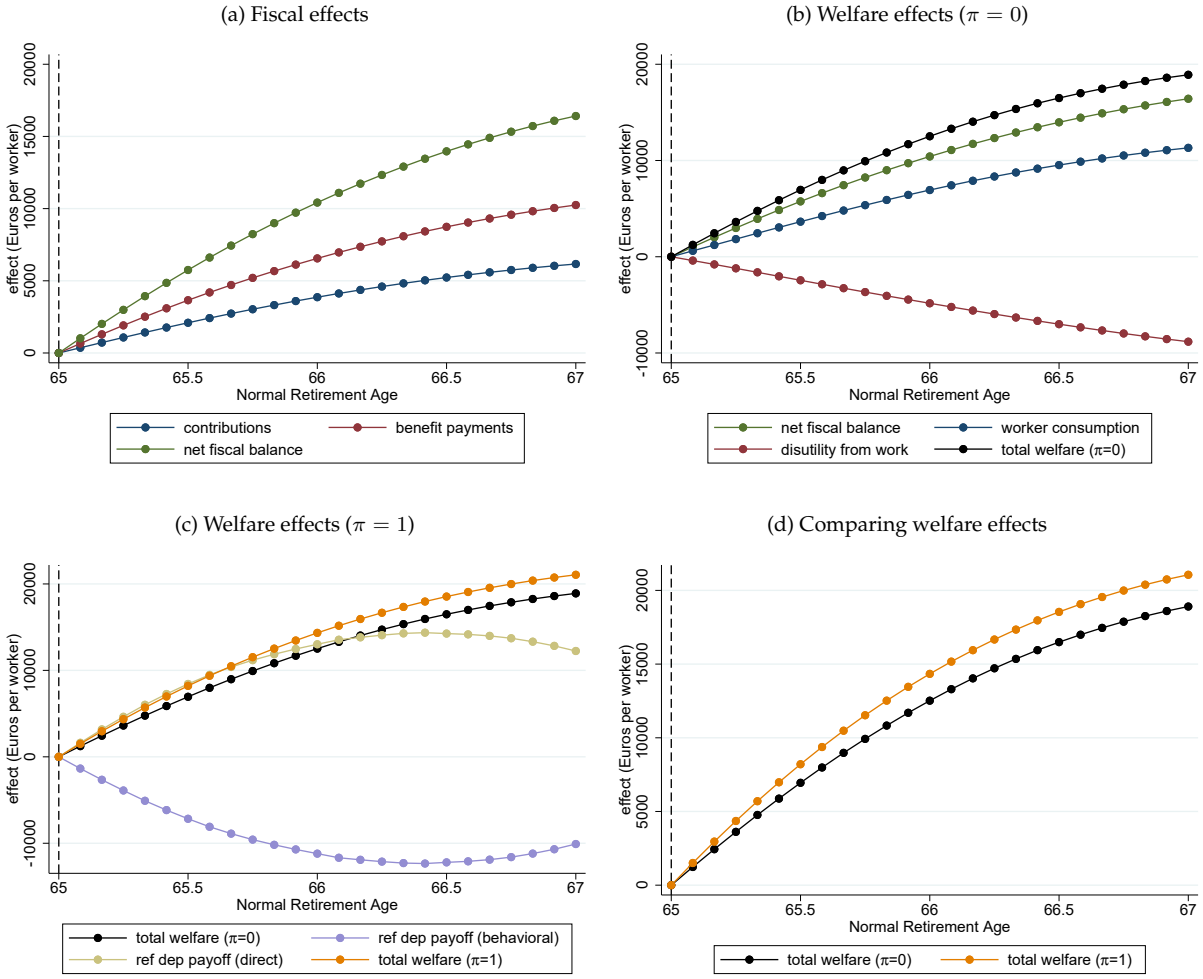


(d) Comparing Welfare Effects



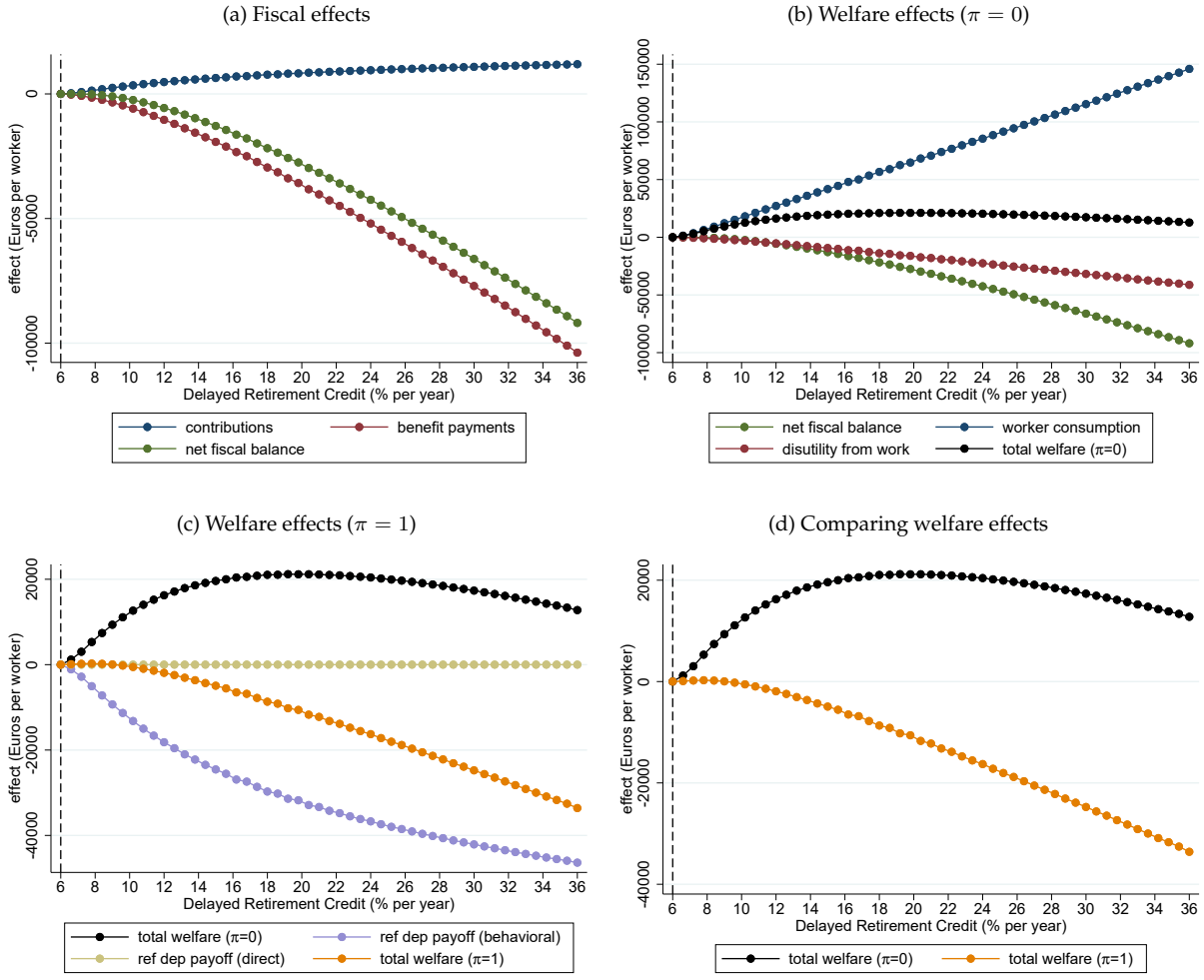
Notes: The figure visually illustrates the aggregation of components of the welfare effects of pension reforms shown in Table 1. All effects are calculated among workers retiring at age 65 and above, and are in Euros per worker, in terms of net present value at age 65.

FIGURE A3: INCREASING THE NORMAL RETIREMENT AGE – EXTENDED SIMULATIONS



Notes: The figure shows simulated fiscal and welfare effects of pension reforms that increase the NRA to ages between 65 and 67 in monthly 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. The signs the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare.

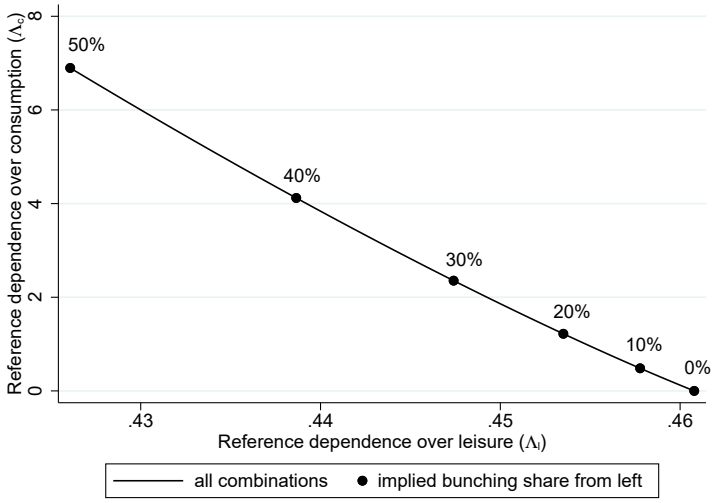
FIGURE A4: INCREASING THE DELAYED RETIREMENT CREDIT – EXTENDED SIMULATIONS



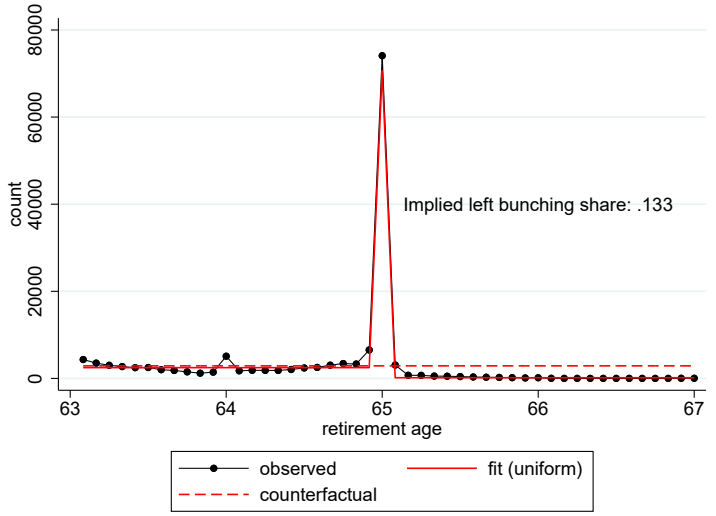
Notes: The figure shows simulated fiscal and welfare effects of pension reforms that increase the Delayed Retirement Credit to values between 6% and 36% per year in half-percentage point increments. Simulations are conducted for birth cohort 1946. All effects are calculated among workers retiring at age 65 and above, and are in Euros per worker, in terms of net present value at age 65. The signs the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare.

