Supplemental Appendix

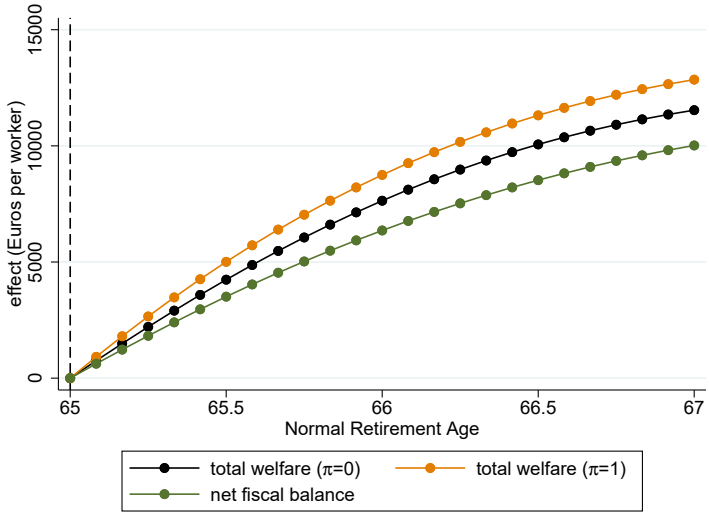
for "The Welfare Economics of Reference Dependence,"

by Daniel Reck and Arthur Seibold

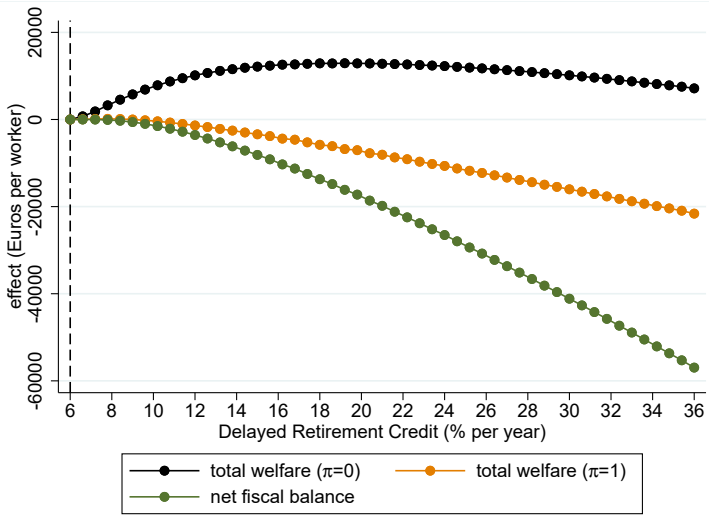
A Additional Figures and Tables

FIGURE A1: EXTENDED POLICY SIMULATIONS

(a) Increasing the Normal Retirement Age



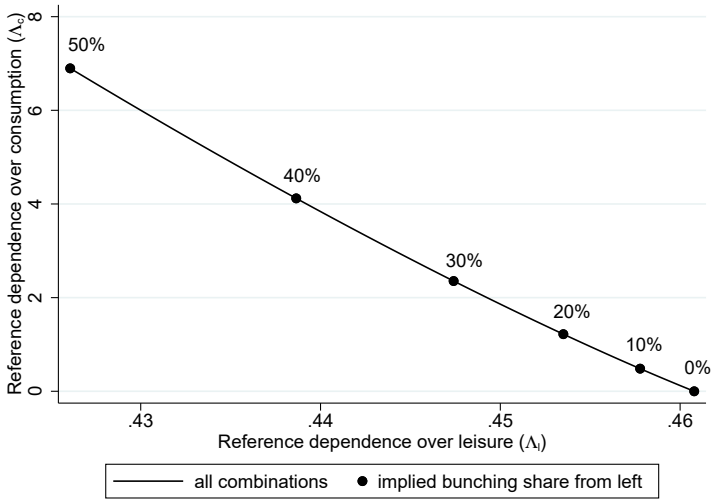
(b) Increasing the Delayed Retirement Credit



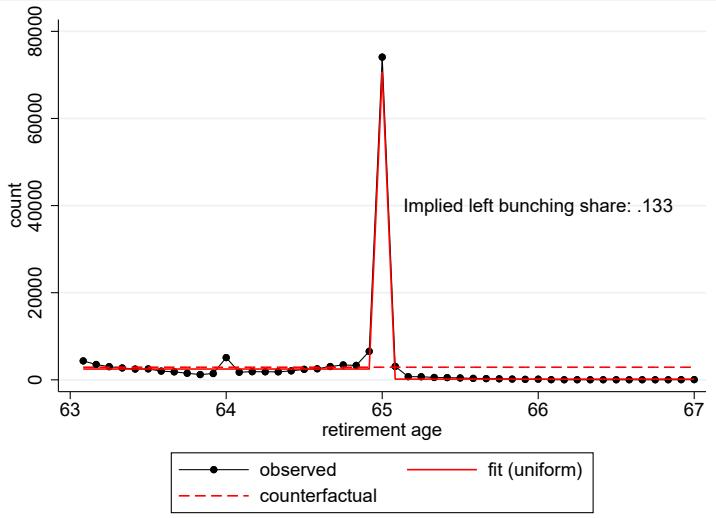
Notes: The figure shows simulated fiscal and welfare effects of pension reforms over an extended range of policies. Panel (a) shows the effects of increasing the Normal Retirement Age to ages between 65 and 67 in monthly increments. Panel (b) shows the effects of increasing the Delayed Retirement Credit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

FIGURE A2: TWO-DIMENSIONAL REFERENCE DEPENDENCE

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

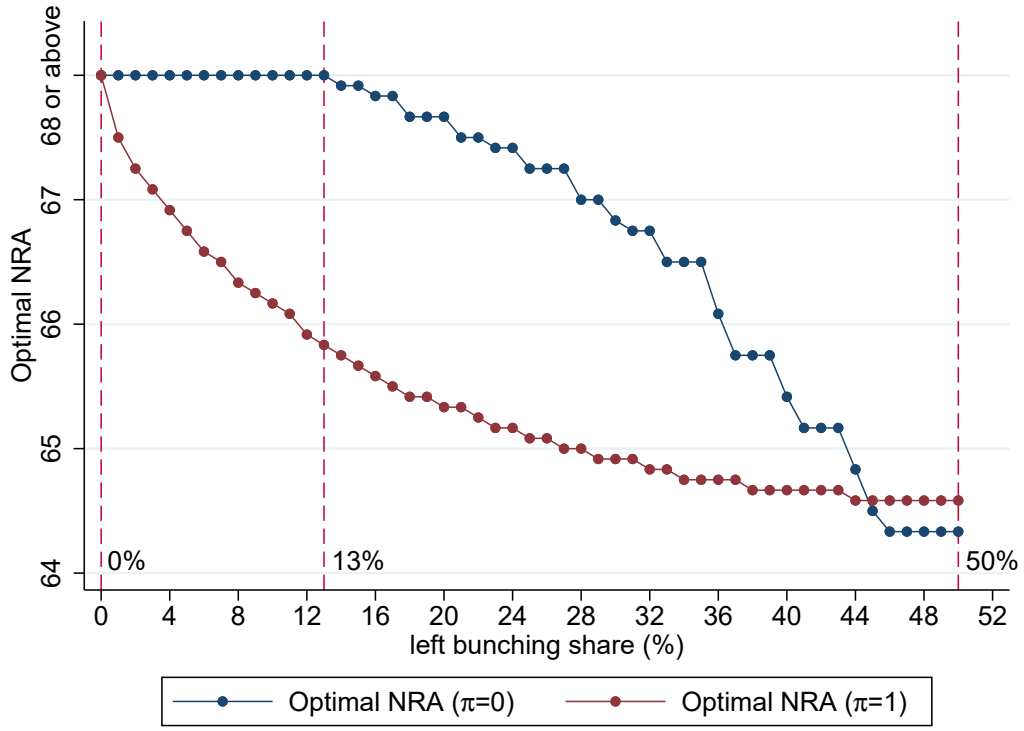


(b) Preferred Bunching Share Estimate



Notes: The figure illustrates the empirical identification of two-dimensional reference dependence parameters. Panel (a) 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.3. 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.