FIGURE A5: TWO-DIMENSIONAL REFERENCE DEPENDENCE

(a) Bunching at the NRA Identifies Combinations of Λ_l, Λ_c



(b) Preferred Bunching Share Estimate



Notes: Panel (a) of the figure shows a simulated range of combinations of reference dependence over leisure Λ_l and reference dependence over consumption Λ_c. Parameter combinations are obtained by gradually moving the left bunching share from zero to 50% as described in Appendix F.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 Λ_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 Λ	0.461 (0.000)
Retirement age elasticity ε	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 (22), 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	+3,815	+3,570
Benefits paid +	6,556	–6,100
Net fiscal effect	+10,371	–2,529
Worker consumption	+6,746	+18,743
Disutility from work	–4,684	–3,162
Worker welfare ($\pi = 0$)	+2,062	+15,582
Ref. dep. disutility from work	–11,139	–13,597
Ref. dep. utility from ref. point	+13,110	0
Worker welfare ($\pi = 1$)	+4,033	+1,984
Total welfare ($\pi = 0$)	+12,434	+13,052
Total welfare ($\pi = 1$)	+14,405	–545
	Panel B: Lognormal Tail	
	Policy 1: Normal Retirement Age to 66	Policy 2: Delayed Retirement Credit to 11.28%
Contributions collected	+3,733	+3,496
Benefits paid	+6,120	–7,074
Net fiscal effect	+9,853	–3,577
Worker consumption	+6,935	+19,501
Disutility from work	–5,281	–3,805
Worker welfare ($\pi = 0$)	+1,655	+15,696
Ref. dep. disutility from work	–9,951	–13,305
Ref. dep. utility from ref. point	+11,325	0
Worker welfare ($\pi = 1$)	+3,028	+2,391
Total welfare ($\pi = 0$)	+11,508	+12,119
Total welfare ($\pi = 1$)	+12,881	–1,187

Notes: The table shows results from pension reform simulations as in 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: DECOMPOSING WELFARE EFFECTS OF PENSION REFORMS BY GROUP

	(1)	(2)	(3)	(4)	(5)	(6)
	R group			L group		
	RR	RG	RL	LR	LG	LL
	Panel A: Policy 1 – Normal Retirement Age to 66					
Share of group (% of all workers)	35.8%	38.9%	0.0%	16.9%	1.0%	7.4%
Contributions collected	+7,280	+2,280		+2,107	+1,402	0
Benefits paid	+10,704	+3,269		+6,661	+2,569	+4,003
Net fiscal effect	+17,984	+5,550		+8,768	+3,971	+4,003
Worker consumption	+14,056	+4,872		+1,479	+4,813	-4,003
Disutility from work	–6,533	–5,305		–2,277	–4,582	0
Worker welfare ($\pi = 0$)	+7,523	–433		-799	+231	–4,003
Ref. dep. utility from ref. point	+27,517	0		+15,185	0	+8,072
Ref. dep. disutility from work	–27,517	0		–8,004	+1,126	0
Worker welfare ($\pi = 1$)	+7,523	–433		+6,382	+1,357	+4,069
Total welfare ($\pi = 0$)	+25,508	+5,117		+7,970	+4,202	0
Total welfare ($\pi = 1$)	+25,508	+5,117		+15,151	+5,328	+8,072
Contribution to overall welfare effect ($\pi = 0$)	+9,141	+1,991	0	+1,344	+40	0
Contribution to overall welfare effect ($\pi = 1$)	+9,141	+1,991	0	+2,554	+51	+599
	Panel B: Policy 2 – Delayed Retirement Credit to 10.44%					
Share of group (% of all workers)	29.0%	0.0%	45.8%	0.0%	0.0%	25.2%
Contributions collected	0		+6,152			+3,340
Benefits paid	0		–9,168			–8,924
Net fiscal effect	0		–3,016			–5,584
Worker consumption	0		+30,231			+21,952
Disutility from work	0		–4,818			–4,577
Worker welfare ($\pi = 0$)	0		+25,413			+17,375
Ref. dep. utility from ref. point	0		0			0
Ref. dep. disutility from work	0		–23,341			–12,871
Worker welfare ($\pi = 1$)	0		+2,073			+4,503
Total welfare ($\pi = 0$)	0		+22,397			+11,791
Total welfare ($\pi = 1$)	0		–944			–1,081
Contribution to overall welfare effect ($\pi = 0$)	0	0	+10,249	0	0	+2,976
Contribution to overall welfare effect ($\pi = 1$)	0	0	–432	0	0	–273

Notes: The table shows results from simulations of pension reforms increasing the NRA to 66 (Panel A) and increasing the Delayed Retirement Credit to 10.44% (Panel B). Effects are shown for following groups of workers: those retiring at the NRA before and after the reform (RR), those retiring at the NRA before and in the gain domain below the NRA after (RG), those retiring above the NRA before and at the NRA after (LR), those retiring above the NRA before and below the NRA after (LG), and those retiring above the NRA before and after (LL). Effects for the GG group who retire below the NRA before and after the reform are not shown, as these workers are unaffected by the reform. 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. The signs the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare. "Contribution to overall welfare effect" denotes the portion of the total welfare effect from Table 1 that can be attributed to the welfare effect on the respective group in each column.

TABLE A4: PARAMETERS FOR SUFFICIENT STATISTICS CALCULATIONS

Parameter	Value
Reference dependence Λ	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 $P(i \in L)$	0.252
Fraction in R group $P(i \in R)$	0.748
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 3.5.

B Theoretical Extensions

This appendix presents the theoretical extensions described in Sections 4.1 and 4.2 more formally.

B.1 Two-Dimensional Reference Dependence

B.1.1 Setup

Behavior. We continue to assume that the individual has quasi-linear preferences over goods x and y as before. We introduce a reference point s for good y and model behavior as follows:

$$\max_{x,y} u_i(x) + y + v_i(x|r) + w_i(y|s) \quad \text{subject to } px + y = z_i \tag{28}$$

The reference-dependent term in the x dimension has the same form as equation (2), and the new reference-dependent term in the y dimension, $w_i(y|s)$, is given by:

$$w_i(y|s) = \begin{cases} 0, & y > s \\ \Gamma_i(y - s) & y \leq s. \end{cases} \tag{29}$$

We continue not to include potential distortions in the gain domain for now, but keep this for the next Section B.2.[35] For simplicity, we restrict our attention to the situation where the two-dimensional reference point (r, s) is on the budget constraint: $pr + s = z_i$. With this restriction $x > r \iff y < s$: the gain domain for good x and the loss domain for y perfectly coincide. As in equations (3) and (4), the first-order conditions for $x > r$ and $x < r$ are now given by

$$\frac{u_i'(x_i^G(p))}{1 + \Gamma_i} = p, \tag{30}$$

$$u_i'(x_i^L(p)) + \Lambda_i = p \tag{31}$$

Behavior is given by equation (5) with the new potential demand curves $x^L(p)$ and $x^G(p)$ implied by equations (30) and (31).

Welfare. As in equation (6), we can specify welfare given a normative judgment $\pi \in \{0, 1\}$ as follows:[36]

$$U_i^*(x, y) = u_i(x) + y + \pi[v_i(x|r) + w_i(y|s)]. \tag{32}$$

Lemma 2. ***The Marginal Internality in the Two-Dimensional Model.*** *Let m_i be the derivative of $U_i^*(x, z_i - px)$ from equation (32) with respect to x, evaluated at $x_i(p, r)$.*[37]

L2.1. *If $x_i(p, r) > r$, $m_i(p, r; pi) = (1 - \pi)\Gamma_i p$.*

L2.2. *If $x_i(p, r) < r$, $m_i(p, r; pi) = -(1 - \pi)\Lambda_i \equiv m_i^L$*

[35] Including an η_i-like term for both dimensions is a straightforward extension, but such a model would be very difficult to identify empirically without some methodological progress. As we discuss in Section 4.2, identifying η_i for a single dimension is already difficult; separately identifying such a parameter for two different dimensions would be even more challenging.

[36] Note that we use the same π in both dimensions. Relaxing this assumption is straightforward, but it is difficult to imagine why one might judge that loss aversion is normative in one dimension and not in another.

[37] In this model, the marginal welfare effect of a change in the endowment z_i is no longer unity, in particular when $x < r$, $\partial w / \partial z_i = 1 + \pi\Gamma_i$. As such, the marginal internalities derived here might not be money metric. It turns out, however, that this does not matter for $\pi \in \{0, 1\}$ because the marginal internality is zero when $\pi = 1$, so that scaling it by $1 + \Gamma$ when $x < r$ is inconsequential.

L2.3. *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 $-\Lambda_i \leq m_i \leq \Gamma_i p$*

Note that when $\pi = 0$, the marginal internality is positive in the gain domain, unlike before, while it continues to be negative in the loss domain. The individual under-consumes x to reduce losses in the y domain when $x > r$, and over-consumes x to reduce losses in the x domain when $x < r$. Figure 5 illustrates demand in this model. We plot demand in the gain and loss domain, x^L and x^G according to equations (30) and (31). The main difference to the simple model is that demand without reference dependence, pinned down by $p = u'(x)$, now falls *between* observed demand in the gain and loss domains. As $p = u'(x)$ describes welfare-maximizing demand when $\pi = 0$, the vertical distance between this demand curve and observed demand equals the marginal internality, which we know from Lemma 2 is now positive in the gain domain and negative in the loss domain. When $\pi = 1$, observed demand is welfare-maximizing and there is no marginal internality.

B.1.2 Main Social Welfare Effects

Proposition 4. ***First-Order Social Welfare Effects in the Two-Dimensional Model.*** *Starting from any given price and an initial reference point, define groups G, L and R based on $x_i(p,r)$.*

P4.1. *The effect of a small change in the reference point of Δr on social welfare in this model is approximately*

$$\begin{aligned} \Delta W \approx & \Delta r \pi \left\{ E\left[\Gamma_i p \mid i \in G\right] P\left[i \in G\right] - E\left[\Lambda_i \mid i \in L\right] P\left[i \in L\right] \right\} \\ & - \Delta r E\left[p - u_i'(r) \mid i \in R\right] P\left[i \in R\right]. \end{aligned} \tag{33}$$

P4.2. *The effect of a small change in price, Δp, on social welfare in this model is approximately*[38]

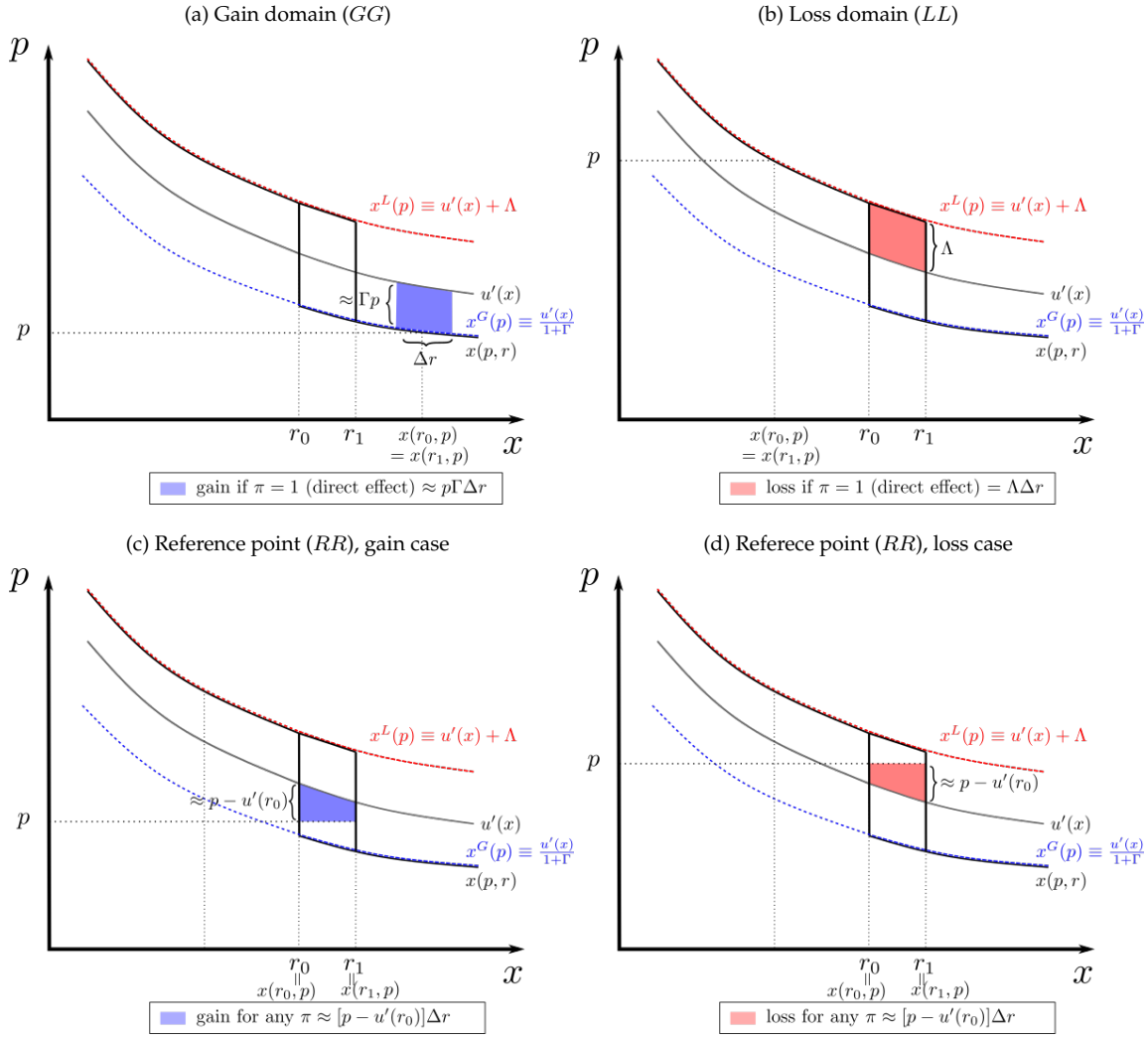
$$\begin{aligned} \Delta W \approx & \Delta p \left\{ E\left[(1-\pi)\Gamma_i p \frac{\partial x_i^G}{\partial p} \;\middle|\; i \in G \right] P[i \in G] - E\left[(1-\pi)\Lambda_i \frac{\partial x_i^L}{\partial p} \;\middle|\; i \in L \right] P[i \in L] \right\} \\ & - E[x_i(p_0, r_0)]\Delta p, \end{aligned} \tag{34}$$

where $\frac{\partial x_i^L}{\partial p}$ and $\frac{\partial x_i^G}{\partial p}$ are evaluated at (p_0, r_0).

Proposition 4 characterizes the main social welfare effects in the two-dimensional model. Note that the proposition nests Propositions 1.2 and 2.2 when $\Gamma_i = 0$ for all i, i.e. when there is no loss aversion over y. Proposition 4.1 considers changes in the reference point, which are also depicted in Figure B1. When the reference point is a point on the budget constraint $s = z - pr$, decreasing r must increase s, which is the policy change we consider in equation (33). The main change relative to the simple model is therefore that we observe an additional, positive direct effect of increasing the reference point for individuals consuming in the gain domain for x (the loss domain over y). Additionally, we should now expect that there are some individuals in the R group for whom $p > u_i'(r)$ and some for whom $p < u_i'(r)$, so the third term in equation (33) is ambiguously signed.

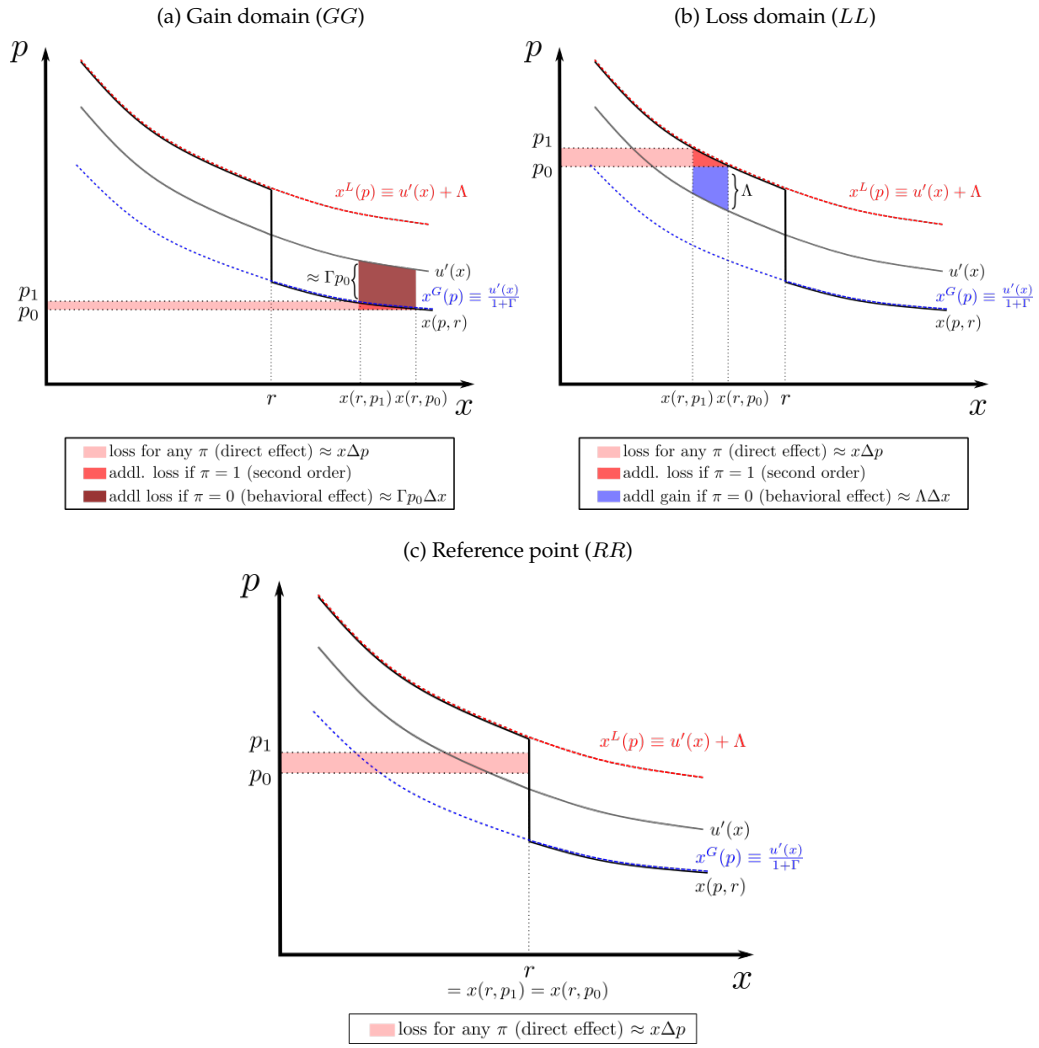
[38] In this approximation, we do not allow the reference point over y, s, to change when the price changes. Doing so would introduce a direct welfare effect of a change in that reference point. In other words, we require that the reference point is a point on the budget constraint at status quo prices, but we do not allow changing the price to have a mechanical effect on the reference point.

FIGURE B1: WELFARE EFFECTS OF CHANGING THE REFERENCE POINT UNDER TWO-DIMENSIONAL REFERENCE DEPENDENCE



Notes: The figure illustrates the welfare effects of changing the reference point for individuals in the domains indicated by the panel titles. Unlike in Figure 3, we assume reference dependence over both x and the background good y. For simplicity, we include only those cases that are relevant for first-order social welfare. We also depict the RR case in two situations, where individual experiences either a gain or loss. We denote observed demand in black, marginal utility $u'(x)$ in grey, and gain and loss domain demand in blue and red, respectively, as in Figure 5. Welfare gains are depicted in blue shaded regions and losses in red shaded regions. Note that because the size of the direct effect on the G group in equation (33) depends on the price, we can no longer illustrate the direct effect of a change in the reference point Panel (a) like in Figure 3, and thus it is depicted slightly differently.

FIGURE B2: WELFARE EFFECTS OF PRICE CHANGES UNDER TWO-DIMENSIONAL REFERENCE DEPENDENCE



Notes: The figure illustrates the welfare effects changing prices for individuals in the domains indicated by the panel titles. Unlike in Figure 3, we assume reference dependence over both x and the background good y. For simplicity, we include only those cases that are relevant for first-order social welfare. We denote observed demand in black, marginal utility $u'(x)$ in grey, and gain and loss domain demand in blue and red, respectively, as in Figure 5. Welfare gains are depicted in blue shaded regions and losses in red shaded regions. The main difference to Figure 4 is the presence of an additional loss due to the behavioral effect in the gain domain, depicted in dark red in panel (a).

In the two-dimensional model, decreasing the reference point along a single dimension – i.e. decreasing r or decreasing s in isolation – would be a robust Pareto improvement, but decreasing the reference point r *along the budget constraint* is not generally a robust social welfare improvement. We can infer from equation (33) that such a decrease in the reference point r would be welfare-improving only if (1) Λ_i is sufficiently large compared to $p\Gamma_i$ on average, (2) there are sufficiently many individuals in the loss domain compared to the gain domain, and (3) individuals consuming at the reference point tend to be over-consuming rather than under-consuming relative to marginal utility u'.[39]

Proposition 4.2 and Figure B2 consider price changes in the two-dimensional model. For a price change, we continue to observe a behavioral welfare effect valued according to the marginal internality m_i, and a direct effect. The main difference to the simple model is that the marginal internality is positive for $x > r$, which implies that decreasing consumption in response to a price change decreases welfare for $i \in G$.