FIGURE A3: WELFARE-MAXIMIZING NORMAL RETIREMENT AGE



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

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 (21), using the bunching observations summarized in Panel A.

TABLE A2: PARAMETERS FOR SUFFICIENT STATISTICS CALCULATIONS

Parameter	Value
Loss aversion parameter 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 $P(i \in L)$	0.154
Fraction in R group $P(i \in R)$	0.456
Leisure demand responsiveness $E\left[\frac{\partial l_i^L}{\partial[w_i(1-\tau_i)]}\right]$	-0.017
Average change in implicit tax rate $E(\Delta\tau_i)$ (main DRC reform)	-0.264
Average change in implicit tax rate $E(\Delta\tau_i)$ (small DRC reform)	-0.029

Notes: The table shows the parameter values entering the sufficient statistics calculations in Section III.D.

TABLE A3: WELFARE EFFECTS OF INCREASING THE NORMAL RETIREMENT AGE – ALTERNATIVE SCENARIOS

	(1)	(2)
	Policy 1: Normal Retirement Age to 66	
	Stylized scenario: without benefit cut	Realistic scenario: with benefit cut
Contributions collected	+2,359	+2,359
Benefits paid	+3,999	+7,658
Net fiscal effect	+6,358	+10,017
Worker consumption	+4,230	+571
Disutility from work	–2,950	–2,950
Worker welfare ($\pi = 0$)	+1,280	–2,379
Ref. dep. disutility from work	–6,835	–6,835
Ref. dep. utility from ref. point	+7,946	+7,946
Worker welfare ($\pi = 1$)	+2,391	–1,268
Total welfare ($\pi = 0$)	+7,638	+7,638
Total welfare ($\pi = 1$)	+8,749	+8,749

Notes: The table shows results from simulations of two pension reforms, both of which are variants of the increase in the Normal Retirement Age from 65 to 66. The first scenario increases the NRA without associated benefit cuts as in Table 2. The second scenario links the NRA increase to a benefit cut, as full pension benefits are only available from the new NRA of 66. 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 of the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare.

TABLE A4: 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 2. 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 of the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare.

B Detailed Analysis of Reference-Dependent Payoff Formulations

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

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

TABLE B1: REFERENCE-DEPENDENT PAYOFF FORMULATIONS

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

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

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

	(1)	(2)	(3)	(4)
	Welfare Effect $w_r(p,r)$ by Domain			
Description	Gain Domain $(x > r)$	Reference Domain $(x = r)$	Loss Domain $(x < r)$	Individually Optimal Reference Points
Simple Loss Aversion	0	$u'(r) - p$	$-\pi\Lambda$	$(-\infty, r^*]$
Loss Aversion with Gain Utility	$-\pi\eta$	$u'(r) - p$	$-\pi(\eta + \Lambda)$	$\pi = 0 : (-\infty, \tilde{r}]$ $\pi = 1 : -\infty$
Utils Formulation (Köszegi-Rabin)	$-\pi\eta u'(r)$	$u'(r) - p$	$-\pi(\eta + \Lambda)u'(r)$	$\pi = 0 : (-\infty, \tilde{r}]$ $\pi = 1 : -\infty$
Gain Discounting	$\pi\Gamma$	$u'(r) - p$	0	$[r^*, \infty)$
Simple Loss Aversion with Diminishing Sensitivity	0	$u'(r) - p$	$-\pi\Lambda(r-x)^{\alpha-1}$ $+(1-\pi)\Lambda(r-x)^{\alpha-1}x_r$	$\pi = 0$: $(-\infty, r^*), +\infty$ $\pi = 1$: $(-\infty, r^*)$
Loss Aversion with Gain Utility & Diminishing Sensitivity	$-\pi\eta(x-r)^{\alpha-1}$ $+(1-\pi)\eta(x-r)^{\alpha-1}x_r$	$u'(r) - p$	$-\pi(\eta + \Lambda)(r-x)^{\alpha-1}$ $+(1-\pi)(\eta + \Lambda)(r-x)^{\alpha-1}x_r$	$\pi = 0$: $-\infty, +\infty$ $\pi = 1 : -\infty$
Two-Dimensional Loss Aversion, (r_x, r_y) on budget constraint	$\pi\Lambda_y p$	$u'(r) - p$	$-\pi\Lambda_x$	$r_x = r_x^*$ $r_y = r_y^*$
Two-Dimensional Loss Aversion with Gain Utility, (r_x, r_y) on budget constraint	$\pi(\eta_y p - \eta_x + \Lambda_y p)$	$u'(r) - p$	$\pi(p\eta_y - \eta_x - \Lambda_x)$	See Appendix B.6
Two-Dimensional Loss Aversion, any (r_x, r_y)	0	$u'(r) - p$ $-1\{y < r_y\}\pi p\Lambda_y$	$-\pi\Lambda_x$	$r_x \in (-\infty, r_x^*]$ $r_y \in (-\infty, r_y^*]$

Notes: The table evaluates welfare effects of changes in the reference point and describes individually optimal reference points for the payoff formulations from Table B1. Columns (1) to (3) evaluates the marginal welfare effect of changing the reference point $w_r(p,r)$ in the Gain, Reference, and Loss Domains. Note that for specifications with diminishing sensitivity, we do not express the behavioral response x_r in terms of primitives in the table due to space constraints. See Appendix B.5 for details. Column (4) shows the set of individually optimal reference points under each formulation, where r^* is the intrinsic optimum characterized by $u'(r^*) = p$ ($u'(r_x^*) = p; r_y^* = z - pr_x^*$ in the two-dimensional case), and $\tilde{r}$ is the reference point at the boundary between the gain and reference domain. Under two-dimensional loss aversion with gain utility and a reference point on the budget constraint, any of the cases from Lemma 1 could apply (see Table B1), and due to space constraints we defer the characterization of optimal reference points in this case to Appendix B.6.

B.1 Simple Loss Aversion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

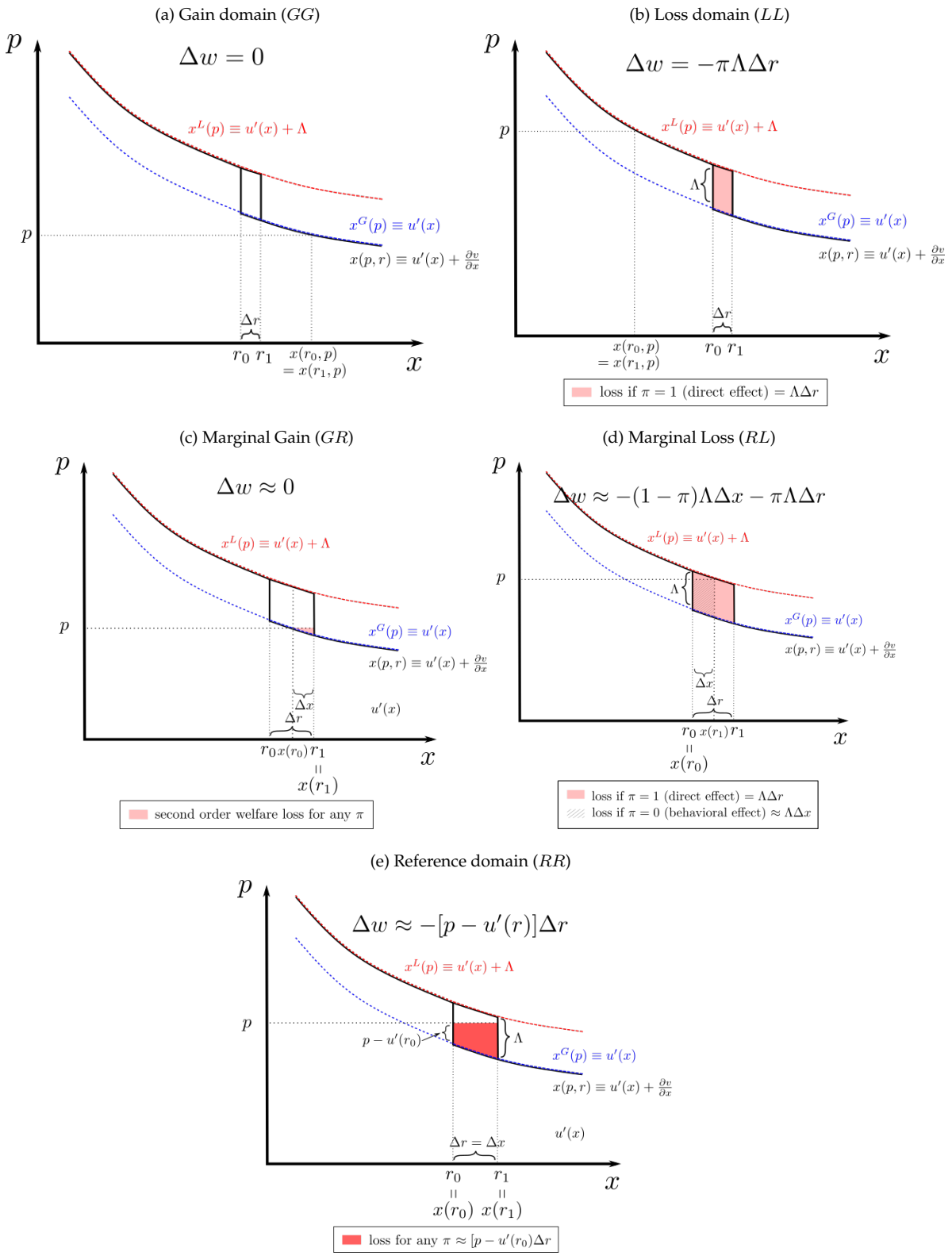
Optimal Corrective Taxes. The corrective tax schedule for good x that maximizes social welfare for a given a reference point r is characterized by