We take a few key lessons away from our model extension allowing for two-dimensional reference dependence. Most importantly, we find that the main welfare effects of interest now depend on the relative strength of reference dependence in the two dimensions as well as the distribution of individuals across gain and loss domains. Thus, applying a two-dimensional model requires empirically differentiating and quantifying reference dependence in these dimensions. In the main text, we consider how two-dimensional reference dependence can be disciplined in our empirical application, where we argue that the empirical retirement age distribution is informative for this purpose. Alternatively, one could impose further theoretical restrictions on the strength of reference dependence in multiple dimensions. For instance, one intuitive and influential type of restriction is proposed by Kőszegi and Rabin (2006), but this imposes substantial structure on welfare effects and when this is empirically justified might be less clear.

We can imagine number of further extensions to analyze multidimensional reference dependence, building on the two-dimensional model. One could be to consider more than two dimensions, for instance to study multi-attribute reference dependence and brand choice (Hardie et al., 1993). Moreover, we assumed that the reference point must be on the budget constraint because we find this restriction simple and intuitive, but it may not be appropriate in all contexts. Finally, one could include the gain domain payoffs from the following section jointly with two-dimensional reference dependence. These extensions would all be theoretically feasible, but whether the resulting models are empirically useful will probably depend on the specific application.

B.2 Incorporating Gain Domain Payoffs

Next, we consider a formulation of $v(x|r)$ which includes gain domain payoffs, taking the model of reference dependence over riskless choice in Tversky and Kahneman (1991) at face value. The main difference between this formulation and our simple model is the presence of a new parameter η_i, which governs the strength of reference-dependent payoffs in the gain domain.[40] As in the simple model, we return to one-dimensional referenced dependence here.

[39] Technically, when $\pi = 0$, only condition (3) matters for welfare. Nevertheless, under typical regularity conditions on the distribution of primitives, conditions (1) and (2) will tend to be satisfied when (3) is satisfied.

[40] A similar friction is sometimes studied with or without reliance on loss aversion in the literature on the relative income hypothesis, (see e.g. Clark et al., 2017). Eliminating loss aversion (setting Λ equal to 0) would be a simple extension of the model we consider here.

B.2.1 Setup

The Tversky and Kahneman (1991) formulation of reference dependence payoffs is

$$v_i(x,r) = \begin{cases} \eta_i(x-r) & x > r \\ \eta_i\lambda_i(x-r) & x \leq r, \end{cases} \tag{35}$$

One can think of η_i as governing the importance of reference dependence overall, while λ_i governs the strength of loss aversion. Note that with reference dependence over just one good x, this model is behaviorally indistinguishable from the simple model in equation (2). We show this formally in Appendix D (see also Barseghyan et al., 2013) Relatedly, the presence and magnitude of the η_i preference parameter in the Tversky and Kahneman (1991) formulation is seldom if ever analyzed empirically.

The η_i parameter makes the individual consume more x by virtue of comparing their consumption to the reference point both in the gain and the loss domain. We first note that we can re-formulate the reference dependence from equation (35) as follows, to make it slightly more comparable to our earlier model:

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

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

As equation (35), this revised formulation is behaviorally equivalent to the simple model, with $\tilde{u}_i(x) = u_i(x) - \eta_i x$ and $\Lambda_i = \eta_i(\lambda_i - 1)$ (see Appendix D for a full proof). Thus, it is instructive to compare how adopting this formulation for welfare (instead of the simple model from Section 2) affects our normative results, holding observed behavior fixed.

In the main text, we scale the reference dependence payoff by a single normative judgment parameter π in the welfare function. With two reference dependence parameters η_i and λ_i, we can consider the question whether each frictions (reference dependence itself and loss aversion) reflects a bias or a normative preference separately. In particular, we use two normative parameters $\pi^{RD} \in \{0,1\}$ and $\pi^{LA} \in \{0,1\}$. The way we model normative judgment in the main text would imply that these are either both zero or both one: $\pi = 0 \implies \pi^{LA} = \pi^{RD} = 0$, and $\pi = 1 \implies \pi^{LA} = \pi^{RD} = 1$. We use the following specification for welfare:

$$\tilde{U}_i^*(x,y) = \tilde{u}_i(x) + y + \tilde{v}_i^*(x|r), \tag{38}$$

$$\tilde{v}_i^*(x|r) = \begin{cases} \pi^{RD}\eta_i(x-r), & x > r \\ [\pi^{RD}\eta_i + \pi^{LA}\Lambda_i](x-r), & x < r. \end{cases} \tag{39}$$

While we allow for the case where $\pi^{RD} = 1$ and $\pi^{LA} = 0$, we do not consider the situation where $\pi^{RD} = 0$ and $\pi^{LA} = 1$. The latter judgment would imply that reference dependence over gains and losses is a bias, but loss aversion is normative, which does not seem sensible. Finally, we denote indirect utility by $\tilde{w}_i(p,r)$, and utilitarian social welfare by $\tilde{W}(p,r)$.

Lemma 3. ***Marginal Internalities with Gain Domain Payoffs.*** *Let $\tilde{m}_i$ be the derivative of $\tilde{U}_i^*(x, z_i - px)$ with respect to x, evaluated at $x_i(p,r)$.*

L3.1. *If $x_i(p,r) > r$, $\tilde{m}_i = -(1-\pi^{RD})\eta_i \equiv \tilde{m}_i^G$.*

L3.2. *If $x_i(p,r) < r$, $\tilde{m}_i = -(1-\pi^{RD})\eta_i - (1-\pi^{LA})\Lambda_i$.*

L3.3. *If $x_i(p,r) = r$,*

- *$\tilde{m}_i$ is undefined when $\pi^{RD} = \pi^{LA} = 1$.*
- *Otherwise, $\tilde{m}_i = \tilde{u_i}'(r) + \pi^{RD}\eta_i - p$.*
- *Moreover, in the cases where $\tilde{m}_i$ is defined, $\tilde{m}_i^L \leq \tilde{m}_i \leq \tilde{m}_i^G$.*

Comparing Lemma 3 to the analogous Lemma 1 helps us understand how adding the friction embodied in η_i changes the model. When $\pi^{RD} = 1$, marginal internalities are all exactly the same as in the simple model, except that π is now denoted π^{LA} – recall that the models are behaviorally isomorphic if $\tilde{u}_i(x) + \eta_i = u_i(x)$. When $\pi^{RD} = 0$, however, reference dependence generates additional distortions, leading to more over-consumption of x than in the simple model. The revised marginal internalities and welfare-maximizing demand are plotted in Figure 8 When $\pi^{RD} = 0$, welfare maximizing demand corresponds to the line depicting $u_i'(x)$. When $\pi^{RD} = 1$, welfare-maximizing demand corresponds to either observed demand or $x_i^G(p)$, depending on whether $\pi^{LA} = 1$ or $\pi^{LA} = 0$.

B.2.2 Main Social Welfare Effects

Proposition 5. ***First-Order Welfare Effects with Gain Domain Payoffs.*** *Starting from an initial price and reference point, define G, R, and L groups. Let ΔW denote the change in social welfare from a reform that we obtain adopting the original formulation, as derived in Propositions 1.2 and 2.2, with $\pi = \pi^{LA}$.*

P5.1. *The first-order social welfare effect of a change in the reference point is approximately*

$$\Delta\tilde{W} \approx \Delta W - E[\pi^{RD}\eta_i \Delta r] - E[(1-\pi^{RD})\eta_i \Delta r \mid i \in R]P[i \in R]. \tag{40}$$

P5.2. *The first-order social welfare effect of a change in price is approximately*

$$\Delta\tilde{W} \approx \Delta W - E[(1-\pi^{RD})\eta_i \Delta x_i], \tag{41}$$

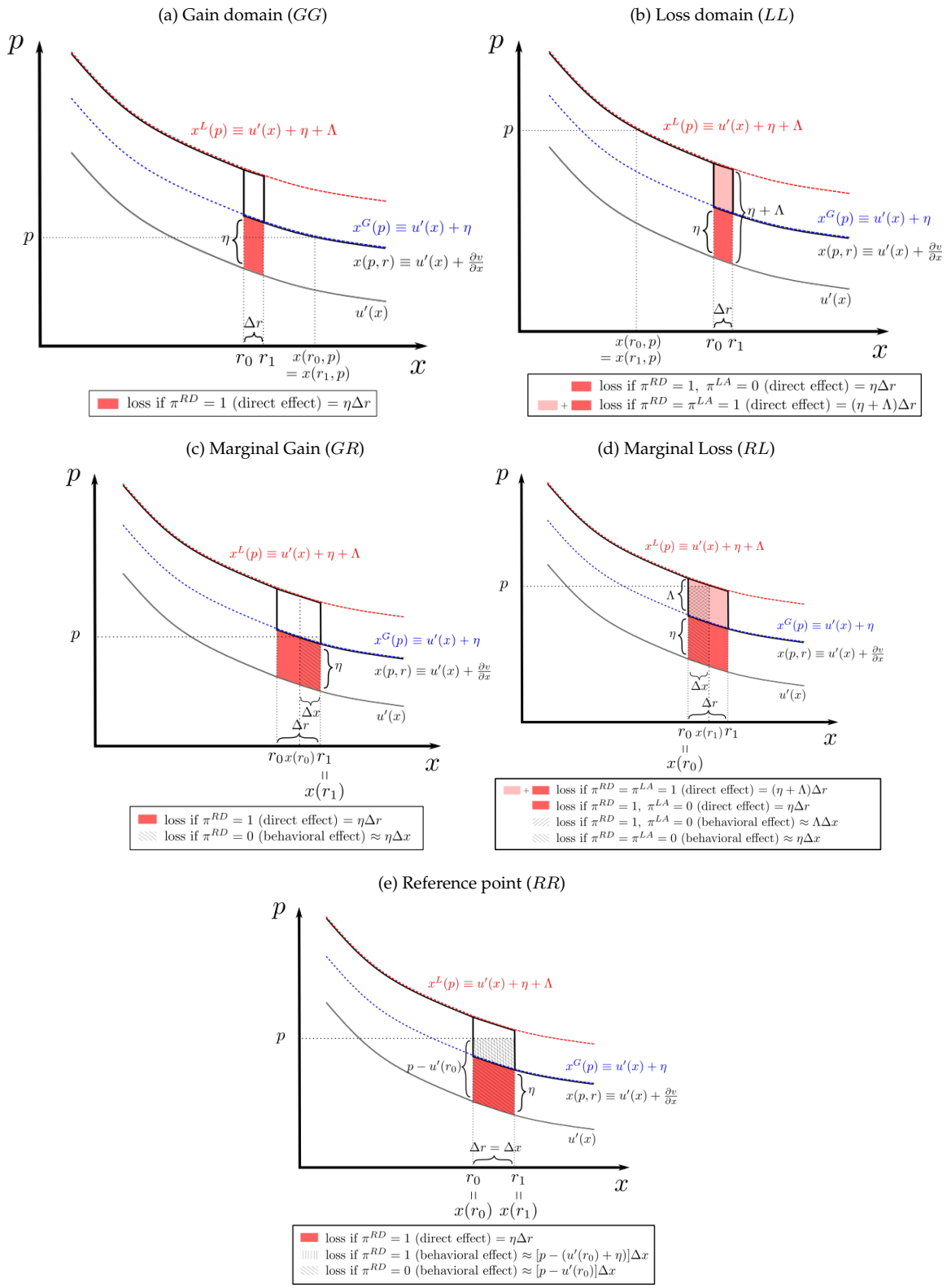
Proposition 5 shows that introducing the potential additional distortion to demand represented by the η_i parameter has two effects on welfare calculations. Appendix Figure B3 illustrates the modified welfare effects of a change in the reference point and Appendix Figure B4 show the modified welfare effects of a price change.