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

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

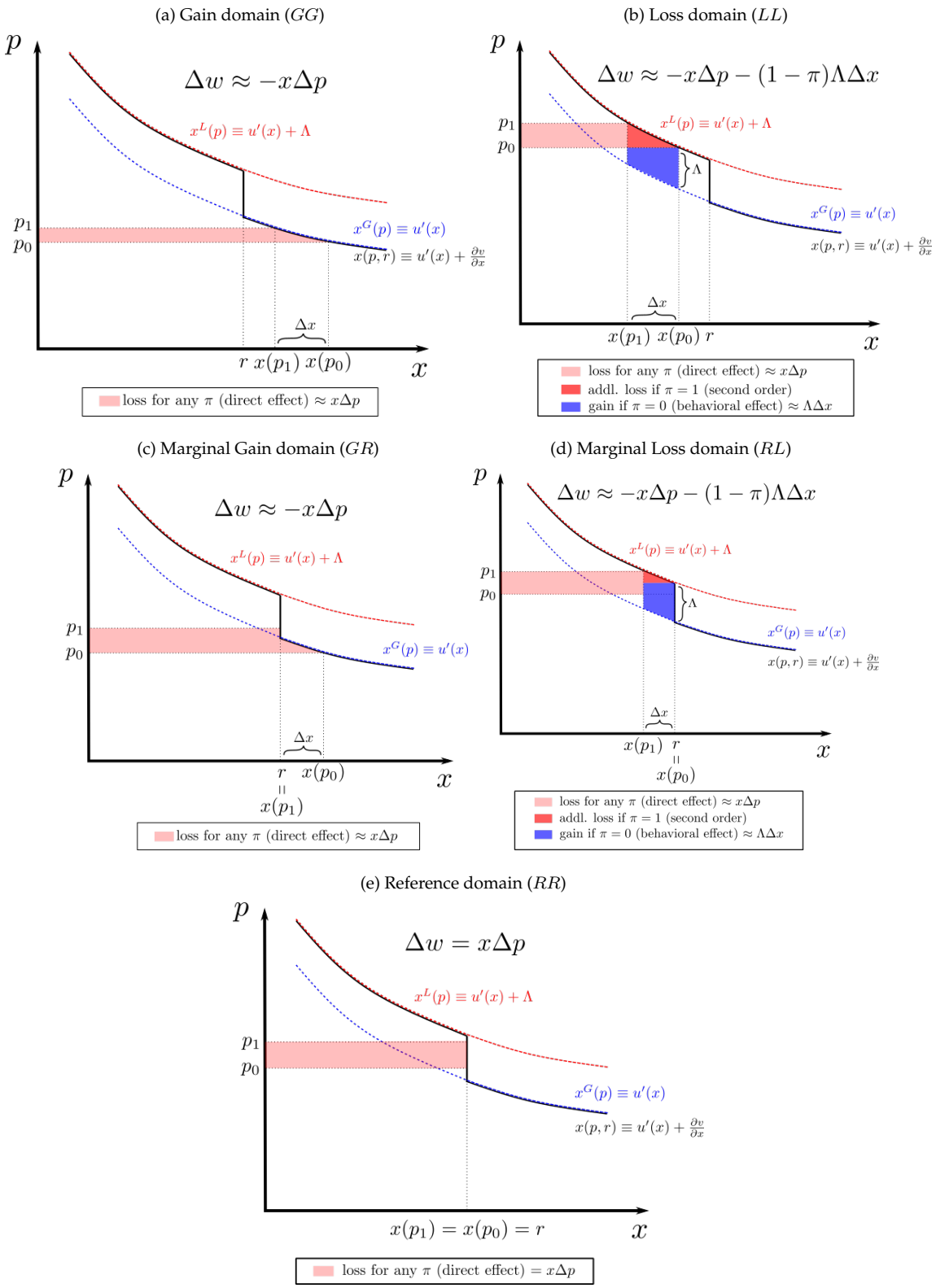
When reference dependence carries full normative weight ($\pi = 1$), there is no scope for corrective taxation, as individuals are making optimal choices in this case. When reference dependence is judged as a bias ($\pi = 0$), on the other hand, it is efficient to tax losses, i.e. to tax consumption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

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



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

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



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

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.[34]

B.2 Loss Aversion with Gain Utility

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

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

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

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

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

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

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

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

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

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

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

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

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

Set any η such that $0 < \eta < \epsilon$.[35] Specify $\tilde{u}$ according to equation (32), i.e. $\tilde{u} = u(x) - \eta x$. Specify λ_i according to equation (33), i.e. $\lambda_i = \frac{\Lambda_i + \eta_i}{\eta_i}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