First, if we judge that reference dependence itself is normative, i.e. $\pi^{RD} = 1$, then the direct effect of a change in the reference point becomes larger and is present even in the gain domain. Apart from this extra direct effect, setting $\pi^{RD} = 1$ is equivalent to adopting our original formulation. Second, if we judge that the new friction is *not* normative, i.e. $\pi^{RD} = 0$, then negative internalities become larger, which means that any changes in demand caused by a change in reference points or prices has larger first-order welfare effects. Conditional on these two effects, the role of π^{LA} is essentially identical to the question of π in the simple model, which makes sense because both parameters represent the same underlying normative judgment about loss aversion. Note that both of the additional welfare effects strengthen the result that lowering the reference point improves individual and social welfare regardless of normative judgments.

That this formulation of reference dependence and the simple model without gain domain payoffs are behaviorally indistinguishable but carry differing implications for welfare raises interesting questions for future empirical research. In general, one might obtain specialized choice data, beyond observed demand

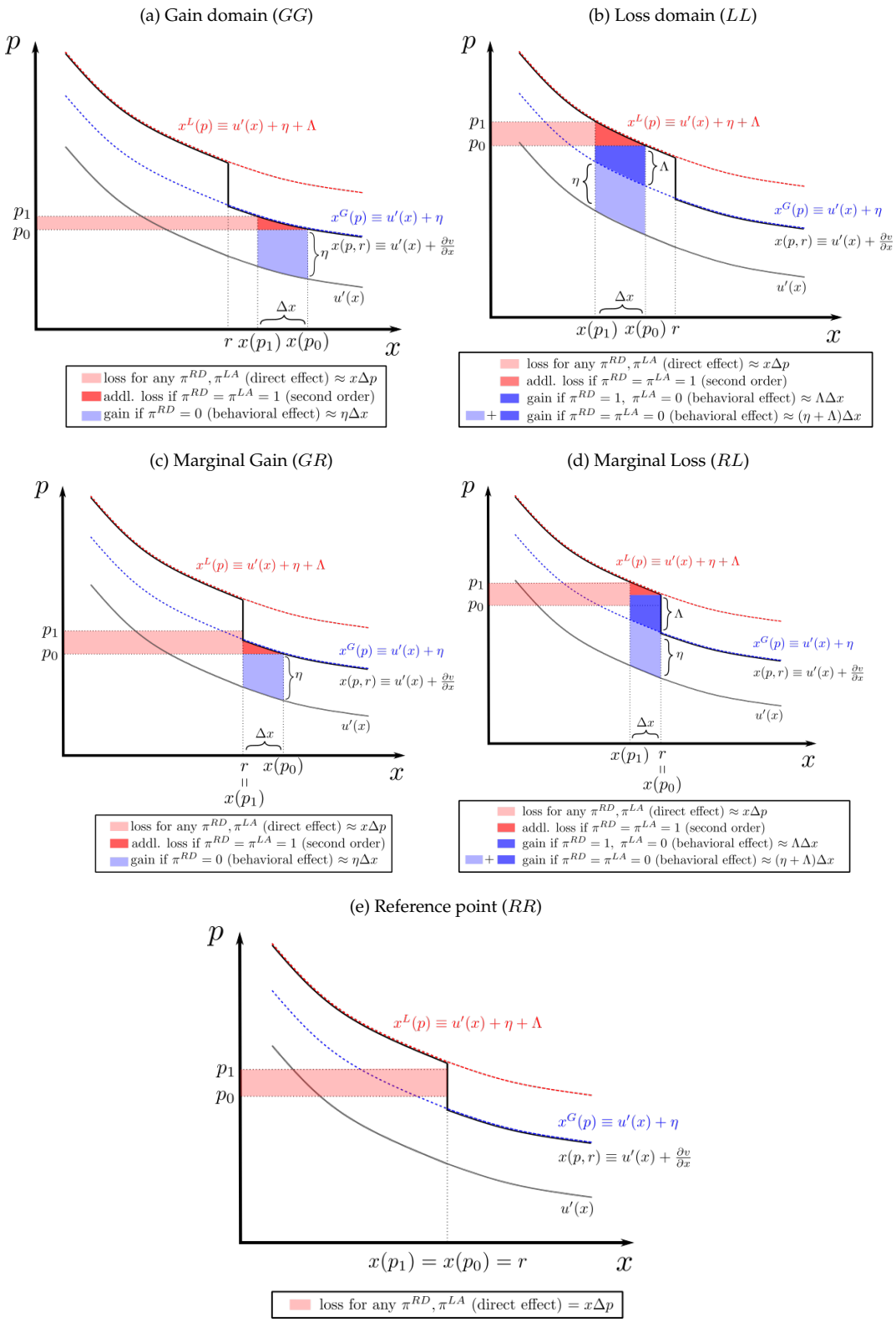
at one or more reference points, to distinguish between the models. For example, the two models yield different predictions for what would happen in a situation where the reference dependence were eliminated by some intervention (see e.g. Sokol-Hessner et al., 2009). Alternatively, we might solicit individuals' willingness to pay to change the reference point. The different formulations of reference dependence we consider make detailed predictions about how individuals would respond to either of these interventions. Additionally, with either of these novel designs, one could potentially identify the parameter η.

FIGURE B3: WELFARE EFFECTS OF CHANGING THE REFERENCE POINT WITH GAIN DOMAIN PAYOFFS



Notes: The figure plots the welfare effect of changing the reference point in the domains indicated by the panel titles. Unlike in Figure 3, , we adopt the formulation of reference dependence with gain domain payoffs. We denote observed demand in black, marginal utility $u'(x)$ in grey, and gain and loss domain demand in blue and red, respectively, as in Figure 8. Direct welfare effects are depicted in red shaded areas and behavioral welfare effects are shaded with diagonal hatching. The size of the direct effect depends on both π^{RD} and π^{LA}.

FIGURE B4: WELFARE EFFECTS OF PRICE CHANGES WITH GAIN DOMAIN PAYOFFS



Notes: The figure illustrates the welfare effects of changing the price in the domains indicated by the panel titles. Unlike in Figure 4, we adopt the formulation of reference dependence with gain domain payoffs. We denote observed demand in black, marginal utility $u'(x)$ in grey, and gain and loss domain demand in blue and red, respectively, as in Figure 8. The negative direct welfare effect of a price change is depicted in red shaded regions, and the positive behavioral effect is plotted in blue. The size of the behavioral welfare effect depends on both π^{RD} and π^{LA}.

C Proofs

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

C.1 Results in the Main Text

Lemma 1. The Marginal Internality.

L 1.1 $x_i(p,r) > r$,

$$m_i^G = \frac{\partial U_i(x, z_i - px)}{\partial x}|_{x=x^G} = u_i'(x^G) - p = 0 \quad \text{because } u_i'(x^G(p)) = p \quad (\text{FOC}^G)$$

L 1.2 $x_i(p,r) < r$,

$$m_i^L = \frac{\partial U_i(x, z_i - px)}{\partial x}|_{x=x^L} = u_i'(x^L) - p + \pi\Lambda_i = -\Lambda_i + \pi\Lambda_i \quad \text{because } u_i'(x^L(p)) + \Lambda_i = p \quad (\text{FOC}^L)$$
$$= -(1-\pi)\Lambda_i$$

L 1.3 $x = r$, $\pi = 1$,

The marginal internality is undefined in this case because of the kink in utility at $x = r$.

$x = r$, $\pi = 0$,

$$m_i = \frac{\partial U_i(x, z_i - px)}{\partial x}|_{x=r} = u_i'(r) - p$$

Proposition 1.1. First-Order Individual Welfare Effect of a Change in the Reference Point.

Recall that $w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r) + \pi v(x_i(p,r)|r)$. The proof proceeds case-wise by the groups defined in the main text.

Case 1: $i \in GG : w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r)$

By first-order Taylor series approximation :

$$\Delta w_i \equiv w_i(p,r_1) - w_i(p,r_0) \approx \frac{\partial w_i(p,r)}{\partial r}|_{r=r_0}\Delta r$$

And

$$\frac{\partial w_i(p,r)}{\partial r}|_{r=r_0} = \frac{\partial x_i^G(p)}{\partial r}(u_i'(x_i^G(p)) - p)$$

As the demand x^G does not depend on r (according to FOC^G), the derivative $\frac{\partial x_i^G(p)}{\partial r}$ is null. Then, $\Delta w_i = 0$.

Case 2: $i \in GR : w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r)$

By first-order Taylor series approximation :

$$\Delta w_i \equiv w_i(p,r_1) - w_i(p,r_0) \approx \frac{\partial w_i(p,r)}{\partial r}|_{r=r_0}\Delta r$$

And

$$\frac{\partial w_i(p,r)}{\partial r}|_{r=r_0} = \frac{\partial x_i(p,r)}{\partial r}(u_i'(x_i^G(p)) - p)$$

According to FOC^G, $u_i'(x_i^G) = p$. Then, $\Delta w_i = 0$.