B.3 Reference Dependence over Utils

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

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

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

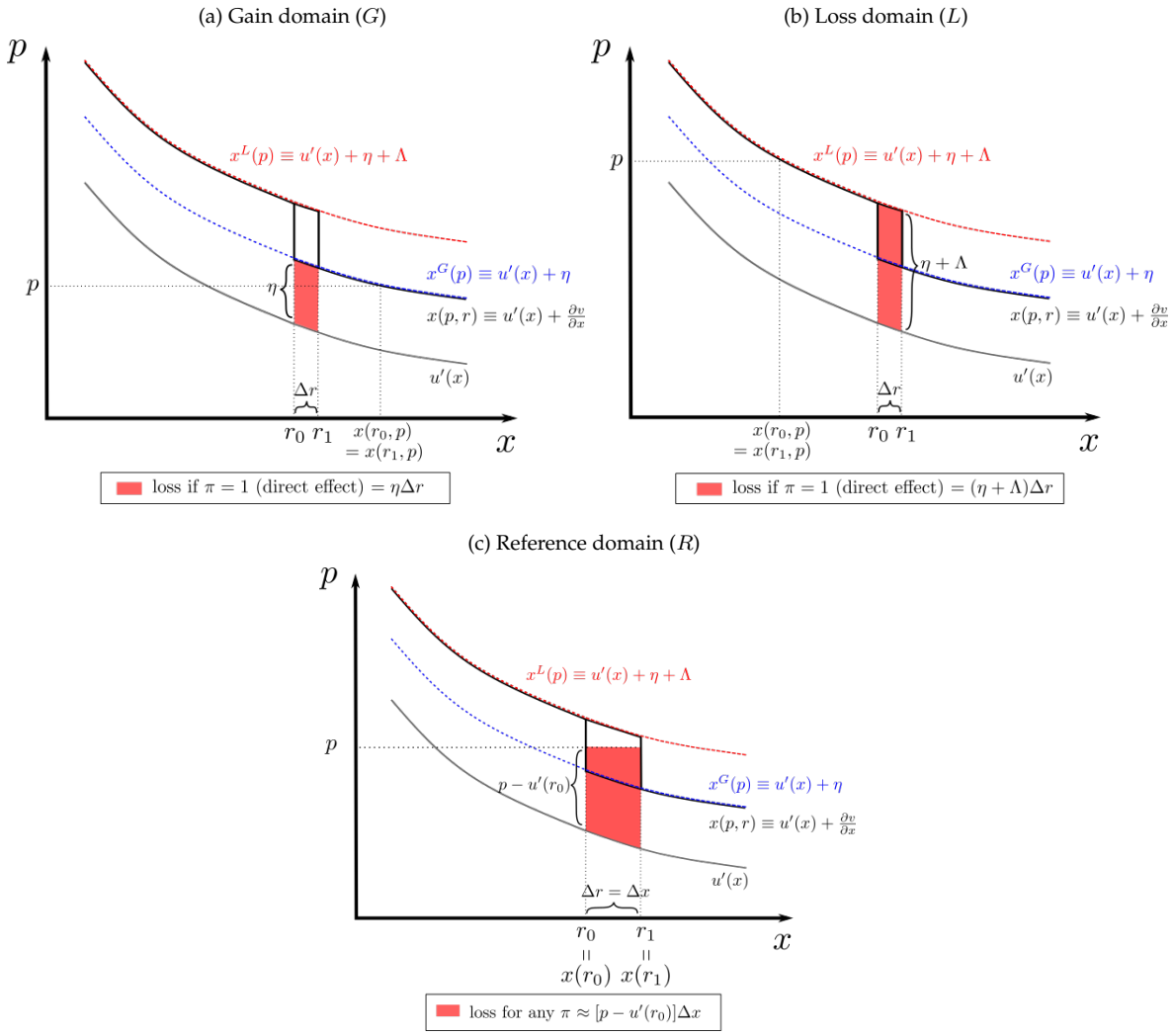
Note that we adopt a structure analogous to Loss Aversion with Gain Utility here, which is in line with Kőszegi and Rabin (2006). Alternatively, a version of Simple Loss Aversion over utils would also be straightforward to analyze.

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

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

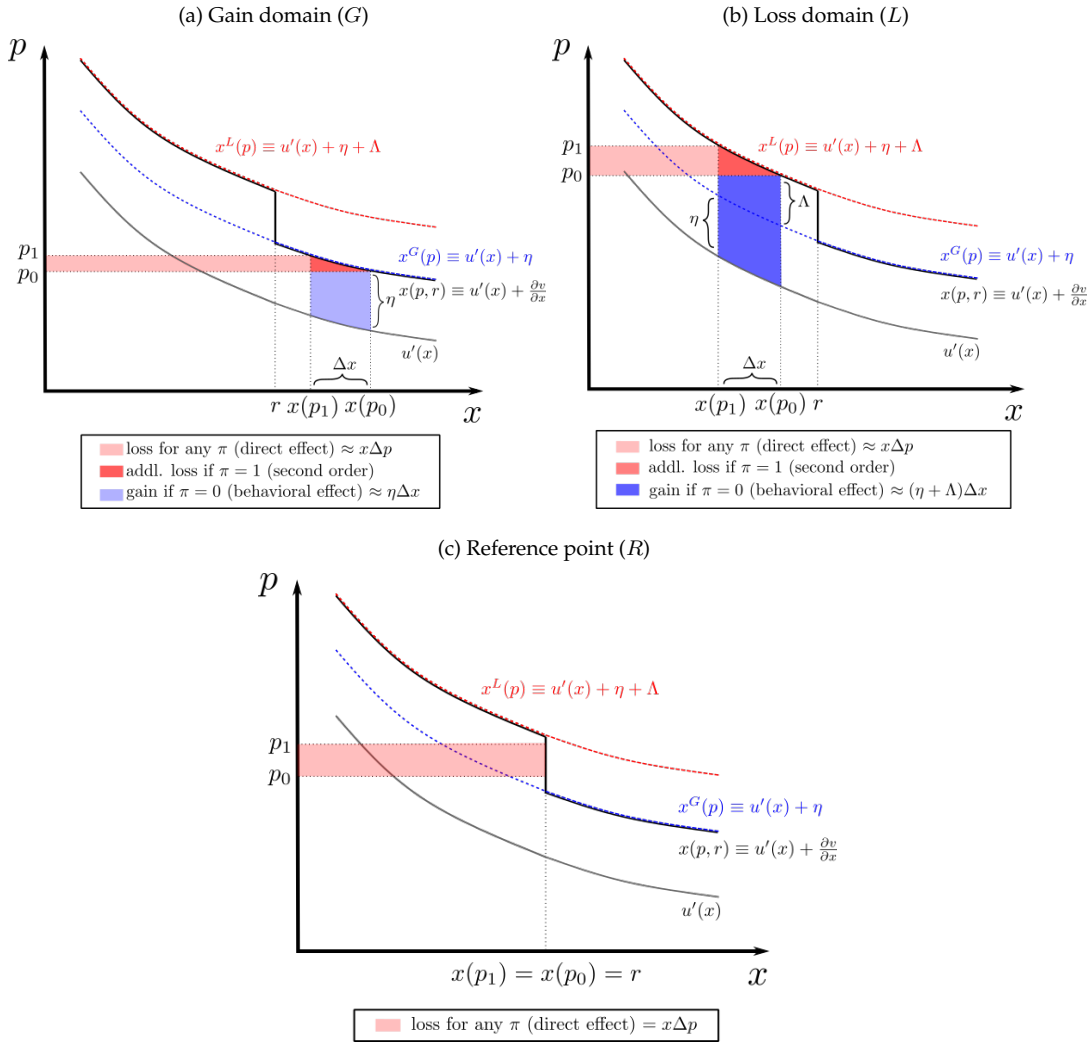
$$u_i'(x_i^L(p))(1 + \eta_i + \Lambda_i) = p. \tag{42}$$

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



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

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



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

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

B.4 Gain Discounting

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

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

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

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

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

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

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

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

effects of price change is given by

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

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

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

B.5 Incorporating Diminishing Sensitivity

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

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

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

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

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

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

In previous formulations, there were no behavioral responses to a marginal change in the reference point in the gain and loss domains ($x_r^G = x_r^L = 0$), but with diminishing sensitivity there are such behavioral

responses. Differentiating the first-order conditions with respect to r, we find

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

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

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

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

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

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

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

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

Intuitively, as r grows to either extreme, the individual stops chasing gains or avoiding losses because they are so far from the reference point that a marginally larger gain or loss does not matter much to them. Behavior thus converges to the intrinsic optimum, as if individuals did not care about reference dependence, and of course the intrinsic optimum is the optimal choice under $\pi = 0$. It is important to note that the lowest

possible reference point is a robust choice in the sense that it is optimal regardless of π. However, the highest possible reference point is optimal only under $\pi = 0$, while it minimizes welfare under $\pi = 1$.[36]

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

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

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

B.6 Two-Dimensional Reference Dependence

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

Two-Dimensional Loss Aversion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The case where the reference point lies in the interior of the budget set can be analyzed along similar lines.

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

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

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

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

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

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

Two-Dimensionsal Loss Aversion with Gain Utility

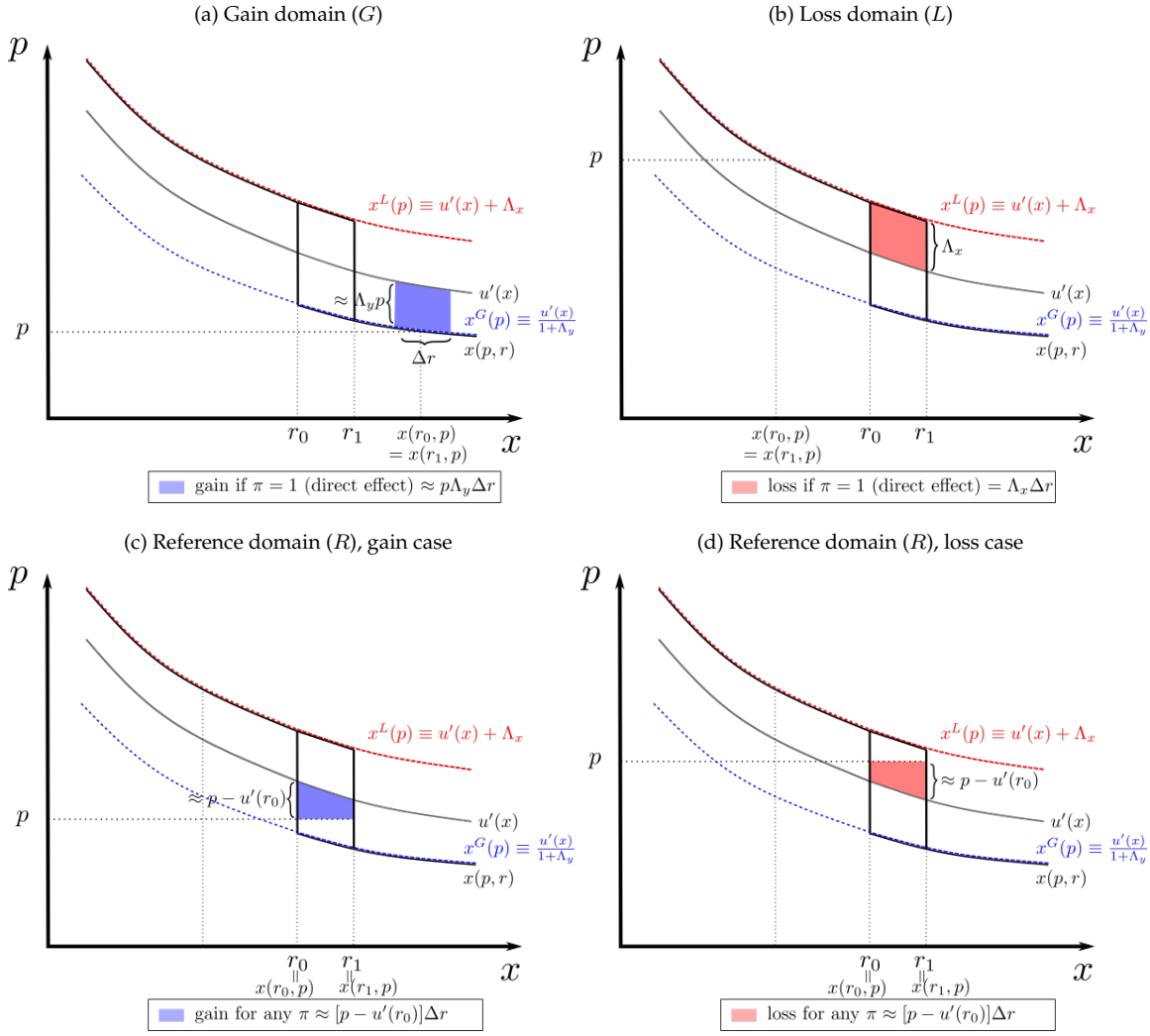
Next, we incorporate gain utility into two-dimensional reference dependence. We consider the following payoff formulation:

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

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

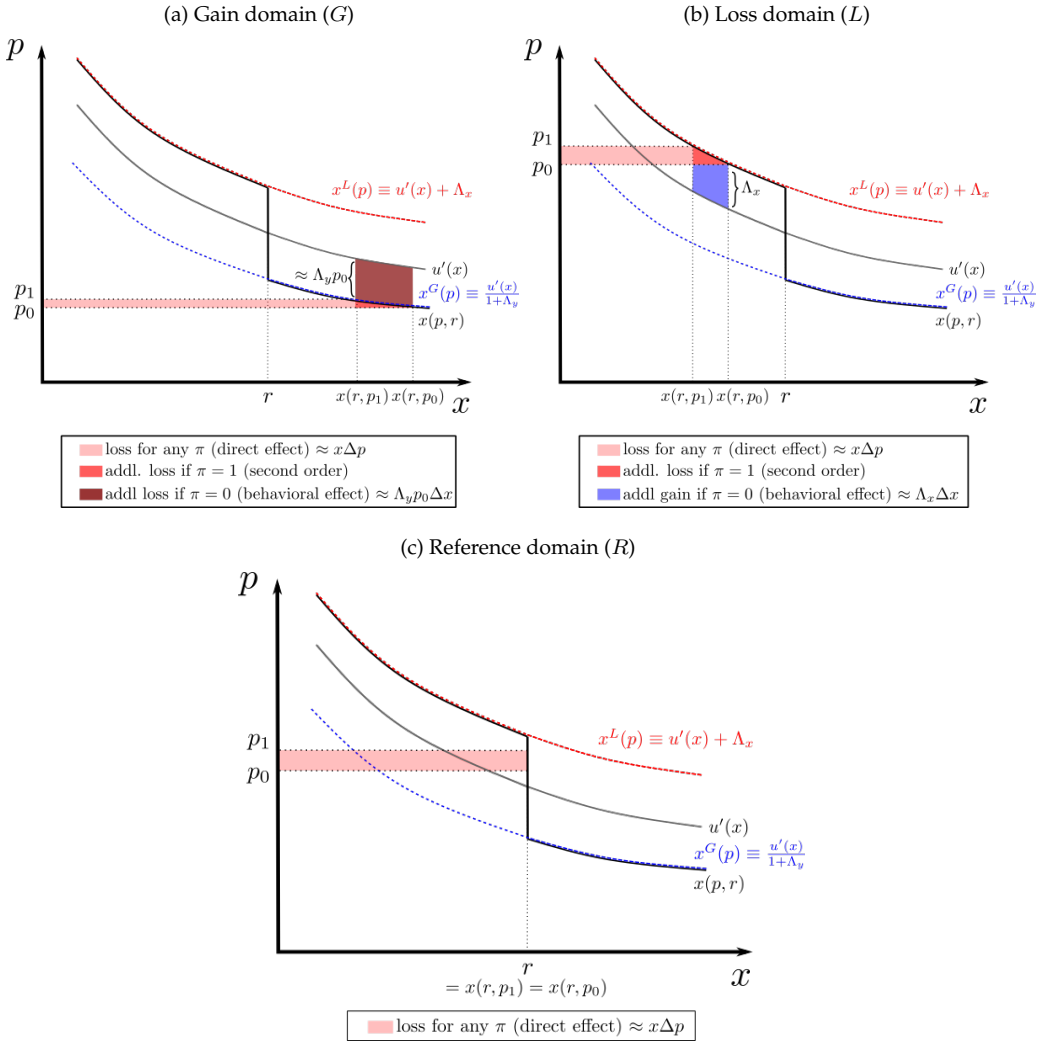
$$v(x,r) = \begin{cases} \eta_x(x - r_x) + p(\eta_y + \Lambda_y)(r_x - x), & x \geq r \\ (\eta_x + \Lambda_x)(x - r_x) + p\eta_y(r_x - x) & x < r. \end{cases} \tag{57}$$

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



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

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



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

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

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

We can sign the derivative in order to find which case of Lemma 1 obtains.[37]

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

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

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

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

B.7 Our Flexible Reduced Form as an Approximation

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

The Flexible Reduced Form from equation (9) is given by

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

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

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

In the following, we show that (i) any formulation satisfying Assumptions 1 and 2 admits a first-order approximation via equation (9) with $\Lambda > 0$ equal to the size of the kink in preferences and $\beta \in \mathbb{R}$, and (ii) if in addition the formulation falls in the Singe-Peaked case from Lemma 1, then $\beta \in [0, 1]$.