Case 3: $i \in RR : w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r) + \pi\Lambda_i(x_i(p,r) - r)$

The kink at $x = r$ results in an indeterminate form for first-order approximation. However, independently

of the change in r, the individual remains with a consumption level equal to his reference point in both situations. Then, there is no direct effect here and we only consider the behavioral effect of the change.
Then, $\Delta w_i = \Delta r(u'(r_0) - p)$.
Case 4: $i \in RL : w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r) + \pi\Lambda_i(x_i(p,r) - r)$
By first-order Taylor series approximation :

$$-\Delta w \equiv w_i(p,r_0) - w_i(p,r_1) \approx -\frac{\partial w_i(p,r)}{\partial r}|_{r=r_1}\Delta r$$

$$\Rightarrow \quad \Delta w_i \approx \frac{\partial w_i(p,r)}{\partial r}|_{r=r_1}\Delta r$$

And

$$\begin{aligned}\frac{\partial w_i}{\partial r}|_{r=r_1} &= \frac{\partial x_i}{\partial r}(u_i'(x_i^L(p)) - p) + \frac{\partial x_i}{\partial r}\pi\Lambda_i - \pi\Lambda_i \\ &= -\frac{\partial x_i}{\partial r}\Lambda_i + \frac{\partial x_i}{\partial r}\pi\Lambda_i - \pi\Lambda_i \quad \text{by FOC}^L \\ &= -(1-\pi)\Lambda_i\frac{\partial x_i}{\partial r} - \pi\Lambda_i \\ \Rightarrow \Delta w_i &\approx -(1-\pi)\Lambda_i\Delta x_i - \pi\Lambda_i\Delta r \ \text{ as, by Taylor approximation } \Delta x_i \approx \frac{\partial x_i}{\partial r}\Delta r\end{aligned}$$

Case 5: $i \in LL : w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r) + \pi\Lambda_i(x_i(p,r) - r)$
By first-order Taylor series approximation :

$$-\Delta w_i \equiv w_i(p,r_0) - w_i(p,r_1) \approx -\frac{\partial w_i(p,r)}{\partial r}|_{r=r_1}\Delta r$$

$$\Rightarrow \quad \Delta w_i \approx \frac{\partial w_i(p,r)}{\partial r}|_{r=r_1}\Delta r$$

And

$$\frac{\partial w_i(p,r)}{\partial r}|_{r=r_1} = \frac{\partial x_i^L(p)}{\partial r}(u_i'(x_{il(p,r_1)}) - p) + \frac{\partial x_i^L(p)}{\partial r}\pi\Lambda_i - \pi\Lambda_i$$

As the demand x^L does not depend on r (according to FOC^L), the behavioral effect $\frac{\partial x_i^L(p)}{\partial r}$ is null. Then, $\Delta w_i = -\pi\Lambda_i\Delta r$.

Proposition 1.2. The First-Order Social Welfare Effect of a Change in the Reference Point.
We begin by taking the derivative of the social welfare function with respect to the reference point, toward a first-order approximation. To analyze social welfare formally, it is useful to introduce some more notation. Without loss of generality, we will model the heterogeneity in the population in terms of a triplet: $(\Lambda_i, \nu_i, \theta_i)$, where $u_i'(r) = \nu_i$ and θ_i is a vector of all other parameters governing heterogeneity in the utility function $u_i(x)$, allowing us to write $u_i(x) = u(x, \nu_i, \theta_i)$. We denote the cdf of Λ_i by F_Λ, we denote the cdf of $\nu_i|\Lambda$ by $F_\nu(\nu)$, and we denote the cdf of $\theta|(\Lambda,\nu)$ by $F_\theta(\theta)$. We can then express social welfare as

$$\begin{aligned}W = \int_0^{+\infty}\Bigg[\int_{-\infty}^{p-\Lambda}\int_\theta\left\{u(x_i^L,\nu,\theta) - px_i^L + \pi\Lambda_i(x_i^L - r)\right\}dF_\theta(\theta)dF_\nu(\nu) + \int_{p-\Lambda}^{p}\int_\theta\{u(r,\nu,\theta) - pr\}\,dF_\theta(\theta)dF_\nu(\nu) \\ + \int_p^{+\infty}\int_\theta\left\{u(x_i^G,\nu,\theta) - px_i\right\}dF_\theta(\theta)dF_\nu(\nu)\Bigg]\,dF_\Lambda(\Lambda)\end{aligned}$$

Taking a derivative with respect to r, we obtain

$$\begin{aligned}\frac{\partial W}{\partial r} &= \int_0^{+\infty}\left[\int_{-\infty}^{p-\Lambda}\int_\theta \{-\pi\Lambda_i\}\, dF_\theta(\theta)dF_\nu(\nu) + \int_{p-\Lambda}^{p}\int_\theta \{\nu - p\}\, dF_\theta(\theta)dF_\nu(\nu)\right.\\ &= \pi E[-\Lambda_i | i \in L]P[i \in L] + E[u_i'(r) - p | i \in R]P[i \in R].\end{aligned}$$

Note that with this parameterization, issues with the boundary between the G, R, and L groups are handled implicitly. In the proof of Proposition 2.2, we show more directly why the welfare effect for the marginal loss and marginal gain groups is second order – utility evolves continuously as an individual transitions from one group to another in response to a change in p or r, so we have a marginal change in utility for a marginal group.

Without the assumption of approximate uniformity of $u_i'(r) = \nu_i$ over the region defined by the R group, we would stop at the expression above for the first-order social welfare effect of a change in r. Under the assumption of an approximately uniform distribution of $\nu_i(r)$ in the R group, we can express the distribution of $u_i'(r)$ given $\Lambda_i = \Lambda$ as $f_u(u) = c_\Lambda$, noting that $P(i \in R|\Lambda) = c_\Lambda * \Lambda$. We plug this into the second term in the welfare effect above and evaluate the inner integral, obtaining:

$$\begin{aligned}E[p - u'(r) | i \in R] &= \int_0^{+\infty}\left[\int_{p-\Lambda}^{p}(\nu - p)\frac{c_\Lambda}{c_\Lambda \Lambda}d\nu\right] f_{\Lambda_i}(\Lambda_i)d\Lambda_i\\ &\approx \int_0^{+\infty}\left[\frac{\Lambda^2}{2}\frac{c_\Lambda}{c_\Lambda\Lambda}\right] f_{\Lambda_i}(\Lambda_i)d\Lambda_i\\ &= E\left[\frac{\Lambda_i}{2}\middle| i \in R\right]\end{aligned}$$

Plugging this into the expression for $\frac{\partial W}{\partial r}$ above and employing a first-order Taylor Series approximation – $\Delta W \approx \frac{\partial W}{\partial r}\Delta r$ – yields the result. Note that the expression used above evaluates the integral from $\nu = -(p - \Lambda)$ to p using the trapezoidal rule for approximating integrals. The trapezoidal rule is exact when ν is (conditionally) uniformly distributed, so that the function we are integrating in the inner integral is linear in ν.

Proposition 2.1. First-Order Individual Welfare Effect of a Change in Price.
Case 1: $i \in GG, GR$

$$\Delta w_i \equiv w_i(p_1, r) - w_i(p_0, r) \approx \frac{\partial w_i(p, r)}{\partial p}|_{p=p_0}\Delta p$$

And

$$\begin{aligned}\frac{\partial w_i(p,r)}{\partial p}|_{p=p_0} &= \frac{\partial x(p,r)}{\partial p}|_{p=p_0}.\frac{\partial U_i(x(p,r), z_i - px(p,r))}{\partial x}|_{p=p_0} - x(p_0, r)\\ &= \frac{\partial x(p,r)}{\partial p}|_{p=p_0} m_i(p_0, r) - x(p_0, r)\\ \Rightarrow \Delta w_i &\approx \frac{\partial x(p,r)}{\partial p}|_{p=p_0} m_i(p_0, r)\Delta p - x(p_0, r)\Delta p\end{aligned}$$

As

$$\Delta x \equiv x(p_1, r) - x(p_0, r) \approx \frac{\partial x(p,r)}{\partial p}|_{p=p_0}\Delta p$$

We finally obtain for $i \in GG, GR$

$$\Delta w \equiv w_i(p_1, r) - w_i(p_0, r) \approx m_i(p_0, r)\Delta x - x(p_0, r)\Delta p$$

Case 2: $i \in LL, RL$

$$-\Delta w_i \equiv w_i(p_0, r) - w_i(p_1, r) \approx \frac{\partial w_i(p, r)}{\partial p}|_{p=p_1}(p_0 - p_1)$$
$$\Rightarrow \Delta w_i \approx \frac{\partial w_i(p, r)}{\partial p}|_{p=p_1}\Delta p$$

And

$$\frac{\partial w_i(p, r)}{\partial p}|_{p=p_1} = \frac{\partial x(p, r)}{\partial p}|_{p=p_1}.\frac{\partial U_i(x(p, r), z - px(p, r))}{\partial x}|_{p=p_1} - x(p_1, r)$$
$$= \frac{\partial x(p, r)}{\partial p}|_{p=p_1} m_i(p_1, r) - x(p_1, r)$$
$$\Rightarrow \Delta w_i \approx \frac{\partial x(p, r)}{\partial p}|_{p=p_1} m_i(p_1, r)\Delta p - x(p_1, r)\Delta p$$

As

$$-\Delta x \equiv x(p_0, r) - x(p_1, r) \approx -\frac{\partial x(p, r)}{\partial p}|_{p=p_1}\Delta p$$
$$\Rightarrow \Delta x \approx \frac{\partial x(p, r)}{\partial p}|_{p=p_1}\Delta p$$

We finally obtain for $i \in LL, RL$

$$\Delta w \equiv w_i(p_1, r) - w_i(p_0, r) \approx m_i(p_1, r)\Delta x - x(p_1, r)\Delta p$$

Case 3: $i \in RR$

The kink at $x = r$ results in an indeterminate form for first-order approximation. However, independently of the change in r, the individual remains with the same consumption level (equal to his reference point) for both prices. Then, there is no behavioral effect here ($\Delta x = 0$) and we only consider the direct effect of the change.

Then, $\Delta w_i = -x(\hat{p}, r)\Delta p$, for any $\hat{p}$.

Proposition 2.2. The First-Order Social Welfare Effect of a Price Change.

As before, we begin by writing social welfare as

$$W = \int_0^{+\infty} \left[\int_{-\infty}^{p-\Lambda} \int_\theta \left\{ u(x_i^L, \nu, \theta) - px_i^L + \pi\Lambda_i(x_i^L - r) \right\} dF_\theta(\theta) dF_\nu(\nu) + \int_{p-\Lambda}^{p} \int_\theta \{u(r, \nu, \theta) - pr\} dF_\theta(\theta) dF_\nu(\nu) \right.$$
$$\left. + \int_p^{+\infty} \int_\theta \left\{ u(x_i^G, \nu, \theta) - px_i \right\} dF_\theta(\theta) dF_\nu(\nu) \right] dF_\Lambda(\Lambda)$$

Differentiating w.r.t. p and applying the Leibniz rule, we find that the boundary terms cancel out. e.g. at $u'(r) = p - \Lambda$, $x^L = r$ and $U^*(x^L, \nu, \theta, \Lambda) = U^*(r, \nu, \theta, \Lambda)$. After this cancellation, we are left with

$$\begin{aligned}\frac{\partial W}{\partial p} &= \int_0^{+\infty} \left[\int_{-\infty}^{p-\Lambda_i} \int_\theta \left\{ -x^L - (1-\pi)\Lambda \frac{\partial x^L}{\partial p} \right\} dF(\nu) dF(\theta) + \int_{p-\Lambda_i}^{p} \int_\theta \{-r\} \, dF(\nu) dF(\theta) \right. \\ &\qquad \left. + \int_{p-\Lambda_i}^{p} \int_\theta \left\{ -x^G \right\} dF(\nu) dF(\theta) \right] dF(\Lambda) \\ &= E[-x_i] - (1-\pi) E\left[\Lambda_i \frac{\partial x_i^L}{\partial p} \middle| i \in L \right],\end{aligned}$$

which yields the desired result after first-order Taylor Series approximation.

Welfare Effects with Fiscal Externalities.

Before we prove Proposition 3, we derive our main welfare effects in the presence of fiscal externalities. Doing so is necessary for the proposition and it helps us to understand the fiscal externality component of welfare effects in our empirical application. Our aim is to understand how incorporating fiscal externalities modifies equations (12) and (16).

With a fiscal externality, we can characterize efficiency using

$$\Delta W = \Delta W^{ind} + \Delta G \tag{42}$$

where ΔW^{ind} is the change in utilitarian social welfare approximated by the above results, ΔG 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 $i \in RR$.

$$\begin{aligned}\Delta W \approx &- \Delta r \pi E[\Lambda_i \mid i \in L(p, r_0)] P[i \in L(p, r_0)] \\ &- \Delta r E\left[\frac{\Lambda_i}{2} \mid i \in R(p, r_0) \right] P[i \in R(p, r_0)] + t \Delta r P[i \in R(p, r_0)]\end{aligned} \tag{43}$$

Simplifying

$$\begin{aligned}\Delta W \approx &- \Delta r \pi E[\Lambda_i \mid i \in L(p, r_0)] P[i \in L(p, r_0)] \\ &+ \Delta r E\left[-\frac{\Lambda_i}{2} + t \mid i \in R(p, r_0) \right] P[i \in R(p, r_0)]\end{aligned} \tag{44}$$

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

$$\begin{aligned}\Delta W \approx &- \Delta r \pi E[\Lambda_i \mid i \in L(p, r_0)] P[i \in L(p, r_0)] \\ &+ \Delta r E\left[-\frac{\Lambda_i}{2} + t_i \mid i \in R(p, r_0) \right] P[i \in R(p, r_0)]\end{aligned} \tag{45}$$

$$\Delta W \approx -\Delta r \pi E[\Lambda_i \mid i \in L(p, r_0)] P[i \in L(p, r_0)] + \Delta r \left\{ -E\left[\frac{\Lambda_i}{2} \mid i \in R(p, r_0)\right] + E[t_i \mid i \in R(p, r_0)] \right\} P[i \in R(p, r_0)] \tag{46}$$

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:

$$\Delta G \approx E[t\Delta x_i + x_i \Delta t] = tE[\Delta x_i] + E[x_i]\Delta t \tag{47}$$

$$= tE\left[\frac{\partial x_i}{\partial p}\right]\Delta t + E[x_i]\Delta t \tag{48}$$

$$= tE\left[\varepsilon \frac{x_i}{p+t}\right]\Delta t + E[x_i]\Delta t \tag{49}$$

For a change in prices operating through a change in tax rates, the individual welfare effect ΔW^{ind} is given by equation (16) with $\Delta p = \Delta t$.

Putting these together:

$$\Delta W \approx \left(-(1-\pi)E\left[\Lambda_i \frac{\partial x_i^L}{\partial p} \;\middle|\; i \in L\right] P[i \in L] - E[x_i(p_0, r)] \right)\Delta t + tE\left[\frac{\partial x_i}{\partial p}\right]\Delta t + E[x_i]\Delta t, \tag{50}$$

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

$$\Delta W \approx \left(-(1-\pi)E\left[\Lambda_i \frac{\partial x_i^L}{\partial p} \;\middle|\; i \in L\right] P[i \in L] \right)\Delta t + tE\left[\frac{\partial x_i}{\partial p}\right]\Delta t, \tag{51}$$

To characterize the new term further, note that obviously,

$$\frac{\partial x_i}{\partial p} = \begin{cases} \frac{\partial x_i^G}{\partial p}, i \in G \\ \frac{\partial x_i^L}{\partial p}, i \in L \\ 0, i \in R \end{cases} \tag{52}$$

We could also express these terms as elasticities, as in the second version of equation (16).

Proposition 3. Corrective Taxes for Reference Dependence.
Consider a change in t affecting individuals in the loss domain only. The above equation becomes simply:

$$\Delta W \approx \left(-(1-\pi)E\left[\Lambda_i \frac{\partial x_i^L}{\partial p} \;\middle|\; i \in L\right] P[i \in L] \right)\Delta t + tE\left[\frac{\partial x_i^L}{\partial p} | i \in L\right] P[i \in L]\Delta t, \tag{53}$$

which simplifies to

$$\Delta W \approx \left(E\left[\{t - (1-\pi)\Lambda_i\}\frac{\partial x_i^L}{\partial p} \;\middle|\; i \in L\right] P[i \in L] \right)\Delta t \tag{54}$$

Expressing this in terms of an elasticity, we have:

$$\Delta W \approx \left(E\left[\{t - (1-\pi)\Lambda_i\}\varepsilon^L \frac{x_i}{p+t} \;\middle|\; i \in L\right] P[i \in L] \right)\Delta t \tag{55}$$

To derive the optimal corrective tax we set the expression in (54) equal to zero and find that

$$t^* = (1-\pi)\left\{ E[\Lambda_i | i \in L] + \frac{Cov\left[\Lambda_i, \frac{\partial x_i^L}{\partial p} \middle| i \in L\right]}{E\left[\frac{\partial x_i^L}{\partial p} \middle| i \in L\right]} \right\} \tag{56}$$

The first term in curly brackets is the optimal tax if $\pi = 0$ and loss aversion and the demand response to a tax/price change do not covary, i.e. just the mean internality among all individuals. The second term accounts for the fact that some individuals might have a larger or smaller demand response, which implies that the optimal tax is not simply the unweighted mean of Λ_i in the L group. The entire expression for the optimal t^* in curly brackets is what Allcott and Taubinsky (2015) call the *average marginal bias*, which can also be written as

$$t^* = \frac{E\left[(1-\pi)\Lambda_i \frac{\partial x_i^L}{\partial p} \middle| i \in L\right]}{E\left[\frac{\partial x_i^L}{\partial p} \middle| i \in L\right]} \tag{57}$$

C.2 Results in Appendix B

Here we present sketches of the proofs of results in Appendix B. The steps of these proofs are virtually identical to the proofs of the analogous result in the main text; we note any important differences.

Lemma 2. The Marginal Internality in the Two-Dimensional Model.
Taking the derivative of $U_i^*(x, z - px)$ with respect to x and applying the first-order conditions from equations (30) and (31), we obtain the desired result in the gain domain and loss domain cases. In the $x_i = r$ case, we continue to have an undefined internality when $\pi = 1$, and when $\pi = 0$, we find that the internality is $u_i'(r) - p$ as in the original model (but note that this term is now ambiguously signed).

Proposition 4. First-Order Social Welfare Effects in the Two-Dimensional Model.
In this model, welfare is given by equation (32), and $v_i = w_i = 0$ locally when $x = r$. Note that we presume $s(r) = z_i - pr$ but we disregard $\partial s / \partial p$, so that a change in p does not effect the reference point s and does not cause a direct welfare effect. Taking derivatives of equation (32) with respect to r or p for the G, R, L cases, substituting for $u_i'(x)$ using the FOCs for the G and L case from equations (30) and (31), and integrating over the three first-order groups for social welfare, we obtain the desired results.

Turning to the the model in Section B.2, the FOC in this case are:

$$\tilde{u_i}'(\tilde{x}^G) + \eta_i = p \tag{FOC^G}$$

$$\tilde{u_i}'(\tilde{x}^L) + \eta_i + \Lambda_i = p \tag{FOC^L}$$

Lemma 3. Marginal Internalities With Gain Domain Payoffs.
Taking the derivative of $\tilde{U}_i^*(x, z - px)$ with respect to x and applying the first-order conditions above, we obtain the desired result in the gain domain and loss domain cases. In the $x_i = r$ case, we continue to have an undefined internality when $\pi = 1$, and when $\pi = 0$, we find that the internality is $u_i'(r) + \pi^{RD}\eta_i - p$.

Proposition 5. First-Order Social Welfare Effects with Gain Domain Payoffs.

In this model, welfare is given by equation (38), and $v_i^* = 0$ locally when $x = r$. Taking derivatives of equation (38) with respect to r or p for the G, R, L cases, substituting for $u_i'(x)$ using the FOCs for the G and L case from the equations above, and integrating over the three first-order groups for social welfare, we obtain the desired results.

D The Behavioral Equivalence of Alternative Formulations

The simple model of reference dependence from Section 2 abstracts from gain domain payoffs, and we consider a model with such payoffs as an extension in Section 4.2 and Appendix B.2. A key fact for our discussion of the model with gain domain payoffs is that these two formulations of reference dependence are behaviorally indistinguishable using observed choices, but that they carry somewhat different implications for welfare. In this appendix, we formalize the sense in which the models are behaviorally indistinguishable. An implicit assumption behind indistinguishability is that choices of x given a reference point (and price and endowment) can be observed, but choices or revealed preferences over chosen options and reference points, i.e. (x, r), cannot be observed jointly. This assumption is consistent with what is observed in typical applications of models of reference-dependent preferences, but it might be relaxed in more stylized experiments. We note that a similar equivalence, implying that the η_i parameter is typically unidentified, is shown for the stochastic case in Barseghyan et al. (2013).

In Section 2, we focus on the following model of behavior, which we will here call *Model 1*:

$$x_i(p, r, z) = \arg\max_x u_i(x) + z - px + \mathbb{1}\{x < r\}\Lambda_i(x - r), \tag{58}$$

with $u_i' > 0, u_i'' < 0$, and $\Lambda_i > 0$.

In Section 4.2 and Appendix B.2, we consider an alternative model in line with the formulation of reference dependence proposed by Tversky and Kahneman (1991), which we here call *Model 2*.

$$x_i(p, r, z) = \arg\max_x \tilde{u}_i(x) + z - px + \begin{cases} \eta_i(x - r) & x > r \\ \eta_i\lambda_i(x - r) & x \leq r, \end{cases} \tag{59}$$

with $\tilde{u}_i' > 0, \tilde{u}'' < 0$, $\eta_i > 0$, and $\lambda_i > 1$.

Consider a demand function $x(p, r, z)$, which describes the choice of x the consumer makes for any (p, r, z). We say $x(p, r, z)$ is *rationalizable* with either 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 Model 1 if and only if there is a utility function $u(x)$ with $u' > 0, u'' < 0$ and a parameter $\Lambda_i > 0$ such that for any (p, r, z) equation (58) obtains. We say $x(p, r, z)$ is rationalizable by Model 2 under analogous conditions.

We make one more modest technical assumption for our result to obtain, which is that the domain of good x is compact. In Model 1, 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. Why we need this assumption will become clear in the proof of the result below.

Proposition 6. ***Behavioral Equivalence of Model 1 and Model 2.*** *A demand function $x_i(p, r, z)$ is rationalizable by Model 1 if and only if it is rationalizable by Model 2.*

Corollary 6.1. ***A Behavioral Isomorphism.*** *If $x_i(p, r, z)$ is rationalizable by Model 1 with utility $u_i(x)$ and*

parameter Λ_i and rationalizable by Model 2 with utility $\tilde{u}_i(x)$ and parameters η_i, λ_i, then we must have

$$u_i(x) = \tilde{u}_i(x) + \eta_i x. \tag{60}$$

$$\Lambda_i = \eta_i(\lambda_i - 1). \tag{61}$$

Proof. First suppose that $x_i(p, r, z)$ is rationalizable by Model 1 with some utility $u_i(x)$ and parameter Λ_i.

Set any η_i such that $0 < \eta_i < \epsilon$.[41] Specify $\tilde{u}_i$ according to equation (60), i.e. $\tilde{u}_i = u_i(x) - \eta_i x$. Specify λ_i according to equation (61), i.e. $\lambda_i = \frac{\Lambda_i + \eta_i}{\eta_i}$.

Because $u' > \eta_i$ for any x by construction, we know that $\tilde{u}_i' = u_i' - \eta_i >= u_i' - \varepsilon > 0$, and $u'' < 0 \implies \tilde{u}_i'' < 0$. Further, by construction $\eta_i > 0$ and $\lambda_i > 1$. With the necessary restrictions satisfied, we only need to show that with these specifications, the optimization problem in equation (58) is equivalent to the optimization problem in (59). As we have guaranteed equations (60) and (61) hold, we can re-express the optimization problem in model 1 as:

$$x_i(p, r, z) = \arg\max_x \tilde{u}_i(x) + \eta_i x + z - px + \mathbb{1}\{x < r\}\eta_i(\lambda_i - 1)(x - r), \tag{62}$$

Next note that as it has no effect on the optimal x, we may freely subtract $-\eta_i r$ from the maximand. Doing so and re-arranging yields Model 2.

For the converse, suppose that $x_i(p, r, z)$ is rationalizable by Model 2 with utility function $\tilde{u}_i(x)$ and parameters $\eta_i > 0$, and $\lambda_i > 1$. Specify $u_i(x)$ using equation (60) and set Λ using (61). Checking the restrictions, we know that $\tilde{u}_i' > 0$ $\eta_> 0$, implying that $u_i' = \tilde{u}_i' + \eta > 0$, and $u_i'' = \tilde{u}_i'' < 0$. And we know that $\Lambda_i > 0$ by $\eta_i > 0$ and $\lambda_i > 1$. We can re-express the optimization problem in Model 2 as

$$x_i(p, r, z) = \arg\max_x \tilde{u}_i(x) + \eta_i x + z - px + 1x > r\eta_i(\lambda - 1)(x - r) - \eta r. \tag{63}$$

The last term has no bearing on the optimum so we can eliminate it. Applying our constructed $u_i(x)$ and Λ_i then yields Model 1. □

E 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 the model in Section 2 into the Bernheim-Rangel framework; a similar line of reasoning can be applied to the extended models in Section 4.

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) write that frames should be 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? A naive guess might be that the reference point itself is a frame, but based on the definition above, this seems inappropriate. We show in the main text that a change in the reference point can have a direct welfare effect - by changing the losses of individuals in the loss domain. Whether this direct effect should carry normative weight is a question of central importance for us, but this question belongs to a later step of the analysis, not the definition of a frame. Similarly, the theory implies that individuals should have a willingness to pay to change the reference point, suggesting that it may have

[41] The fact that we can choose such an arbitrary η_i in this step is directly related to the fact that η_i is typically unidentified from observations of observed demand.

a direct bearing on well-being. As such, we do not conceive of the reference point as a frame. 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 called observed demand in our analysis comes from choices under a single frame, f_1. This frame is analogous to what Bernheim et al. (2015) call 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:

$$u(x, y, r, f_1) = u_i(x) + y + v_i(x|r), \tag{64}$$

where v_i takes the simple form described in equation (2).

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. 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 no-reference-dependence state by encoding a frame f_0. Choices under f_0 maximize

$$u_i(x, y, r, f_0) = u_i(x) + y. \tag{65}$$

Obviously, choices under f_0 are difficult to directly observe in positive empirical analysis, 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, or by inducing individuals to use an arbitrarily low reference point (recall that $v_i = 0$ in the gain domain).

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

$$u_i(x, y, r, f) = u_i(x) + y + f \cdot v_i(x|r), \tag{66}$$

The frame f now obviously plays a very similar role in the model to π, but here we are conceiving of the two different frames purely in terms of choices in different situations.

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 in 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, the content of the results 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 (64) or

equation (65)) 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. Recall that we used the term "robust" to describe situations where whether one situation or the other was better for welfare did not depend on π. 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 robust welfare comparisons, i.e. those under which some option (x_0, y_0, r_0) is preferred to some other option (x_1, y_1, r_1) for any $\pi \in \{0, 1\}$.

As discussed in the main text, we find that decreases in the reference point tend to improve welfare for either value of π. Through the lens of the Bernheim-Rangel framework, this suggests that even if we include choices under 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 lower reference points.

Note that because $v_i = 0$ in the gain domain, we find that holding the reference point fixed, equations (64) and (65) express the same preferences in the gain domain. Moreover the reference point has no direct impact on welfare in the gain domain for either value of f or π. If we restrict our attention to individuals making choices in the gain domain, therefore, all welfare comparisons will be robust. The only potential deviations from revealed preference come from individuals choosing in the loss domain or at the reference point. This finding is the basis for the statement in Section 2 that we respect revealed preference in the gain domain. The potential deviations from revealed preferences in the naturally occurring frame that we consider are material for individuals in the loss domain or at the reference point only. Obviously, the extended models in Section 4 do not necessarily have this property. Importing those extended models requires modifications to the above. For instance, we would need to introduce three frames rather than two to import the model in Section 4.2 into the Bernheim-Rangel framework. But the general approach by which we could give those models a behavioral revealed preference interpretation remains the same.

F Empirical Application

F.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. 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 $\hat{R}$.

F.2 Two-Dimensional Reference Dependence in the Empirical Application

F.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 (19) to include consumption reference dependence:

$$U = C - \frac{n}{1+\frac{1}{\epsilon}}\left(\frac{R}{n}\right)^{1+\frac{1}{\epsilon}} - \begin{cases} 0 & R < \hat{R} \\ \widetilde{\Lambda}_l(R-\hat{R}) & R \geq \hat{R}, \end{cases} - \begin{cases} \Lambda_c(\hat{C}-C) & C < \hat{C} \\ 0 & C \geq \hat{C}, \end{cases} \tag{67}$$

where $\hat{C} = C(\hat{R})$ is the consumption reference point, which is assumed to correspond to the consumption level at the NRA. The parameter Λ_l captures the strength of reference dependence over leisure and Λ_c captures the strength of reference dependence in the consumption dimension.[42] Such loss aversion in consumption may arise for instance because "full" pension benefits become available at the NRA, and individuals perceive the associated consumption level as a reference point (Behaghel and Blau 2012).[43]

[42] Λ_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.

[43] 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

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 6 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 4.1.2, 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 3.2. 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

$$\frac{b}{\hat{R}} = \left[\left(\frac{1-\tau}{1-\tau-\Delta\tau-\Lambda_l}\right)^\varepsilon - 1\right] + \left[1 - \left(\frac{1}{1+\Lambda_c}\right)^\varepsilon\right] \tag{68}$$

Hence, bunching has two components. The first term in equation (68) 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.

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.

F.2.2 Parameter Estimation and Simulations

Analogously to equation (21), bunching observed at a threshold i, which may be the Normal Retirement Age or a pure financial incentive discontinuity, can be written as

$$\frac{b_i}{\hat{R}_i} = \left[\left(\frac{1-\tau_i}{1-\tau_i-\Delta\tau_i-\Lambda_l\cdot D_i}\right)^{\varepsilon}-1\right]+\left[1-\left(\frac{1}{1+\Lambda_c\cdot D_i}\right)^{\varepsilon}\right]+\xi_i \tag{69}$$

where D_i is an indicator for the Normal Retirement Age and ξ_i is an error term. A key issue with the estimation is that Λ_l and Λ_c cannot be separately identified based solely on equation (69). 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 Λ_l and Λ_c.

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

$$\frac{b_i^+}{\hat{R}_i} = \left[\left(\frac{1-\tau_i}{1-\tau_i-\Delta\tau_i-\Lambda_l\cdot D_i}\right)^{\varepsilon}-1\right]+\xi_i^+ \tag{70}$$

and bunching from the left is

$$\frac{b_i^-}{\hat{R}_i} = \left[1-\left(\frac{1}{1+\Lambda_c\cdot D_i}\right)^{\varepsilon}\right]+\xi_i^- \tag{71}$$

where $b_i = b_i^+ + b_i^-$. Denoting $\alpha_i = b_i^+/b_i$ the share of excess mass originating from the right, this share ranges between a minimum $\hat{\alpha}_i$ and 1. The minimum right bunching share $\hat{\alpha}_i$ is given by 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 Λ_l and Λ_c. First, we can simulate the full range of possible combinations of the two parameters by gradually moving the share of right bunching at the NRA from its minimum to 1 and estimating equations (70) and (71) using the implied values of b_i^+ and b_i^-. Panel (a) of Appendix Figure A5 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 7.

As a second approach, we aim at obtaining a set of preferred "point" estimates of Λ_l and Λ_c. For this, an empirical estimate of α_i 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.[44]

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 A5 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 $\alpha_i = 0.867$. Thus, the estimated share of bunching from the right due to reference dependence over leisure is 86.7% and the share of bunching from the left due to reference dependence over consumption is 13.3%. Finally, the parameters Λ_c and Λ_l can be estimated by plugging the bunching shares into equations (70) and (71). 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.

[44] 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.