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

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

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

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

Let $\Lambda = \nu'_-(0)\mu'(r_0) - \nu'_+(0)\mu'(r_0)$. Note that this is the implied size of the kink in preferences around $x = r_0$ above. And let $\beta = -\nu'_+(0)\mu'(r_0)/\Lambda$. Note that $1 - \beta = \nu'_-(0)\mu'(r_0)/\Lambda$. Then the approximate formulation (59) becomes equation (9). All that remains to check are the parametric restrictions. We have $\Lambda > 0$ by Assumption 1.3. The parameter β then turns on which of the cases from Lemma 1 obtains:

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

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

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

C Proofs

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

P2.1. *The first-order social welfare effect of a change in the reference point is given by*

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

P2.2. *If the distribution of $u_i'(r) - p$ is independent of (β_i, Λ_i) and distributed uniformly conditional on $i \in R(p,r)$, the first-order social welfare effect of a change in the reference point is determined by*

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

P2.3. *The first-order social welfare effect of a change in price is given by*

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

where ε_i is the price elasticity of demand for good x.

Proof. Proof of 2.1. There are two main steps in this proof. The first step is to derive the individual welfare effect under the Flexible Reduced-Form specification (9) in each of the three domains. The second step is to show that welfare effects at the boundary of these cases are irrelevant for the first-order welfare effect, so that the social welfare effect is a simple aggregation of the welfare effects in the three domains.

To do this, we need only evaluate the derivatives in equation (4). As noted in the previous proof, we have $x_{i,r} = 0$ for $i \in G, L$ because equation (9) satisfies Assumption 2.2. For these cases therefore we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Evaluating the derivatives for the individual welfare effect of a price change in equation (8) under formulation (9), we find:

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

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

Proposition 3. ***Identification from Bunching.*** *Define a random variable $\Delta_i = u_i'(r) - p$ and denote its density by f_Δ. Assume that Δ, Λ and β are mutually independent, and that Δ_i is uniformly distributed conditional on i R.*

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

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

P3.2. *If $\beta_i \geq 0$ $\forall i$,*[38] *the share of bunching that comes from the right – the share of individuals who would choose to consume more than r in the absence of reference-dependent payoffs – is*

$$P[r_i^* > r | i \in R] = E[\min\{\beta_i, 1\}].$$

[38]Note that we prove a more general version of Proposition 3.2 here. The proof of the version shown in the main text follows analogously.

If $\beta_i \leq 1$ for every $\forall i$, the share of bunching that comes from the right is

$$P[r_i^* > r | i \in R] = E[\max\{\beta_i, 0\}].$$

Proof. We begin by building on the characterization of the composition of the three groups in terms of β, Λ, and Δ.

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

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

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

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

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

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

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

So far, our derivations allow any $\beta \in \mathbb{R}$. In Proposition 3.2 in the main text, we state a result that is implied by the statement of the proposition above. Here we provide a proof for the more general statement, using a truncation of β_i to characterize the right bunching share.

We begin with the first condition above, supposing $\beta \geq 0 \; \forall i$. An individual bunches from the right – meaning $i \in R$ and $r_i^* > r$ – if $\Delta > 0$, $\Delta < \beta\Lambda$, and $\Delta < \Lambda$, i.e. if $0 < \Delta < \min\{\beta, 1\}\Lambda$. Similarly to the characterization of $P[i \in R]$ above, the fraction of the population who bunch from the right, i.e. those $i \in R(p, r)$ for whom $r_i^* > r$, is

$$P[r_i^* > r \;\&\; i \in R] = \int_{\beta,\Lambda} [f_\Delta(0)(\min\{\beta, 1\}\Lambda - 0)] \, dF_{\beta,\Lambda}(\beta,\Lambda) \approx f_\Delta(0)E[\min\{\beta_i, 1\}\Lambda_i].$$

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

$$P[r_i^* > r | i \in R] = \frac{P[r_i^* > r \;\&\; i \in R]}{P[i \in R]} \approx E[\min\{\beta, 1\}].$$

The proof for the other case, where $\beta_i \leq 1 \; \forall i$, is analogous. □

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

With a fiscal externality, we can characterize efficiency using

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

where 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 $i \in R$.

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

Simplifying

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

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

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

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

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

$$\Delta G \approx E[t \Delta x_i + x_i \Delta t] = tE[\Delta x_i] + E[x_i] \Delta t$$

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

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

Putting these together:

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

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

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

To characterize the new term further, note that obviously,

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

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

D Relationship to Bernheim and Rangel (2009)

Bernheim and Rangel (2009) propose a general framework for decision-theoretic behavioral welfare economics. This appendix describes in detail the relationship between our analysis and this framework. We focus on mapping our Flexible Reduced-Form specification from equation (9) into the Bernheim-Rangel framework; a similar line of reasoning can be applied to other formulations.

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

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

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

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

where v takes the form described in equation (9), or some other formulation from Appendix B.

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

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

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

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

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

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

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

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

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

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

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

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

Alternative Approach Under $\pi = 0$. Suppose, contrary to our preferred line of reasoning above, that we wish to conceive of the reference point as a frame. In this case, we could actually think of demand in the gain domain and demand in the loss domain as demand under two different frames. Note that under $\pi = 0$, with equation (9) we nest the case where demand in the gain domain is normative by $\beta = 0$, as in this case we have $v(p, r) = 0$ for $(p, r) \in G$. Similarly, demand in the loss domain is normative when $\beta = 1$. Thus, we could think of the parameter β as capturing normative ambiguity over whether gain or loss domain demand are normative, provided we are willing to also assume that reference-dependent payoffs generally are not

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

E Social Welfare with Heterogeneous Reference Points

When we derive our main social welfare results in Section II, we assume that the reference point itself is homogeneous. In this appendix, we relax this assumption. Note that for our results characterizing individual welfare in Section I, the reference point is generally allowed to be individual-specific. However, this does not imply we can simply replace r with r_i in our social welfare results in Section II. Heterogeneous reference points raise some fundamental theoretical issues and complications for empirical identification that we discuss below.

Comparability and Heterogeneity. How should we compare welfare across individuals with different reference points? The individual welfare metric – the function $w(p, r; \theta_i, \Gamma_i, \beta_i)$ in our model – is identical to equivalent variation from a baseline situation in which an individual chooses the reference point. With homogeneous reference points, this means we are evaluating equivalent variation relative a baseline situation in which everyone chooses the same option.[39] With heterogeneous reference points, we can still derive a utility function in equivalent variation units, but the baseline from which this equivalent variation is derived varies between individuals with different reference points. As we do not examine distributional welfare effects in this paper, one straightforward possibility is to assume the value of a dollar is identical across individuals with different reference points. Under this assumption, we can continue to aggregate utility in equivalent variation units to evaluate social welfare. However, if we relax this normative assumption, the marginal value of a dollar may co-vary with the reference point r_i, and this covariance will matter for welfare. We defer a full consideration of this and other distributional matters to future work.

Assuming our welfare function is comparable in level and units across individuals, we can express welfare as an integral over a welfare function with heterogeneous reference points, just as in the social welfare function from equation (11), but with an individual-specific reference point r_i. In other words, under this assumption, we obtain the same expressions for welfare at a given status quo, but with r_i. Note that when we decompose welfare into the G, L and R groups, the expression for expected welfare in each group is the same, but now the composition of these groups depends on r_i. This does not create any problems for the theory but entails new empirical challenges that we discuss below.

Influence of Policy on the Reference Point. When we introduce policy variation and characterize the welfare effects of a reform like a change in the NRA, r_i is now a heterogeneous preference parameter rather than a homogeneous policy parameter. This presents another challenge.

Let $X \in \mathbb{R}$ be a policy like the NRA that influences behavior by changing reference points. Formally, to introduce heterogeneity in r, a simple approach is to imagine that given the policy X, r_i depends on the unstructured individual preference parameter we have already introduced, θ_i. So we suppose there is a function $\tilde{r}$ such that $r_i = \tilde{r}(\theta_i, X)$. Assuming this function is differentiable almost everywhere, for a policy perturbation ΔX, we have

$$\Delta r_i = \frac{\partial \tilde{r}}{\partial X} \Delta X.$$

We obtain all the results in the main text under a homogeneity restriction that we can now formalize: for any θ_i, $\tilde{r}(\theta_i, X) = X$. Homogeneity implies that $\Delta r_i = \Delta X$, so when we aggregate welfare using expectations, we can pull out the term ΔX in equation (12). Without this restriction, we must account for the covariance between Δr_i and the other components of the welfare effect in our expressions.

[39] The baseline conditions and comparability assumptions that justify our approach are formally articulated in Naik and Reck (2025).

Social Welfare Effects. With heterogeneous reference points, the expression for the first-order welfare effect of a change in X – the analogue of equation (12) – becomes

$$\begin{aligned}\Delta W \approx \pi \{&E[\beta_i \Lambda_i \Delta r_i \mid i \in G] P[i \in G] - E[(1-\beta_i)\Lambda_i \Delta r_i \mid i \in L] P[i \in L]\} \\ &+ E\left[(u_i'(r_i) - p) \Delta r_i \;\middle|\; i \in R \right] P[i \in R]. \end{aligned} \tag{64}$$

Notice that Δr_i now appears inside the expectations. Making the simplifying assumption from Proposition 2.2, we obtain the following analogue of equation (13):

$$\begin{aligned}\Delta W \approx \pi \{&E[\beta_i \Lambda_i \Delta r_i \mid i \in G] P[i \in G] - E[(1-\beta_i)\Lambda_i \Delta r_i \mid i \in L] P[i \in L]\} \\ &+ E\left[\Lambda_i \left(\beta_i - \frac{1}{2} \right) \Delta r_i \;\middle|\; i \in R \right] P[i \in R] \end{aligned} \tag{65}$$

$$\begin{aligned}&= \Delta X \pi \left\{ E\left[\beta_i \Lambda_i \frac{\partial \tilde{r}}{\partial X} \;\middle|\; i \in G \right] P[i \in G] - E\left[(1-\beta_i)\Lambda_i \frac{\partial \tilde{r}}{\partial X} \;\middle|\; i \in L \right] P[i \in L] \right\} \\ &+ \Delta X E\left[\Lambda_i \left(\beta_i - \frac{1}{2} \right) \frac{\partial \tilde{r}}{\partial X} \;\middle|\; i \in R \right] P[i \in R] \end{aligned} \tag{66}$$

In words, the welfare effect is now a weighted average of the individual-specific effects from equation (13), where the weights depend on individuals' sensitivity of r_i to policy, $\frac{\partial \tilde{r}(\theta_i, X)}{\partial X}$. To see how this matters quantitatively, consider the first term in this expression, for the G group, which we can re-write as

$$E\left[\beta_i \Lambda_i \frac{\partial \tilde{r}}{\partial X} \;\middle|\; G \right] = E[\beta_i \Lambda_i \mid G] E\left[\frac{\partial \tilde{r}}{\partial X} \middle| G \right] + \text{Cov}\left(\beta_i \Lambda_i, \frac{\partial \tilde{r}}{\partial X} \;\middle|\; G \right).$$

How does the welfare effect of a policy reform under heterogeneity compare to the homogeneous case? First, we observe that if the covariance between the reference dependence parameters and sensitivity of the reference point to policy is negligible – a sufficient condition for this would be that θ_i and (Λ_i, β_i) are independent – what we need to know in addition to our initial sufficient statistics is how much the reference point shifts in each group, i.e. $E\left[\frac{\partial \tilde{r}}{\partial X} \middle| G \right]$ and analogous terms for the L and R groups. Outside of this case, one also needs to account for the covariance between reference point sensitivity and other preference parameters as in the expression above.

Empirical Implementation. If the distributions of r_i and Δr_i were known, applying these characterizations of welfare effects would be straightforward. Using information about r_i and observed choices, we would know which individuals are in the G, L, and R groups, and we would have all the information needed to evaluate the expressions above. Matters become more complicated when individual reference points and their dependence on policy are not observed. Intuitively, if we do not know what reference point individual i is using, then even if we observe their choice x_i, we do not know whether $x_i < r_i$, so we do not know which group the individual is in. Moreover, we would be unable to evaluate the $\frac{\partial \tilde{r}}{\partial X}$ terms for the different groups.

Empirically identifying heterogeneous reference points and their responsiveness to policies is challenging. One of the main advantages of our empirical setting is that there the Normal Retirement Age is a single salient number that many individuals use as a reference point, and that policy can directly influence reference points by shifting the NRA. This plausibly suggests a model with homogenous reference points that respond one-to-one to changes in the NRA. More generally, accommodating heterogeneity in an empirically implementable model would require imposing structure on reference point formation. In this appendix, we clarify what information one would need to evaluate welfare under heterogeneous reference points, but we

defer these questions of empirical identification to future work.

F Empirical Application

F.1 Simulation Methods

In order to obtain the results shown in Section III.E, we simulate the welfare effects of pension reforms, building on Seibold (2021), who calculates effects of similar reforms on retirement behavior and fiscal balances. The first reform is an increase in the NRA from 65 to 66. We simulate two variants of the NRA reform, without and with associated benefit cuts. While the former is useful in isolating the effect of changing reference points, the latter more accurately captures a "realistic" pension reform. The other type of reform we consider is an increase in the DRC. In order to anchor this change in financial incentives, we increase the credit from the current level of 6% to 10.44% per year, which yields the same effect on the average retirement age as the first reform. In addition, we simulate a small DRC increase from 6% to 6.48% which serves mainly as a demonstration that the sufficient statistics approach is more accurate for small reforms.

The policy simulations proceed in the following steps. First, we require a counterfactual distribution of retirement ages – a distribution of retirement ages in the absence of reference dependence. We follow the standard approach to obtain this counterfactual distribution and fit a polynomial to the observed distribution, excluding the bunching region around the NRA. In the absence of reference dependence, individuals bunching at the NRA would be distributed across retirement ages above the NRA, and we simulate this de-bunching by distributing the bunching mass across ages 65 and above. Unfortunately, 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. Figure F1 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 F1 shows. We then assign counterfactual retirement ages to individuals in the data based on ranks of actually observed retirement ages.

Second, we simulate optimal retirement ages for each individual under the baseline policy environment where the NRA is 65 and the DRC is 6% per year. Third, we simulate optimal retirement ages under each counterfactual policy scenario. For this, we simulate individual lifetime budget constraints from equation (19) as in Seibold (2021), based on observed individual earnings and contribution histories, and choose the retirement age that maximizes utility from equation (18) subject to the budget constraint and the reference point given by the NRA.

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

F.2 Decomposing Reference Dependence Payoffs

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

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

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

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

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

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

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

F.3 Two-Dimensional Reference Dependence in the Empirical Application

Two-Dimensional Model

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

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

where $\hat{C} = C(\hat{R})$ is the consumption reference point, which is assumed to correspond to the consumption level at the NRA. Thus, the two-dimensional reference point lies on the budget constraint. The parameter Λ_l captures the strength of reference dependence over leisure and Λ_c captures the strength of reference dependence in the consumption dimension.[40] 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).[41]

[40] 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.

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

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

Equation (68) yields the exact amount of bunching under the utility funciton we assume. Taking a first-order Taylor approximation about the point $(\Lambda_l, \Lambda_c) = (0, 0)$ under $\Delta_t = 0$, we obtain the following approximation of the excess mass at a two-dimensional reference point without a local financial incentive

available at the NRA.

kink:

$$\frac{b}{\hat{R}} \approx \varepsilon(\Lambda_l + \Lambda_c), \tag{69}$$

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

Parameter Estimation and Simulations

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

$$\frac{b_j}{\hat{R}_j} = \left[\left(\frac{1-\tau_j}{1-\tau_j-\Delta\tau_j-\Lambda_l \cdot D_j}\right)^{\varepsilon} - 1\right] + \left[1 - \left(\frac{1}{1+\Lambda_c \cdot D_j}\right)^{\varepsilon}\right] + \xi_j \tag{70}$$

where D_j is an indicator for the Normal Retirement Age and ξ_j is an error term. As discussed above, a key issue with the estimation is that Λ_l and Λ_c cannot be separately identified based solely on equation (70). 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

$$\frac{b_j^+}{\hat{R}_j} = \left[\left(\frac{1-\tau_j}{1-\tau_j-\Delta\tau_j-\Lambda_l \cdot D_j}\right)^{\varepsilon} - 1\right] + \xi_j^+ \tag{71}$$

and bunching from the left is

$$\frac{b_j^-}{\hat{R}_j} = \left[1 - \left(\frac{1}{1+\Lambda_c \cdot D_j}\right)^{\varepsilon}\right] + \xi_j^- \tag{72}$$

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

We follow two approaches in order to obtain joint estimates of 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 left bunching at the NRA from zero to its maximum and estimating equations (71) and (72) using the implied values of b_j^+ and b_j^-. Panel (a) of Appendix Figure A2 shows resulting parameter combinations. The negative slope of the relationship illustrates the intuition that the two types of reference dependence are substitutes in terms of rationalizing observed excess mass. The labeled dots in in the figure mark a range of implied left bunching shares between 0 and 50%. These results allow us to simulate the welfare effects of pension reforms as a function of the relative strength of consumption reference dependence, which are shown in Figure 6.

As a second approach, we aim at obtaining a set of preferred "point" estimates of Λ_l and Λ_c. For this, an empirical estimate of β 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.[42]

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

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

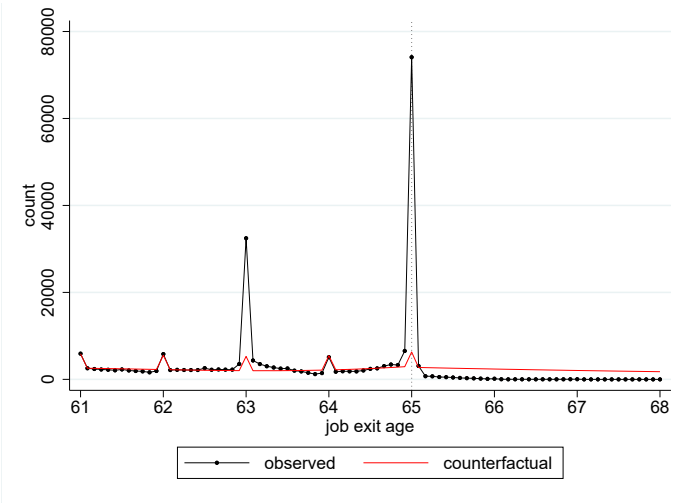
From this, the implied bunching shares from both sides can be computed as $B^- = (h_0 - h_-)(\hat{R} - R^-)$ and $B^+ = (h_0 - h_+)(R^+ - \hat{R})$ because bunching from either side must be equal to the total density shift on that side.

Panel (b) of Appendix Figure A2 illustrates this procedure. The solid red line shows the average empirical retirement density on both sides in a window of +/-2 years around the NRA, h_+ and h_-. The dashed red line shows the implied counterfactual density h_0 calculated as described above. The figure shows that the difference between the observed density and the counterfactual density is much larger on the right, indicating that most "missing density" is on this side, and thus most bunching appears to originate from above. We obtain an estimate of $\beta = 0.133$. Thus, the estimated share of bunching from the left due to reference dependence over consumption is 13.3% and the share of bunching from the right due to reference dependence over leisure is 86.7% . Finally, the parameters Λ_c and Λ_l can be estimated by plugging the bunching shares into equations (71) and (72). We obtain estimates of $\Lambda_c = 0.672$ and $\Lambda_l = 0.457$. The simulations shown in Table A4 are conducted based on these parameter estimates.

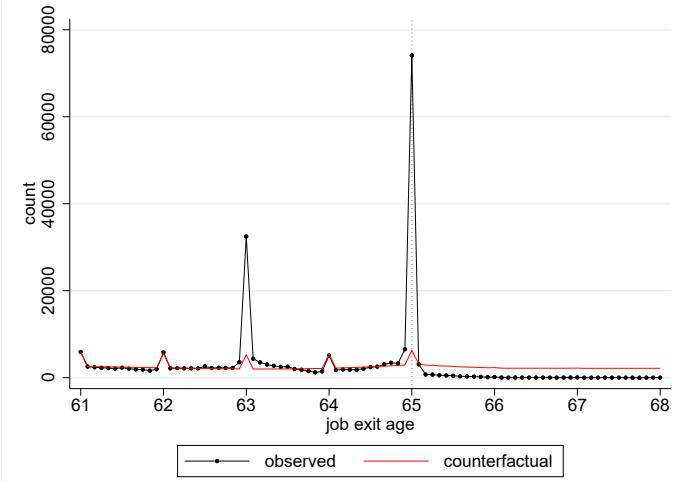
[42] Besides, although theory predicts individuals responding to the threshold along the entire density in principle, it is unclear in practice whether those far from the threshold respond in the same way as those closer.

FIGURE F1: COUNTERFACTUAL RETIREMENT AGE DISTRIBUTION

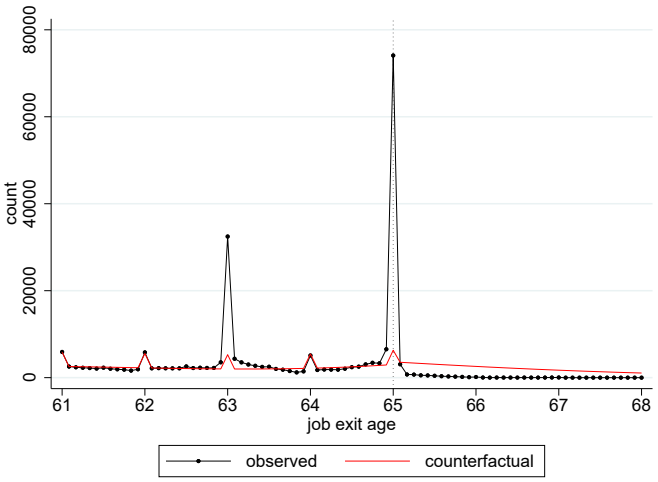
(a) Baseline: Pareto Tail



(b) Alternative 1: Uniform Tail



(c) Alternative 2: Lognormal Tail



Notes: The figure shows counterfactual retirement distributions under different assumptions about the shape of the upper tail of the distribution. In all panels, the counterfactual distribution up until the Normal Retirement Age (age 65) is obtained by fitting a seventh-order polynomial to the observed retirement age distribution, allowing for round-age effects. Panel (a) shows the baseline distribution we use in the simulations, where the upper tail is given by a fitted Pareto distribution. Panels (b) and (c) show alternative counterfactual distributions, where the upper tail is given by a uniform and lognormal distribution, respectively. Appendix Table F1 shows that our simulation results are robust to the shape of the upper tail of the counterfactual distribution.

TABLE F1: WELFARE EFFECTS OF PENSION REFORMS: ALTERNATIVE COUNTERFACTUAL DISTRIBUTIONS

	(1)	(2)
	Panel A: Uniform Tail	
	Policy 1: Normal Retirement Age to 66	Policy 2: Delayed Retirement Credit to 10.20%
Contributions collected	+2,363	+2,278
Benefits paid	+4,061	–3,879
Net fiscal effect	+6,425	–1,601
Worker consumption	+4,179	+11,937
Disutility from work	–2,901	–2,094
Worker welfare ($\pi = 0$)	+1,278	+9,843
Ref. dep. disutility from work	–6,900	–8,670
Ref. dep. utility from ref. point	+8,121	0
Worker welfare ($\pi = 1$)	+2,499	+1,173
Total welfare ($\pi = 0$)	+7,702	+8,242
Total welfare ($\pi = 1$)	+8,923	–428
	Panel B: Lognormal Tail	
	Policy 1: Normal Retirement Age to 66	Policy 2: Delayed Retirement Credit to 11.28%
Contributions collected	+2,261	+2,173
Benefits paid	+3,707	–4,386
Net fiscal effect	+5,967	–2,213
Worker consumption	+4,201	+12,103
Disutility from work	–3,198	–2,434
Worker welfare ($\pi = 0$)	+1,002	+9,669
Ref. dep. disutility from work	–6,027	–8,268
Ref. dep. utility from ref. point	+6,859	0
Worker welfare ($\pi = 1$)	+1,834	+1,401
Total welfare ($\pi = 0$)	+6,970	+7,456
Total welfare ($\pi = 1$)	+7,802	–811

Notes: The table shows results from pension reform simulations as in Table 2 under alternative assumptions about the upper tail of the retirement age distribution 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 of the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare.