The Welfare Economics of Reference Dependence *

Daniel Reck
London School of Economics

Arthur Seibold
University of Mannheim

March 2021

Abstract

Empirical evidence suggests that in many contexts, individuals evaluate options relative to a reference point, placing disproportionate weight on losses. In this paper, we analyze welfare under reference dependence. We explicitly model the central normative ambiguity over whether the influence of the reference point on choices reflects a bias or a normative preference, and we describe how judgments regarding this ambiguity affect welfare calculations. We find that policies that decrease the reference point generally improve welfare, absent countervailing externalities or other behavioral biases. In contrast, the welfare effect of price or tax changes in the presence of reference dependence depends strongly on normative judgments. We illustrate our findings with an empirical application to reference dependence exhibited in retirement decisions of German workers. Simulation results lend some support to increasing the Normal Retirement Age rather than using financial incentives in order to induce workers to postpone retirement.

*d.h.reck@lse.ac.uk and seibold@uni-mannheim.de.

We thank Jack Fisher, Jacob Goldin, Xavier Jaravel, Camille Landais, Alex Rees-Jones, Johannes Spinnewijn, Charlie Sprenger, Neil Thakral, and Dmitry Taubinsky for helpful discussion and comments. Canishk Naik and Baptiste Roux provided excellent research assistance.

"Blessed is he who expects nothing, for he shall never be disappointed." - Alexander Pope

1 Introduction

Reference-dependent preferences are a cornerstone of behavioral economics. In a vast array of settings, decision-makers apparently evaluate options relative to a reference point, and they evaluate losses relative to the reference point more strongly than equivalent gains - *loss aversion*. Early evidence of such behavior came from now-classic laboratory experiments by Tversky and Kahneman (1974). Since then, the experiments have been replicated and extended in myriad ways, in parallel with a rich theoretical literature seeking to model reference dependence (see O'Donoghue and Sprenger (2018) for a review). Researchers have found evidence of reference dependence in experiments around the world (Ruggeri et al., 2020), and in studies of observational data from diverse contexts including the hourly labor supply of cab drivers (Camerer et al., 1997; Crawford and Meng, 2011; Thakral and Tô, 2020b) and bicycle messengers (Fehr and Goette, 2007), job search (DellaVigna et al., 2017), behavioral responses to taxation (Homonoff, 2018; Rees-Jones, 2018), and the timing of retirement (Seibold, 2020).

As policymakers take notice of mounting evidence of the fundamental importance of reference dependence, thorny questions loom large. How should we evaluate welfare in the presence of reference dependence? What are the policy implications of all the evidence that reference dependence matters? As in other behavioral settings, the influence of reference points on behavior creates ambiguity over welfare. One possibility is that individuals simply have strange normative preferences, so that reference dependence and loss aversion should be viewed simply as a part of revealed preferences (Samuelson, 1938). Alternatively, one might suppose that these factors distort behavior relative to what is welfare-maximizing. Perhaps because of this difficulty, the vast literature to date on reference dependence has virtually entirely avoided welfare analysis.

This paper studies the welfare economics of reference dependence, overcoming the main challenge by explicitly embracing and analyzing the normative ambiguity inherent in the theory. We view the resolution of the ambiguity over welfare as a normative judgment that must be made by an observer or social planner, and we map such judgments to welfarist quantities. Specifically, we characterize the welfare impact of changes in the reference point, and of changes in prices (or, very similarly, taxes), under varying judgments about welfare. Goldin and Reck (2020) applied a similar approach to optimal default policy. We illustrate the findings in the retirement setting of Seibold (2020), where statutory retirement ages set by public policy influence individuals' reference points and implicit prices are given by financial retirement incentives.

Beyond normative ambiguity, a second challenge in analyzing welfare in the presence of reference dependence is the multitude of varieties of models of reference dependence. The basic approach of embracing normative ambiguity and mapping judgments to welfare quantities can be used in any of the popular models, but more complicated models introduce more nuanced normative ambiguities and they become more difficult to map to empirical data. Following much prior literature (see e.g. the discussion in O'Donoghue and Sprenger (2018)), our approach is to start with the simplest model capable of delivering some insight, and then develop extensions. Specifically, we start with a simple version of the model of Tversky and Kahneman (1991), where the only deviation from classical preferences comes from loss aversion, without additional frictions from reference dependence over gains or diminishing sensitivity. We consider this model in a static, non-stochastic setting, with reference dependence along a single dimension, i.e. over the consumption of a single good. These simplifications allow us to analyze welfare and develop our intuition in a simple and tractable way. We then extend the model in straightforward ways later on.

Within this model, we begin by analyzing the effect of a policy that influences the reference point, holding all else fixed.[1] We show that a change in the reference point has two potential first-order welfare effects; which of these matters depends on whether the planner judges that loss aversion is normative. If the planner judges that loss aversion is normative, the sole first-order welfare effect of a change in the reference point is the *direct effect* of a change in the reference point on gain-loss utility, holding behavior fixed. The effect of a change in the reference point on behavior in this case has no bearing on welfare due to the envelope theorem. In contrast, if the planner judges that loss aversion is not normative, then the sole first-order welfare effect comes from the *behavioral effect* of a change in the reference point.

Characterizing these direct and behavioral effects, we find that in simple, commonly used models of reference dependence, *decreasing the reference point is a robust Pareto improvement*: regardless of the view of welfare one takes, a reduction in the reference point is a Pareto improvement. When loss aversion is normative, the direct effect implies that loss averse agents are better off when the reference point decreases because they incur smaller losses. When loss aversion is not normative, individuals consuming at the reference point or in the loss domain are over-consuming the good to reduce their losses - i.e. there is a *negative internality* (Mullainathan et al., 2012). Decreasing the reference point decreses consumption, mitigating over-consumption.

Empirical evidence suggests that reference dependence also affects the behavioral response to price instruments, like taxes (e.g. Homonoff 2018; Rees-Jones 2018). Motivated by such evidence, we analyze the welfare effect of a price change in our model. Like a change in the reference point, a price change has straightforward direct and behavioral effects. Just as in standard models, a price increase has a direct, first-order, negative welfare effect, holding behavior fixed. Once again, when loss aversion is judged to be rational, the change in behavior has no first-order welfare implications. When loss aversion is causing over-consumption in the loss domain, however, the decrease in consumption caused by a price increase has a positive first-order behavioral welfare effect. This logic implies that when loss aversion is irrational, the optimal corrective tax would equal the amount of loss aversion in the loss domain and zero in the gain domain.[2]

We illustrate these theoretical results with an empirical application to retirement policy in Germany, building on Seibold (2020). The setting has two important advantages for our purposes. First, commonly used policies in the retirement context correspond closely to the types of interventions we analyze theoretically. On the one hand, pension systems typically feature statutory retirement ages such as a Normal Retirement Age. These age thresholds are framed as a "normal" time to retire and are perceived as reference points for individuals' retirement decisions. On the other hand, pension systems provide financial retirement incentives, influencing the marginal return to working longer or the implicit price of leisure. The second advantage of the empirical setting is that the relevant parameters governing individual behavior can be transparently estimated. In particular, we use high-quality administrative data on German retirees and exploit the bunching strategy of Seibold (2020) in order to estimate the responsiveness of these individuals to financial incentives and to the Normal Retirement Age as a reference point.

We simulate the effects of two types of pension reforms often discussed as policy options to induce workers to postpone retirement. The first reform is an increase in the Normal Retirement Age by one year.

[1]The origins of reference points are a subject of much discussion in the literature. We discuss this issue in detail in Section 2.2.1.

[2]We note that this conclusion relies on our assumption that choices in the gain domain are normative, i.e. that individuals are over-valuing the mitigation of losses and not under-valuing the intensification of gains. See also footnote 6 below. Caveats involving heterogeneity are also discussed below.

This reform increases the reference age of retirement, or equivalently lowers the reference point in terms of leisure, the corresponding good. We show that in addition to the positive fiscal effects it entails, such a reform increases overall welfare regardless of the planner's normative judgment of reference dependence. If reference dependence is judged as a bias, a lower reference point in terms of lifetime leisure counteracts some of the initial sub-optimal early retirement behavior, bringing individuals closer to their optimal retirement age. If reference dependence is judged to be rational, a lower reference point yields additional direct welfare gains, as individuals compare their lifetime leisure more favorably to the higher Normal Retirement Age.

The second reform we consider is an increase in the Delayed Retirement Credit, that is higher actuarial pension adjustment for working beyond the Normal Retirement Age. Again, this type of reform is the subject of real-world policy debates around incentivizing individuals to work longer, but it also relates directly to our theoretical results. In particular, a higher Delayed Retirement Credit increases the marginal return to working, implying a higher implicit price of leisure in the loss domain above the Normal Retirement Age. We show that such a policy can improve welfare when reference dependence is judged as a bias, since incentivizing workers to retire later reduces the sub-optimality from early retirement due to reference dependence. In fact, it is possible to find an optimal level of corrective subsidies for later retirement in this case. Due to the relatively strong estimated reference dependence in our setting, such a subsidy would have to be very large in order to sufficiently correct behavior, and it would be fiscally costly. If reference dependence is judged to be rational, on the other hand, the welfare effects of the Delayed Retirement Credit are much more muted. Moderate actuarial adjustment can help correct fiscal externalities in the pension system, while an overly large Delayed Retirement Credit can distort retirement behavior, reduce the fiscal balance of the pension system and ultimately lower welfare.

Finally, we return to the theory and extend it in a number of ways, building on the simple results employing loss aversion only. First, an implication of the theory as it is often written (Tversky and Kahneman, 1991; Köszegi and Rabin, 2006) is that reference dependence may affect decisions over gains, by giving the individual extra utility from comparing the option to the reference point even in the gain domain.[3] Although reference dependence is often formulated in this way in the literature, including reference dependence over gains is not necessary to explain the patterns in behavioral data typically attributed to reference dependence.[4] However, we show that including reference dependence over gains in the model does affect welfare calculations, depending on whether it is judged to be normative. The main changes are that 1) there will now be potential negative internalities in the gain domain, and 2) increasing the reference point can have a direct negative welfare effect on individuals in the gain domain if these frictions are normative. Decreasing the reference point remains a robust Pareto improvement - in fact the welfare gains from lower reference points would be larger if we adopt this formulation.

We next consider the dimensionality of reference dependence. To fix ideas by way of an example, in our empirical application one could suppose that there is reference dependence over labor supply/leisure or over consumption or over both, as some previous work on reference dependence and labor supply points out (Crawford and Meng, 2011). We justify our initial decision to include only the leisure dimension in our empirical application by studying the pattern of bunching around the Normal Retirement Age more closely. The empirical shape of the retirement age distribution is well in line with reference dependence in the labor supply/leisure dimension, but less well in line with alternative specifications. However, in other contexts, two- or multi-dimensional models may be more appropriate. We sketch a way forward

[3] A similar friction is sometimes studied without loss aversion in the literature on relative social comparisons, see e.g. Gali (1994).

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

for normative analysis in such contexts by deriving our basic social welfare effects of interest in a two-dimensional model of reference dependence.

A third extension considers the deliberate setting of reference points by individuals. Our findings up to this point suggest that individuals will generally prefer low reference points, but in some contexts individuals may set high reference points out of some other concern, like a separate behavioral bias they wish to overcome. We explicitly study this in a model of goal-setting like Koch and Nafziger (2011). Unsurprisingly, when individuals optimally set their own reference points, inducing them to set a higher or lower reference point imposes no first-order welfare effects, due to the envelope theorem, and a second-order welfare loss. Whether individuals do optimally set their own reference point turns on the question of whether they are sophisticated in their knowledge of their own biases (DellaVigna and Malmendier, 2006). This type of reasoning does not seem applicable to our empirical application, because we observe policy changes influencing reference points in a fashion inconsistent with individuals' deliberately and optimally setting their own reference points. Nevertheless, in other settings where e.g. individuals use reference points to exert self control, this type of model may be applicable.

We finally discuss a number of considerations for future work based on possible extensions of the model. These include adding "diminishing sensitivity" to the model, studying reference dependence under risk and uncertainty, and the question of "narrow bracketing" (Tversky and Kahneman, 1981). A common theme of many of these possible extensions is that they are technically feasible, in that one can write down the model and characterize welfare, but they are also difficult to implement empirically, especially outside of highly stylized experiments.

The remainder of this paper proceeds as follows. In Section 2, we introduce the basic model and derive our main theoretical results, Section 3 presents the empirical application to retirement behavior, Section 4 discusses extensions of the theory, and Section 5 concludes.

2 Simple Model

In this section, we lay out a simple model of reference dependence and characterize behavior and welfare within this model. As described above, the modern theory of reference dependence is in reality a family of theories. We will focus here on the simplest model that delivers useful insights that are applicable in many empirical contexts. We will then enrich the theory in several ways in Section 4.

2.1 Setup

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

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

As it generates behavior, Kahneman et al. (1997) would call $U(x, y)$ *decision utility*, to distinguish it from *experienced utility*, or welfare. We assume an interior solution, and that $u_i' > 0$ and $u_i'' < 0$ always.

Empirical patterns in observed choices documented in the literature suggests that reference dependence generates a kink in marginal utility over x at the reference point r, so that losses incur a penalty on the margin relative to gains.[5] In our starting model, the only deviation from classical models we adopt is the modification to decision utility necessary to rationalize such an empirical observation. We therefore specify the deviation from the classical model in the $v_i(x|r)$ term as follows:

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

The parameter $\Lambda_i > 0$ governs the extent of *loss aversion*, the size of the penalty that losses incur relative to gains on the margin.[6] The domain of x where $x > r$ is called the *gain domain*, and the domain where $x < r$ is called the *loss domain*.

The model described by Equations (1) and (2) is slightly different from that proposed canonically by Tversky and Kahneman (1991).[7] The simplification we use here is common in the literature and immaterial for explaining behavior, but it may matter for welfare. The Tversky and Kahneman (1991) formulation is

$$v_i(x, r) = \begin{cases} \eta_i(x - r) & x > r \\ \eta_i\lambda_i(x - r) & x \leq r, \end{cases} \tag{3}$$

where η_i governs the relative importance of reference dependence overall, and λ_i governs loss aversion. With reference dependence over just one good x, this model is behaviorally indistinguishable from the model in Equation (2). We show this formally in Appendix C; Barseghyan et al. (2013) demonstrate a similar equivalence with a focus on the stochastic case.

The presence and size of the η_i parameter in the Tversky and Kahneman (1991) formulation is seldom if ever analyzed empirically. As such, our simple model focuses on the case where the main friction in the model is loss aversion, with no reference-dependence effects on demand in the gain domain. In revealed preference terms, this means we are assuming that choices in the gain domain are deemed "welfare relevant," while choices in the loss domain are more suspect (Bernheim and Rangel, 2009) - see Appendix D for further explanation.[8] We consider the question of whether reference dependence in the gain domain might justify an additional deviation from revealed preference in Section 4.1. We also disregard *diminishing sensitivity*, which would require that $v_i'' > 0$ in the loss domain and $v_i'' < 0$ in the gain domain. We discuss this further in Section 4.4.

[5]There have been some suggestions in the literature of a notch in utility rather than a kink, see e.g. Allen et al. (2017). The kink formulation is more commonly adopted, so we adopt it here, and defer consideration of alternative forms of reference dependence to future work. Moreover in our empirical application below, a kink fits the observed bunching better than a notch, which one would expect to generate noticeable "missing mass."

[6] We assume that gain-loss utility affects decision utility in the loss domain. An alternative would be to keep loss-domain demand to zero in $v(.)$ and include a term in the gain domain with negative marginal utility. Such a specification would be equivalent for predicting behavior but different for welfare. Nevertheless we believe this specification is the correct one for welfare analysis because a sizable literature in psychology and neuroeconomics suggests that loss aversion is driven by a negative emotional response to the incursion of perceived losses. We review this literature in detail below. For one particularly compelling piece of evidence on this question, suggesting that reference dependence affects loss payoffs and not gain payoffs, see Figure 4 of Sokol-Hessner et al. (2013).

[7]In Tversky and Kahneman (1991), gain-loss utility is posited as the sole component of preferences, while later applications incorporated non-reference-dependent concerns as we do in the first part of equation 1 (see e.g. Köszegi and Rabin (2006); O'Donoghue and Sprenger (2018)).

[8]Appendix D also contains a thorough accounting of how we can fit our analysis into the general behavioral revealed preference framework of Bernheim and Rangel (2009).

We now characterize demand $x_i(p, r)$ in the simple model. Following our assumption of quasi-linear preferences, we suppress z_i as an input to demand and other functions. We first characterize two potentially optimal choices as follows:

$$u'(x_i^G(p)) = p, \tag{4}$$

$$u'(x_i^L(p)) + \Lambda_i = p. \tag{5}$$

Because $u_i'' < 0$ and $\Lambda_i > 0$, we know that for any i and any p, $x_i^G(p) < x_i^L(p)$. Demand for a given individual will therefore be

$$x_i(p, r) = \begin{cases} x_i^G(p), \text{ if } x_i^G(p) > r & (G) \\ x_i^L(p), \text{ if } x_i^L(p) < r & (L) \\ r, \qquad \text{otherwise.} & (R) \end{cases} \tag{6}$$

At any given price and reference point in this model, there are three groups of individuals, group G in the gain domain, group L in the loss domain, and group R at the reference point. Furthermore, we know that for individuals in group R, $x_i^G(p) \leq r \leq x_i^L(p)$.

Individual Welfare. The planner must judge whether reference-dependent decision utility should be given normative weight, i.e. whether to respect loss aversion or regard it as a bias. We parametrize this decision by $\pi \in \{0, 1\}$, where $\pi = 1$ if the planner respects loss aversion. We focus on the cases where $\pi = 0$ and $\pi = 1$ for clarity, but the analytic expressions we derive below could be evaluated for intermediate values of $\pi \in [0, 1]$ as well, as in Goldin and Reck (2020). We express normative preferences according to

$$U_i^*(x, y) = u_i(x) + y + \pi v_i(x|r), \tag{7}$$

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

$$w_i(p, r) \equiv U_i^*(x_i(p, r), z_i - px_i(p, r)). \tag{8}$$

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

Mechanisms and the Correct Value of π. The psychological mechanisms behind loss aversion has some bearing on the question of what judgments planners should make about π. Broadly speaking, what evidence we have suggests that loss aversion has emotional origins (see Rick (2011) for a review). The findings of an influential study by Kermer et al. (2006) suggested that loss aversion derives from an *affective forecasting error*: people wrongly project that they will experience emotional pain if they incur a loss, so they try to avoid losses. In this case, it is arguably appropriate to set $\pi = 0$. However, more recent evidence suggests that the emotional pain of incurring losses is real rather than a forecasting error and that emotional regulation strategies mitigated loss aversion (Sokol-Hessner et al., 2009). This notion has been further borne out by neurological evidence associating activity in the amygdala with loss aversion and the incursion of perceived losses in a number of ways (De Martino et al., 2010; Sokol-Hessner et al., 2013; Sokol-Hessner and Rutledge, 2019). If we accept this premise, the question of π becomes a deeper question about whether individuals should let negative emotions influence their choices, or whether individuals should make decisions dispassionately. Understanding the mechanisms at play here can lead to an interesting philosophical discussion, but it does not resolve the normative ambiguity over π. So henceforth we remain totally agnostic about the value of π.

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

$$W(p,r) = \int_i w_i(p,r)di. \tag{9}$$

Due to our assumption of quasi-linear preferences, maximizing this social welfare function is equivalent to maximizing the sum of compensating or equivalent variation, relative to any arbitrary benchmark. One could relax the assumption of strict utilitarian preferences and quasi-linearity with a straightforward application of Saez and Stantcheva (2016), or with a more sophisticated approach. Likewise, we do not address distributional concerns here, but defer them to future work.

2.2 Results

This section lays out the main theoretical results of the paper for the simple model. We begin with an intuition-building characterization of welfare, and then derive the effects of references points and prices on welfare. Finally, to help us understand the fiscal externality in our empirical application, we discuss additional social welfare effects in the presence of an externality.

The Marginal Internality. A key statistic for welfare analysis is the *marginal internality* (Mullainathan et al., 2012; Allcott and Taubinsky, 2015; Allcott et al., 2019). In our context, this is the welfare effect of a marginal change in x along the budget constraint, evaluated at observed demand. Using the first-order conditions in Equations (4) and (5) and the behavioral characterization in (6), it is straightforward to derive the following:

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

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

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

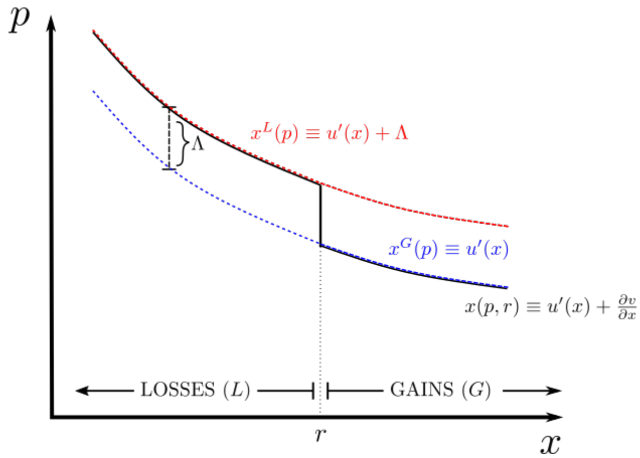
L1.3. *If* $x_i(p,r) = r$,

- $m_i(p,r)$ *is undefined when* $\pi = 1$.
- $m_i(p,r) = u_i'(r) - p$ *when* $\pi = 0$, *with* $-\Lambda_i \leq m_i \leq 0$

We interpret the marginal internality as the marginal welfare effect of paternalistically inducing the consumer to choose a little bit more of good x, starting from observed demand. When the planner judges that observed demand is welfare-maximizing, $\pi = 1$ and there are no marginal internalities as a consequence of the envelope theorem. The marginal internality is undefined when $x = r$ in this case because of the kink in utility at $x = r$, but it remains the case that no induced change in behavior could improve welfare. When $\pi = 0$, in contrast, individuals consuming $x_i \leq r$ are over-consuming good x out of loss aversion. Forcing individuals to consume more x would harm them, so the marginal internality is negative.

Building on Lemma 1 and our characterization of behavior, Figure 1 plots individual demand for good x, which coincides with welfare-maximizing demand when $\pi = 1$, along with $u_i'(x)$, which coincides with welfare-maximizing demand when $\pi = 0$. We also plot demand according to $x^G(p)$ and $x^L(p)$, as defined in Equations (4) and (5), for illustration. Given the judgment by the planner about what is welfare-maximizing, the marginal internality is the vertical difference between observed demand and welfare-maximizing demand in Figure 1.

FIGURE 1: OBSERVED DEMAND, WELFARE-MAXIMIZING DEMAND, AND MARGINAL INTERNALITIES



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

2.2.1 Reference Point Effects

We next characterize the welfare effect of a discrete or marginal change in the reference point.

Can We Change Reference Points? In our model the reference point is exogenous and known to the observer or social planner, but the origin of reference points is the subject of much discussion in the literature. In the design of their early experiments, Kahneman and Tversky apparently viewed reference points as exogenous features of the environment by which gains versus losses were evaluated by decision-makers. The psychological literature suggests that reference points may have multiple cognitive origins, including salient options or features of options (Rosch, 1975) and goals set by the self or others (Heath et al., 1999). Recent economic theory, in contrast, attempts to systematically endogenize reference points in terms of beliefs or expectations, though much of this work focuses on the uncertainty case (see e.g. Köszegi and Rabin, 2006; 2007).

We argue that characterizing the welfare effect of a change in the reference point is important regardless of the precise origins of the reference point in a given context. For any policy P that affects the reference point, we may wish to know the effect of this policy on welfare, which expressing social welfare as a function of P and a reference point $r(P)$ would be $\frac{dW}{dP} = \frac{\partial W}{\partial P} + \frac{\partial W}{\partial r}\frac{\partial r}{\partial P}$. The first term captures any direct effect of the policy on welfare, while the second effect captures the welfare effect that this policy has because it changes the reference point. Our analysis here will characterize the $\frac{\partial W}{\partial r}$ term, abstracting away from other effects policy changes might have.

In some practical settings, it appears to be the case that policymakers can in fact shift reference points, consistent with psychological theories positing that salient features of the environment can induce reference dependence. For example, Seibold (2020) found significant reference dependence in relation to statutory retirement ages that can be changed by pension policy. Given compelling evidence that the behavioral response to changes in these statutory ages exhibits all the signature characteristics of a change in the reference point, we can therefore ask what would happen to welfare if the government changed the Normal Retirement Age, for instance, which is our empirical application below. Some other examples we cited above also suggest that policy can shift reference points. Homonoff (2018) found that framing a plastic bag tax as a tax penalty rather than a discount for re-usable bag use had a large positive effect on the use of re-usable bags. We can interpret this effect as coming from a change in the reference point whereby consumers evaluate the discount in the gain domain and the tax in the loss domain. Likewise, income tax withholding rates can be manipulated by policy and they appear to create a reference point at zero tax due at the time of tax filing (Rees-Jones, 2018).

Consistent with expectations-based origins of reference points, there is experimental evidence that changing expectations can change reference points (Abeler et al., 2011; Ericson and Fuster, 2011). The literature has not yet settled on the best way to model expectations-based reference points. Doing so is a challenging theoretical problem (Masatlioglu and Raymond, 2016), and experimental evidence suggests that simple models of expectations-based reference points have limited explanatory power (Gneezy et al., 2017; Goette et al., 2020). Relatedly, DellaVigna et al. (2017) and Thakral and Tô (2020b) use theory and evidence to suggest that past experiences can influence reference points, but whether this is driven by belief updating or some other cognitive process is not entirely clear. In any case, a policy that changes expectations or some similar determinant of a reference point might have effects on welfare that are not driven by reference dependence itself, as above, but in order to characterize the full welfare effect we would still need to consider the effect on welfare through the reference dependence channel, which we do here.

Generic changes in the reference point Let r_1 and r_0 denote two arbitrary reference points, where $r_1 > r_0$ without loss of generality. Based on (6), we know that there are three cases to consider under a given reference point. It is straightforward to show that $x(p, r_1) \geq x(p, r_0)$, so we have 6 potential cases for behavior under these two reference points. Let group GR be the set of individuals for whom $x_i(p, r_0) = x_i^G$ and $x_i(p, r_1) = r_1$, and define the other groups analogously. The change in individual welfare from a change in the reference point is:

$$w_i(p, r_1) - w_i(p, r_0) = \begin{cases} 0, \ i \in GG \\ u_i(r_1) - u_i(x_i^G) - p(r_1 - x_i^G), \ i \in GR \\ u_i(x_i^L) - u_i(x_i^G) - p(x_i^L - x_i^G) + \pi\Lambda_i(x^L - r_1), \ i \in GL \\ u_i(r_1) - u_i(r_0) - p(r_1 - r_0), \ i \in RR \\ u_i(x^L) - u_i(r_0) - p(x^L - r_0) + \pi\Lambda(x_i^L - r_0), \ i \in RL \\ -\pi\Lambda_i(r_1 - r_0), \ i \in LL \end{cases} \tag{10}$$

The effect of the change in the reference point on social welfare will be the summation of the individual change in welfare across all six of these groups. We denote the probability of an individual being in a given group by, e.g., $P[i \in GG]$. Signing the welfare effects in all six cases yields the following result.

Proposition 1. ***The Desirability of Low Reference Points.*** *Consider a change from r_0 to $r_1 > r_0$. Denote the effect on individual and social welfare given the planners' judgments by $\Delta w_i(\pi) \equiv w_i(p, r_1) - w_i(p, r_0)$ and $\Delta W(\pi) \equiv W(p, r_1) - W(p, r_0)$.*

P1.1. *For any i and any $\pi \in \{0, 1\}$, $\Delta w_i(\pi) \leq 0$.*

P1.2. *If $P[i \in GG] < 1$, then $\Delta W(1) < 0$, and r_0 Pareto dominates r_1 for $\pi = 1$.*

P1.3. *If $P[i \in GG] + P[i \in LL] < 1$, then $\Delta W(0) < 0$ and r_0 Pareto dominates r_1 for $\pi = 0$.*

Proposition 1 implies that in the model we have laid out so far, notably including no externalities, a benevolent social planner would always prefer to decrease the reference point, holding all else fixed. So long as the reference point has any non-trivial effect on behavior, decreases in the reference point are robust Pareto improvements, meaning that they are Pareto improvements for any value of π. In Appendix D, we show that this notion of a robust improvement in welfare is closely related to the welfare criterion proposed by Bernheim and Rangel (2009); in other words, we find that their welfare criterion also suggests that lower reference points are welfare-improving. A full proof of Proposition 1 is provided in the Appendix.

The basic intuition of Proposition 1 is as follows. A change in the reference point has two effects: a direct effect and a behavioral effect. The direct effect comes from the fact that in the loss domain, $v_i(x|r)$ is decreasing in r. This implies that when $\pi = 1$, increasing the reference point decreases welfare holding behavior fixed. The behavioral effect comes from the impact of changing the reference point. As increasing the reference point tends to increase x and, as discussed above, individuals tend to over-consume in this model, the behavioral effect tends to decrease welfare, particularly when $\pi = 0$ so that over-consumption occurs.

To shed further light on these two effects, we next consider a first-order approximation of Equation (10), that is, the effect of a marginal change in r.

Proposition 2. ***First-Order Individual Welfare Effect of a Change in the Reference Point.*** *Consider a change in the reference point that is small enough that the GL group is negligible. Let $\Delta r = r_1 - r_0$, and let $\Delta x_i = x_i(p, r_1) - x_i(p, r_0)$. The first-order individual welfare effect of this change in the reference point is approximately:*

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

To understand how these decompose into the behavioral and direct effects, consider that in every case the change in welfare is approximately

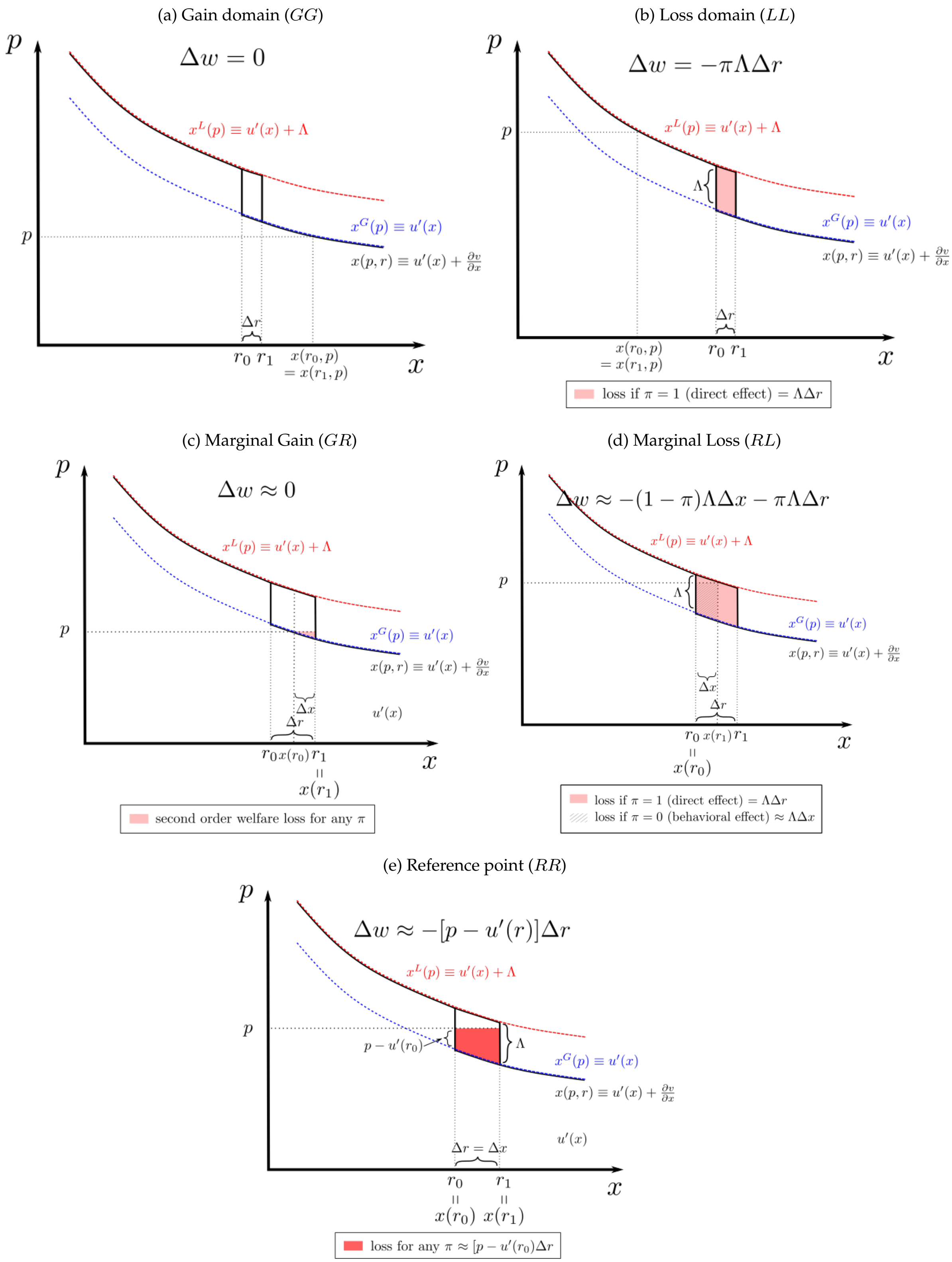
$$\Delta w_i \approx m_i(\pi)\Delta x + \pi \left.\frac{\partial v_i}{\partial r}\right|_{x=x_i(p,r)} \Delta r \tag{12}$$

In the GG and LL cases, $\Delta x = 0$, so the behavioral effect vanishes. The direct effect is zero in the GG case because $\left.\frac{\partial v_i}{\partial r}\right|_{x=x_i(p,r)} = 0$, but it equals $-\pi\Lambda_i \Delta r$ in the LL case. In the GR case, the direct effect similarly vanishes and $m_i = 0$ causes the behavioral effect to vanish. In the LR group we can explicitly see both effects in equation (11). The RR group is more complicated because m_i may not be defined in this case, which invalidates the approximation in equation (12), but in the case where $\pi = 0$ we observe the behavioral effect in equation (11).

Figure 2 depicts the welfare effect through the behavioral response for the many different cases, building on Figure 1. In every case, we integrate the marginal internality over the the change in consumption to obtain the effect of the change in behavior on welfare. The direct effect is visualized (in the cases where $\pi = 1$ and it is present) by noting that the vertical distance between the $u_i'(x)$ and the observed demand curve equals $\frac{\partial v_i}{\partial x} = -\frac{\partial v_i}{\partial r}$.

We now turn from individual to social welfare for a change in the reference point. Because welfare and behavior are continuous in this model, the change in social welfare from the GR and RL groups is

FIGURE 2: WELFARE AND CHANGES IN THE REFERENCE POINT



Note: This figure plots the welfare effect of changing the reference point in each domain. We denote observed demand in black and gain and loss domain demand in blue and red, respectively, as in Figure 1. All welfare changes here are losses, shaded in red, reflecting the main result that increasing the reference point unambiguously decreases welfare. Welfare losses from the direct effect are depicted in light red shaded regions, while losses due to the behavioral effect are shaded with diagonal hatching. In panel (e), we find that the change in welfare in the RR case is the same regardless of π, but whether the depicted welfare loss actually represents a behavioral welfare effect or a direct welfare effect does depend on π, so we use dark red shading.

second-order - a marginal change in welfare for a marginal group. Building on Proposition 2, we obtain the following result.

Proposition 3. ***The First-Order Social Welfare Effect of a Change in the Reference Point.*** *Starting from any given price and an initial reference point, define groups G, L and R. The effect of a small change in the reference point of Δr on social welfare is approximately*

$$\begin{aligned} \Delta W \approx & -\Delta r \pi E[\Lambda_i \mid i \in L] P[i \in L] \\ & -\Delta r E[p - u_i'(r) \mid i \in R] P[i \in R]. \end{aligned} \tag{13}$$

Proposition 3 underscores the results in Proposition 1, that social welfare is robustly decreasing in the reference point but the magnitude of the welfare effect of a decrease in the reference point depends on normative judgments. We knew from Proposition 1 that policy should always seek to decrease reference points in this model, but when we add other motives, such as an externality below, Proposition 3 helps us understand how the planner should balance externality concerns with these private welfare concerns, and how that trade-off depends on normative judgments.

2.2.2 Price Effects

We next consider the effect of a change in price on welfare. For simplicity, we will formally derive the first-order welfare effects, but we illustrate the exact welfare effects in our figures.

Observed demand in Figure 1 is negatively inclined, and it is straightforward to show formally that for a price change from p_0 to p_1 with $p_1 > p_0$, $x_i(p_1, r) \leq x_i(p_0, r)$. As with a change in the reference point, there are six potential cases in general, which we denote using similar notation to before as GG, GR, GL, RR, RL and LL. The characterization of these groups is obviously different here because we are comparing different values of p rather than r.

Proposition 4. ***First-Order Welfare Effect of a Change in Price.*** *Consider a change in price that is small enough that the GL group is negligible. Let $\Delta p = p_1 - p_0$, and let $\Delta x_i = x_i(p_1, r) - x_i(p_0, r)$. The first-order welfare effect of a change in price is approximately:*

$$w_i(p_1, r) - w_i(p_0, r) \approx m_i(\hat{p}, r)\Delta x - x_i(\hat{p}, r)\Delta p, \tag{14}$$

where $m_i(\hat{p}, r)$ is defined as in Lemma 1, $\hat{p} = p_0$ if $i \in GG, GR$, and $\hat{p} = p_1$ if $i \in LL, RL$. For $i \in RR$, $\hat{p}$ can be either price for this first-order approximation, and the first term in (14) is zero even when m_i is undefined - in this group $\Delta x = 0$.

A price change has two first-order effects in this model, which are illustrated in Figure 3. First, there is a conventional direct, mechanical effect, analogous to the first-order decline in consumer surplus that would occur in a classical model. To value this we simply multiply the change in price by observed demand. Second, there is a change in the internality, which we value by multiplying the relevant marginal internality from Lemma 1 by the change in demand. In the case where demand does not change because the individual sticks at the reference point before and after the price change, this term is absent. When $\pi = 1$, the marginal internality is always 0 and the envelope theorem once again implies the change in x has no first-order welfare effect. In this case, increases in price strictly decrease welfare. When $\pi = 0$, however, the individual over-consumes in the loss domain, and inducing the individual to consume less via a higher price can improve welfare by counteracting the internality. In this case, whether the price change improves or harms individual welfare is ambiguous.

As before, the marginal groups are second order when we turn to the first-order social welfare effect of a change in price. We obtain the following characterization of the first-order welfare effect of a change in price:

Proposition 5. ***The First-Order Social Welfare Effect of a Price Change.*** *Starting from any given reference point r_0 and an initial price p_0, define groups G, L and R. The effect of a small price change Δp on social welfare is approximately*

$$\Delta W \approx \left\{ E[-(1-\pi)\Lambda_i \frac{\partial x_i^L}{\partial p} \Delta p \mid i \in L] \right\} P[i \in L] - E[x_i(p_0, r_0)]\Delta p, \tag{15}$$

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

Corollary 5.1. ***Corrective Taxes for Reference Dependence.*** *The efficient non-linear, person-specific tax on x in the model given a reference point, $t_i(x, r)$, is*

$$t_i(x, r) = \begin{cases} -(1-\pi)\Lambda_i(x-r), & x < r \\ 0 & x \geq r \end{cases} \tag{16}$$

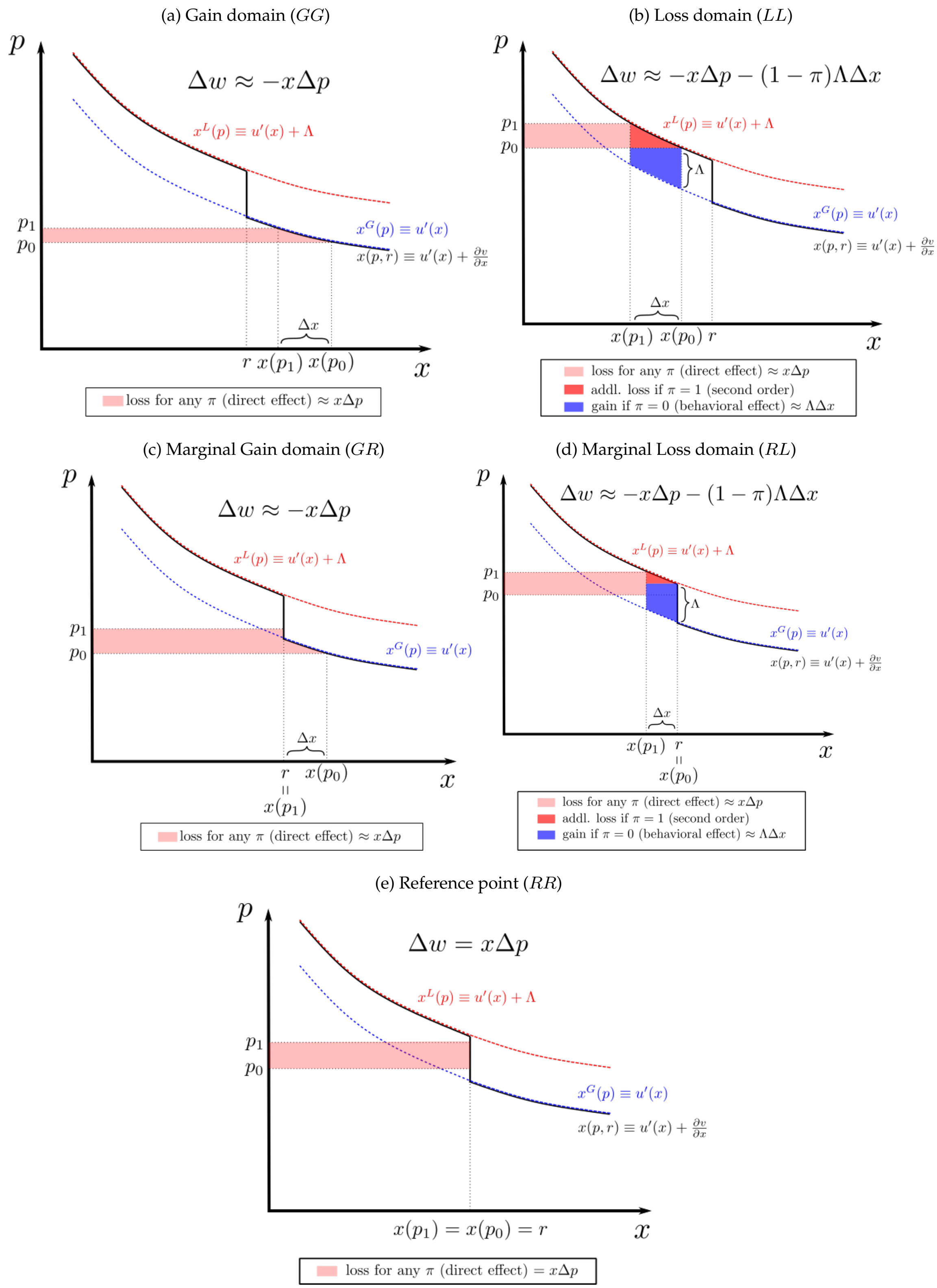
The first term in Equation 15 comes from the negative internality in L group from Lemma 1. The final term is the conventional direct first-order effect of a price change, i.e. the mechanical effect on total expenditure. As this final term is equal to the effect on tax revenues from the introduction of tax, Proposition 3 has straightforward implications for optimal corrective taxes in the presence of reference dependence, which we express in Corollary 5.1.

As in other contexts, the optimal corrective tax tends to equal the marginal internality. In this case, the marginal internality is non-linear and only nonzero in the loss domain, so the optimal corrective tax has these properties as well. And naturally, when $\pi = 1$, there are no internalities to correct and the optimal corrective tax is zero. For simplicity, we presume the planner knows Λ_i for each individual. This assumption is obviously unrealistic if Λ_i is highly heterogeneous. In this case, one would need to account for the covariance between the demand elasticity and the internality embodied by Λ_i to set an optimal corrective tax (see e.g. Allcott and Taubinsky (2015); Allcott et al. (2019)). Likewise, we presume the government knows r, which precludes cases where some individuals use different reference points than others in a given situation, and the planner does not know each individual's reference point. We defer further consideration of these issue to future work.

Adding Externalities. In many settings, including our empirical setting, we may wish to incorporate a fiscal or other externality into our welfare framework. For a simple linear atmospheric externality equal to $\alpha E[x_i]$, the social welfare effect of either a change in price or a change in the reference point will be the social welfare effect described above plus $\alpha E[\Delta x_i]$, i.e. the marginal externality times the change in aggregate demand for good x.

Consider, for example, the plastic bag incentives studied by Homonoff (2018). Framing the incentive as a tax loss rather than a bonus discount effectively raises the reference point, so that the cost of using a plastic bag is evaluated in the loss domain. Without an externality, our results above suggest that this intervention

FIGURE 3: WELFARE AND PRICE CHANGES



Note: This figure plots the welfare effect of a price change in each domain. We denote observed demand in black and gain and loss domain demand in blue and red, respectively, as in Figure 1. The direct, mechanical negative effect of a price change is depicted with red shaded regions. The positive behavioral welfare effect attributable to internalities, the normative significance of which depends on π, is depicted with blue shaded regions.

would decrease welfare, especially when $\pi = 1$. However, with a large enough negative externality for plastic bag use, which of course was the motivation for introducing the incentive to begin with, the policy implication could go in the opposite direction, in favor of the loss framing. We can further infer from equation (13) that the size of the externality needed to justify the loss framing is larger when $\pi = 1$ than when $\pi = 0$, which reflects that in the $\pi = 1$ case changing the reference point has a direct welfare effect.

3 Empirical Application: Reference Dependence in Retirement Behavior and the Welfare Effects of Pension Reforms

In this section, we present an empirical application of our theoretical results. Retirement behavior is arguably one of the most important empirical contexts in which reference-dependent preferences have been documented in recent literature (Behaghel and Blau, 2012; Seibold, 2020). Our empirical setting is that of Seibold (2020), who documents large bunching in the retirement distribution around *statutory retirement ages* in Germany and argues that this phenomenon can be explained by workers perceiving those ages as reference points in their retirement decision. In this context, our goal is to characterize the welfare effects of changes to the Normal Retirement Age, and of financial incentives for delayed retirement. These policy reforms are closely related to the welfare effects analyzed in previous sections, and to the types of pension reforms often debated in practice.

3.1 Institutional Setting and Data

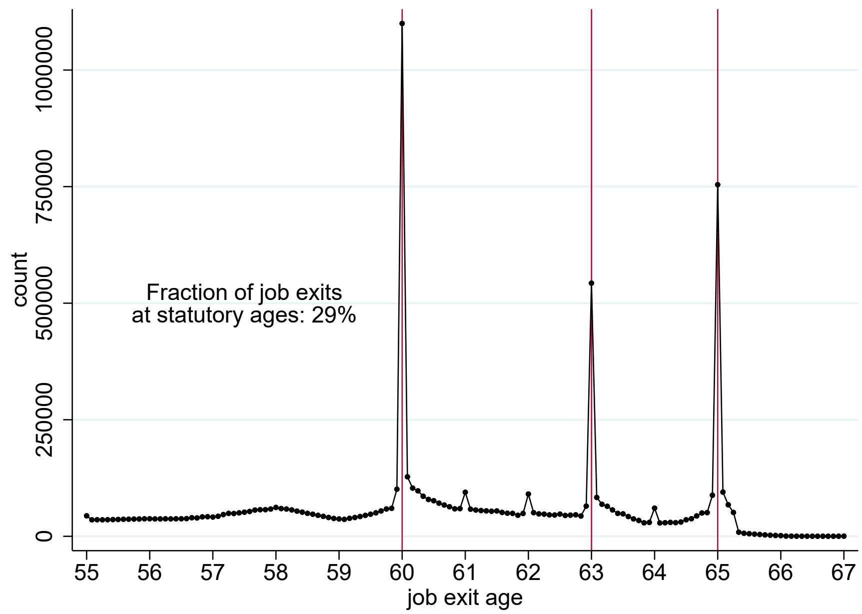
Germany has a pay-as-you-go pension system sharing many of its key characteristics with public pension systems in other developed countries. The vast majority of German workers are covered by public pensions, as enrollment is mandatory for all private-sector employees. Pension contributions are levied as a payroll tax on gross earnings. Benefits are defined according to a pension formula based on a worker's lifetime contribution history. Pension benefits are roughly proportional to lifetime income and there is relatively little redistribution. The average net replacement rate is just over 50% (OECD, 2019), and public pensions are the main source of income for most recipients.

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

The NRA is the most salient and latest statutory retirement age, but there are others in the system. In addition to the NRA, there is a Full Retirement Age (FRA) from which a "full" pension is available. For most workers in the birth cohorts considered in our analysis, the Normal and Full Retirement Ages coincide, but they can differ for some. Thirdly, the pension system has an Early Retirement Age (ERA), the earliest age from which a pension can be claimed, which we do not analyze directly. Overall, statutory retirement ages induce strong retirement responses. Figure 4 shows the retirement age distribution among German workers born between 1933 and 1949, among which 29% retire exactly in the month when they reach a statutory retirement age.[9]

[9] The precise location of all three statutory retirement ages differs across workers, for example based on birth cohort and gender, but

FIGURE 4: RETIREMENT AGE DISTRIBUTION IN GERMANY



Note: The figure shows the retirement (job exit) age distribution among German workers born between 1933 and 1949. The vertical red lines indicate the main locations of statutory retirement ages faced by these workers. Source: Seibold (2020).

The second key policy dimension are financial retirement incentives. Similarly to other pension systems, there is actuarial adjustment of pension benefits as a function of an individual's retirement age. Hence, when a worker chooses to retire later, there is an explicit upward adjustment of pension benefits in addition to the increase in their baseline pension due to additional contributions. In Germany, actuarial adjustment is relatively low, however. Pensions increase by 3.6% per year of later retirement below the FRA, and there is no explicit adjustment between the FRA and the NRA, should they differ for a worker. The largest actuarial adjustment occurs above the NRA, where a *Delayed Retirement Credit* of 6% per year applies.

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

In the empirical analysis, we use the data set of Seibold (2020), which is based on administrative data covering the universe of German retirees who claim a public pension between 1992 and 2014 provided by the German State Pension Fund (Forschungsdatenzentrum der Rentenversicherung (FDZ-RV), 2015).[10] We

the main locations are 60, 63 and 65. See Seibold (2020) for more details of the institutional setting.

[10]Due to current logistical constraints, the results shown in this version of the paper had to be obtained from a 1% sample. Final results based on the full data will likely differ slightly.

apply the same sample restrictions as Seibold (2020) and additionally restrict the sample to birth cohort 1946. The main reason to focus on one birth cohort is to simplify the analysis, as different cohorts face different statutory retirement ages and benefit schedules due to various cohort-based pension reforms.

3.2 Model and Parameter Estimation

From our theoretical analysis, the most important factors for welfare analysis are the strength of reference dependence (Λ above), the number of individuals in the G, R, L groups, and the behavioral response to price changes. We capture these three components parsimoniously in a static model of retirement behavior with reference dependence, as in Seibold (2020). Preferences of a reference-dependent agent are[11]

$$U_i(C,R) = C - \frac{n_i}{1+\frac{1}{\varepsilon}}\left(\frac{R}{n_i}\right)^{1+\frac{1}{\varepsilon}} - \begin{cases} 0 & R < \hat{R} \\ \widetilde{\Lambda}(R-\hat{R}) & R \geq \hat{R}, \end{cases} \tag{17}$$

where C is lifetime consumption, R is the worker's retirement age relative to a career starting age normalized to 0. The parameter $\widetilde{\Lambda}$ captures the strength of reference dependence. The heterogeneous parameter n_i reflects earnings ability at old age, where low ability increases disutility from postponing retirement; for our purposes the distribution of n_i will determine whether an individual is in the $G, R,$ or L group in any given situation. The parameter ε is the elasticity of the retirement age with respect to the implicit net-of-tax rate, which is the relevant elasticity to price changes for our context. Loosely speaking, $\widetilde{\Lambda}$ can be identified by the behavioral response to statutory retirement ages holding financial incentives fixed, ε can be identified by the behavioral response to financial incentives, and the distribution of n_i is specified to fit the variation in retirement ages.

Importantly, equation (17) assumes *loss aversion in lifetime leisure*. The last term in this equation captures reference dependence as in the simple model presented in Section 2. Intuitively, our formulation implies that marginal disutility from increasing labor supply beyond the retirement reference point $\hat{R}$ is greater than marginal disutility from approaching $\hat{R}$ from the left, and the parameter $\widetilde{\Lambda}$ determines the size of this kink in the utility function. Such reference dependence in terms of the retirement age can be interpreted as loss aversion in lifetime leisure, where workers perceive postponing retirement as a loss relative to a normal time to retire.

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

$$C(R) = \sum_{t=0}^{R-1} \delta^t w_t(1-\tilde{\tau}_t) + \sum_{t=R}^{T} \delta^t B(R) \tag{18}$$

where w is the gross wage per period, $\tilde{\tau}$ is the payroll tax/pension contribution rate, T is the time of death, and δ is the discount factor.[12] The slope of the budget constraint, that is the marginal gain in lifetime consumption possibilities C from delaying retirement by one period, defines the implicit net wage $w^{net} = dC/dR$. Expressing the consumption gain as a fraction of the gross wage, the *implicit net-of-tax rate* is $1-\tau = w^{net}/w$.

Bunching methods can be used to transparently identify key parameters of the model.[13] As Seibold

[11] Assuming quasi-linear utility in consumption and iso-elastic in lifetime labor supply is convenient for the bunching strategy described below, and, though not strictly necessary, it matches our theory above.

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

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

(2020) shows, the model predicts bunching at the Normal Retirement Age when it is perceived as a reference point by workers. One can identify a marginal bunching individual, whose indifference curve would be tangent to the budget line at some retirement age R^* without reference dependence, and who is tangent exactly at $\hat{R}$ with reference dependence. All workers initially located between $\hat{R}$ and R^* bunch at the reference point, while all individuals initially to the left of the reference point leave their retirement age unchanged and all individuals initially to the right of R^* stay above the reference point. Hence, the bunching mass B at a retirement age reference point is given by

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

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

$$\frac{b}{\hat{R}} = \left(\frac{1-\tau}{1-\tau-\Delta\tau-\Lambda}\right)^{\varepsilon} - 1, \tag{19}$$

where $\Lambda = \widetilde{\Lambda}/w$ is the reference dependence parameter normalized by the wage per period and $\Delta\tau$ is the size of the budget constraint kink that may be present at the threshold.[14]

Figure 5 shows the empirical retirement age distribution around the Normal Retirement Age. There is sharp, large bunching at age 65, the location of the NRA among these workers. The presence of bunching is in line with the NRA serving as a reference point for retirement. While sizable bunching at the NRA has been documented across a number of countries, it is particularly striking in the German case because there is a non-convex kink of size -0.28 at the NRA, providing a negative incentive to retire exactly at this age. The figure also shows a counterfactual density fitted as a polynomial to the empirical distribution, excluding the bunching region. Expressing the bunching mass relative to the counterfactual, the excess mass at the NRA is estimated around 31, implying that workers are roughly thirty times more likely to retire exactly in the month of the NRA than we would expect from the smooth counterfactual distribution.

We use the identification strategy of Seibold (2020) in order to estimate Λ and ε. In particular, we leverage the fact that bunching is observed at the Normal Retirement Age, but also at some standard, "pure" financial incentive discontinuities, i.e. budget constraint kinks or notches without the presence of a statutory age. Indexing these various thresholds by i, bunching can be written as

$$\frac{b_i}{\hat{R}_i} = \left(\frac{1-\tau_i}{1-\tau_i-\Delta\tau_i-\Lambda\cdot D_i}\right)^{\varepsilon} - 1 + \xi_i \tag{20}$$

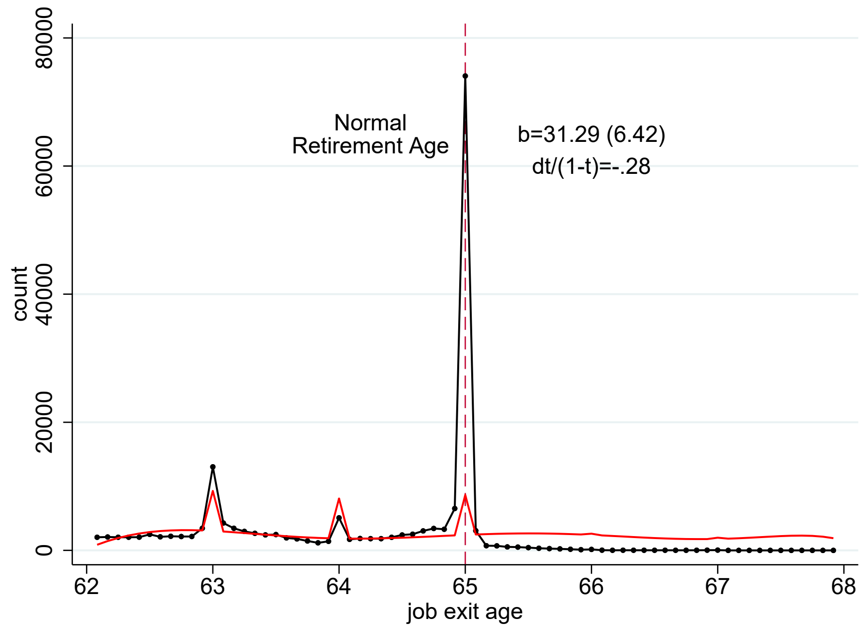
where D_i is an indicator for the Normal Retirement Age and ξ_i is an error term.[15]

Appendix Table A1 shows bunching estimates and resulting parameter estimates for birth cohort 1946. The bunching estimates are a subset of the estimates from Seibold (2020), where bunching is estimated for groups of workers defined by birth cohorts and retirement pathways, the level at which statutory ages and financial incentives vary. In Panel A of the table, the average excess mass at the NRA is 31.3, although there is a negative local financial incentive to retire corresponding to a kink size of –0.28. At other, pure financial incentive discontinuities faced by the same workers, the average excess mass of 6.73 is smaller, although these entail sizable financial incentives to retire with an average kink size of 0.47. These bunching

[14]We assume this transformed parameter Λ is homogeneous for simplicity. Heterogeneity in Λ is difficult to identify empirically given the design used here, but we acknowledge the limitation.

[15]The empirical specification also controls for whether the Normal Retirement Age coincides with the Full Retirement Age.

FIGURE 5: BUNCHING AT THE NORMAL RETIREMENT AGE



Note: The figure shows the pooled distribution of retirement (job exit) ages around the Normal Retirement Age among workers born in 1946. The black connected dots show the actual distribution, while the red line shows the counterfactual density estimated as a seventh-order polynomial excluding the bunching region. The counterfactual density also allows for round-number bunching and features an upward correction to the right of the NRA, where a shift of the retirement age distribution is predicted. The parameter b denotes the average excess mass at the NRA, corresponding to the estimate from Appendix Table A1.

observations can be used to estimate equation (20), yielding the estimates of $\Lambda = 0.46$ and $\varepsilon = 0.06$ shown in Panel B of the table. Thus, the parameter estimates for birth cohort 1946 are similar to the estimates reported in Seibold (2020) for a broader range of cohorts.

3.3 Policy Simulations

3.3.1 Pension Reforms, Reference Points and Prices

In the light of demographic change and resulting fiscal challenges for pension systems, two types of pension reforms are often considered in order to induce workers to postpone retirement. A first common policy is an increase in the Normal Retirement Age (or similar statutory retirement ages). For example, the NRA will be increased to age 67 in the U.S. by 2027, to 67 in Germany by 2031, and to 68 in the U.K. by 2046. This type of reform entails large effects on retirement behavior (Mastrobuoni, 2009; Staubli and Zweimüller, 2013; Manoli and Weber, 2016; Cribb et al., 2016), which is largely driven by shifting individuals' reference points to a higher retirement age (Behaghel and Blau, 2012; Seibold, 2020).

Two important aspects are worth noting about NRA reforms. First, while an increase in the NRA sets the reference point at a higher retirement age, such a reform corresponds to decreasing the reference point in terms of lifetime leisure in the model from Section 3.2. Thus, we should conceptually think of a reform that increases the NRA as one that *lowers individuals' reference points* in the sense of Proposition 1. Second, while our theory considered changes to reference points holding all else fixed, changes to the NRA typically entail some change in individuals' lifetime budget constraints, because pension benefit schedules are linked to the NRA. In the German context, the Delayed Retirement Credit is only available from the NRA onward. If this feature is maintained, increasing the NRA would also move the non-convex kink in the budget constraint to the new NRA. Moreover, if the NRA coincides with the FRA, the age from which the "full" pension is available may move upwards with the NRA, such that increasing the NRA effectively implies a benefit cut across the board.

The second type of policy often considered for pension reforms are changes to financial incentives. In particular, a natural way to incentivize workers to retire later is to offer higher marginal pension benefit increases for later retirement. This is typically done via increases in the Delayed Retirement Credit, providing higher actuarial adjustment to workers retiring beyond the NRA. For instance, the U.S. Delayed Retirement Credit has been gradually increased from 3% to 8% per year over the last decades. Conceptually, a higher Delayed Retirement Credit corresponds to a higher marginal return to work, or a *higher price of lifetime leisure in the loss domain* above the NRA. Whether intentionally or not, the Delayed Retirement Credit can thus be interpreted as a corrective "tax" in the sense of Corollary 5.1, which can incentivize individuals to move away from the reference point of the NRA by increasing their retirement age.

3.3.2 Simulations

We simulate the effects of two pension reforms of the two types discussed above. The first reform is an increase in the NRA from 65 to 66. The reform shifts individuals' retirement reference points and entails a (relatively small) change in the budget constraint. In order to maintain the feature of a budget constraint kink at the NRA, the Delayed Retirement Credit only applies above the new NRA in the simulation. However, the counterfactual scenario does not feature a benefit cut across the board below the NRA in order to avoid confounding the effects of influencing reference points with large mechanical fiscal and consumption effects. The second reform is an increase in the Delayed Retirement Credit. In order to anchor the second reform, we increase the credit from the current level of 6% to 10.2% per year, which yields the same effect on the average retirement age as the first reform.

The policy simulations proceed in the following steps. First, we require a counterfactual distribution of retirement ages – a distribution of retirement ages in the absence of reference dependence. We obtain this counterfactual distribution by fitting a polynomial to the observed distribution, excluding the bunching region around the Normal Retirement Age. The estimation yields an estimate of the bunching mass at the NRA. In the absence of reference dependence, this mass would be distributed across retirement ages above the NRA, and we simulate this un-bunching by distributing the bunching mass uniformly across the age range 65 to 68. 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 Delayed Retirement Credit is 6% per year. Third, we simulate optimal retirement ages under the two counterfactual policy scenarios. For this, we simulate individual lifetime budget constraints from equation (18) as in Seibold (2020), based on observed individual earnings and contribution histories, and choose the retirement age that maximizes utility from equation (17) given the level of $C(R)$ pinned down by the budget constraint and the location of the NRA $\hat{R}$.

Fourth, we compute the difference between each counterfactual scenario and the baseline scenario for each of the following outcomes: contributions to the pension system, benefits paid to workers, workers' lifetime consumption, all in terms of net present value at age 65. Moreover, we calculate the effects on disutility from working and reference dependence loss disutility given the preferences in equation (17). 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.

3.3.3 Main Results

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

Increasing the NRA. Column (1) shows the effects of the NRA increase. Shifting the NRA by one year increases average actual retirement ages by 7.4 months. As shown in Seibold (2020), such a reform improves the fiscal balance of the pension system. The positive fiscal effect arises due to a combination of a higher net present value of contributions collected and a lower value of benefit payments, both of which arise when individuals work longer and postpone retirement. The magnitude of the net fiscal effect is around +€11k per worker. Next, the reform affects individual welfare. Lifetime consumption increases by around +€7k along with later retirement. Disutility from work becomes larger due to an additional year of work, which enters negatively into individual welfare. However the increase in consumption outweighs the disutility from work, which mainly reflects the key result from Propositions 2 and 3 when $\pi = 0$. In words, the individual is consuming too much leisure when $\pi = 0$, so decreasing the reference point over leisure by increasing the NRA has a corrective effect that improves individual welfare. Thus we find that worker welfare improves when $\pi = 0$ in the table.

In addition, there is a decrease in reference dependence loss disutility due to the lower reference point in terms of lifetime leisure, which enters positively into welfare only if the planner places normative weight on reference dependence ($\pi = 1$). Hence, worker welfare also increases under $\pi = 1$. We can conceive of the overall change in reference dependence loss disutility of +€2k as the sum of two components: a negative component from the increase in the chosen R and a positive component from the increase in the reference point, $\hat{R}$. The first of these approximately offsets the effect on worker welfare when $\pi = 0$, which represents the theoretical idea that envelope conditions eliminate the behavioral welfare effect of a change in the reference point when $\pi = 1$. The second component is the direct welfare effect, which is the main first-order determinant of welfare when $\pi = 1$.

Finally, the effect on total welfare is given by the sum of the individual welfare effect and the net fiscal effect. Under $\pi = 0$, total welfare increases by around +€13k per worker. This reflects the fact that workers retire inefficiently early when the NRA is 65, and they move towards their optimal retirement ages when the NRA is increased. When $\pi = 1$, the total welfare gain of around +€15k is even larger due to the direct effect of a change in the reference point.

Increasing the Delayed Retirement Credit. Column (2) of the table shows the effects of the increase in the Delayed Retirement Credit to 10.2%. By construction, this policy achieves a sizable increase in the average retirement age like the NRA increase. However, a first important difference to the NRA reform is the fiscal effect. The net fiscal effect is negative at –€2.6k per worker. Workers also contribute for longer in this scenario, but the positive effect on contributions is more than offset by the large increase in benefit payments. Due to the higher pension benefits and the additional income from working another year, worker consumption increases strongly, as workers retiring later receive large pension benefit increases. Disutility from work becomes larger, but less so than the change in the NRA because workers account for their own marginal disutility of work in deciding just how much later to retire under the higher credit. Thus, there is

Table 1: Welfare Effects of Pension Reforms

	(1) Policy 1: Normal Retirement Age to 66	(2) Policy 2: Delayed Retirement Credit to 10.2%
Contributions collected	+4,042	+3,706
Benefits paid	+6,870	–6,298
Net fiscal effect	+10,912	–2,592
Worker consumption	+7,178	+19,389
Disutility from work	–4,969	–3,251
Worker welfare ($\pi = 0$)	+2,209	+16,138
Ref. dep. loss disutility	+1,945	–14,108
Worker welfare ($\pi = 1$)	+4,154	+2,030
Total welfare ($\pi = 0$)	+13,121	+13,545
Total welfare ($\pi = 1$)	+15,066	–562

Note: The table shows results from simulations of two pension reforms, an increase in the NRA from 65 to 66 and an increase in the Delayed Retirement Credit to 10.2%. Both reforms yield the same effect on the average actual retirement age (+7.4 months). Simulations are conducted for birth cohort 1946. All effects in Euros per worker, in terms of net present value at age 65. The signs the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare.

a large positive effect of +€16k on worker welfare under $\pi = 0$. However, the large increase in reference dependence loss disutility reduces individual welfare by –€14k if it carries normative weight under $\pi = 1$. This sizable negative effect arises because workers increase their retirement ages relative to an unchanged reference point, moving further into the loss domain. Taking the additional reference dependence effect into account, individual welfare increases only by +€2k under $\pi = 1$. Finally, the total welfare effect is positive at +€13.5k under $\pi = 0$, as the large gain in individual welfare strongly dominates the negative fiscal effects. However, the total welfare effect turns negative under $\pi = 1$, when workers experience large disutility from being "pulled away" from the reference point.

The sizable difference in worker welfare between the $\pi = 0$ and the $\pi = 1$ case is directly related to the theoretical results in Propositions 4 and 5. When $\pi = 0$, there is a sizable internality present, from workers consuming too much leisure out of loss aversion. Increasing the Delayed Retirement Credit can have a large positive welfare effect by (partially) correcting this internality. In contrast, when $\pi = 1$, the change in worker welfare is much smaller because this internality is not present. Moreover, in this case, the basic intuition of the envelope theorem implies that first-order effects on worker welfare will be virtually entirely offset by the net fiscal effect.[16]

3.3.4 Extended Simulations

We next provide some extended simulations depicting continuous changes to reforms. Some caution is warranted in interpreting some of these results as we are extrapolating further from observed data. Nev-

[16]Technically, this envelope theorem logic would apply when we start from an actuarially fair schedule and perturb the budget constraint. The initial 6% credit seems to be close enough to actuarially fair such that the envelope logic is approximately correct. We explore this further in subsequent extended simulations below.

ertheless, extending our simulations to a wider range of policies illustrates provide additional insights into the relationship between the policy simulations and our theoretical results.

While Table 1 considers a specific change to the Normal Retirement Age, Figure 6 shows results for a range of simulated counterfactual NRAs. The figure is based on simulations for NRAs between 65 and 67 in monthly increments. Overall, the figure confirms the robust positive welfare effects of increasing the NRA. To begin with, Panel (a) shows that the fiscal balance of the pension system increases with the NRA, as more contributions are collected and pension benefits are paid for shorter periods. In Panel (b), individual consumption increases with the NRA, but workers also experience higher disutility from working longer. Adding up those two components of standard preferences, individual worker welfare under $\pi = 0$ increases with the NRA up to a point, but then flattens out and starts decreasing slightly. This occurs because, contrary to a policy that only changes the reference point, some workers suffer pension benefit decreases as a result of changes to the NRA, since the Delayed Retirement Credit is only paid from the new NRA onwards in our simulation.[17] In Panel (c), we add reference dependence loss disutility in order to obtain worker welfare under $\pi = 1$. A higher NRA corresponds to a lower reference point in terms of leisure, and thus workers experience lower loss disutility, such that worker welfare is more positively affected under $\pi = 1$. Finally, Panel (d) shows that total welfare increases monotonically with the NRA under $\pi = 0$ and Panel (e) shows that, as with the basic reform in Table 1, the total welfare increase is even stronger under $\pi = 1$.[18]

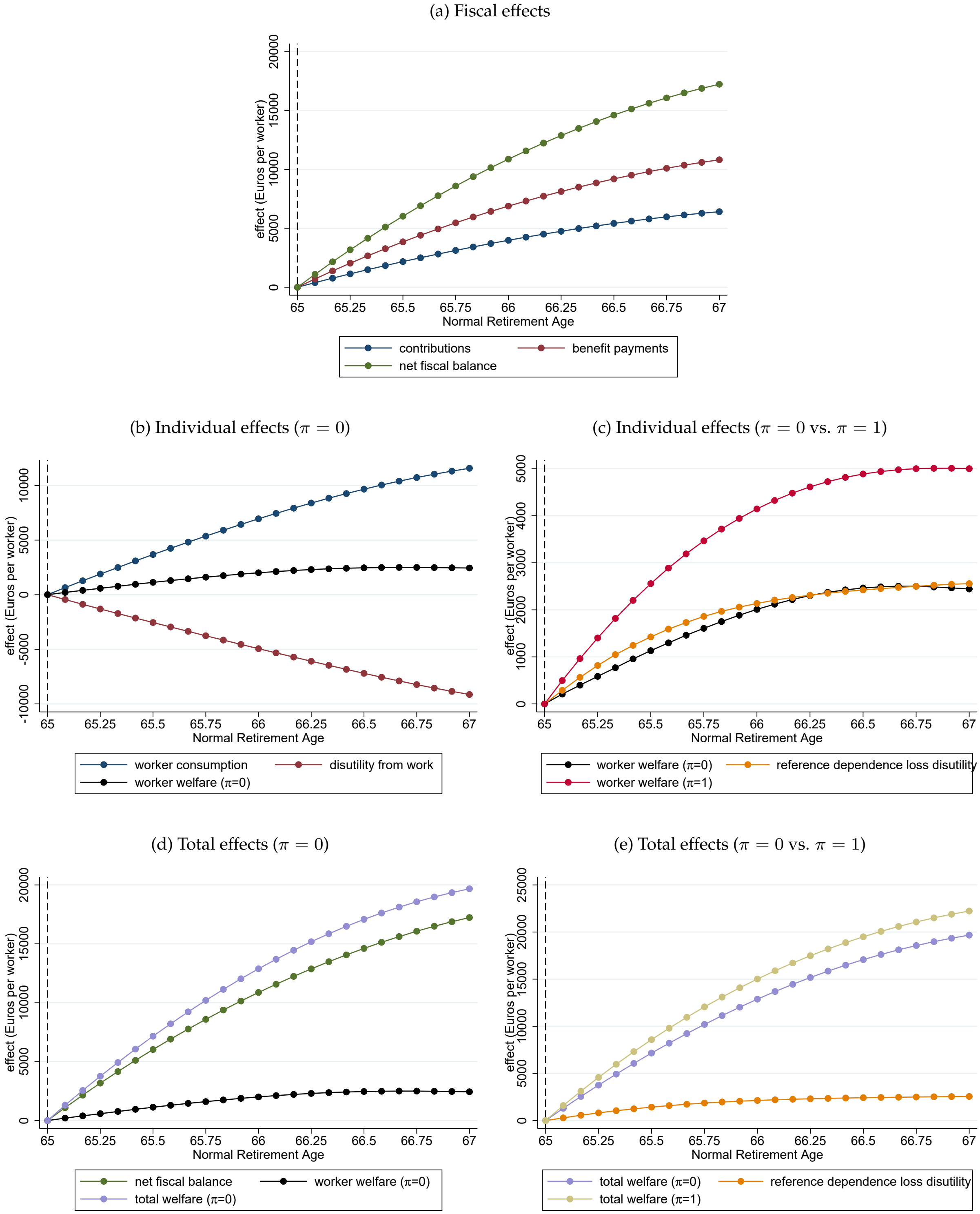
Similarly, Figure 7 shows results for a range of simulated values of the Delayed Retirement Credit. We simulate credits between 3% and 36% per year in half-percentage point increments. In Panel (a), the fiscal effects of increasing the Delayed Retirement Credit tend to be large and negative, since the large increases in pension benefit payments dominate increases in contributions received by the pension system. There is, however, a small range just above the current value of 6% over which the net fiscal effect of slightly increasing the credit is positive, as the pension system moves closer to actuarial fairness. Panels (b) and (c) show that workers always benefit from a higher Delayed Retirement Credit, since their consumption increases by more than disutility from working longer. Under $\pi = 1$ the positive effect on worker welfare is substantially dampened, however, as they experience strongly increased reference dependence loss disutility when moving further into the loss domain at older retirement ages.

A key difference between increasing the NRA and changing the Delayed Retirement Credit is that the total welfare effects of the latter reforms are not monotonic. While increasing the NRA always increases total welfare, both under $\pi = 0$ and $\pi = 1$, Panels (d) and (e) of Figure 7 show that it is possible to find an optimal level of the Delayed Retirement Credit. Importantly, the welfare-maximizing credit depends strongly whether the planner places normative weight on reference dependence. In Panel (d), total welfare is maximized at a very large Delayed Retirement Credit of 19.2% p.a., more than three times its current level. This results speaks to a possible role for the Delayed Retirement Credit to correct inefficiently early retirement under $\pi = 0$, as in Corollary 5.1. Such a large marginal financial return to working longer, or implicit price of leisure, can induce workers to retire later and move towards their optimal retirement age. In Panel (e), the optimal level of the Delayed Retirement Credit is much lower under $\pi = 1$. Intuitively,

[17] Whether individual welfare increases or decreases as a result of increasing the NRA depends on the exact change to pension adjustment. An NRA increase that entails no change at all in the budget constraint would increase individual welfare (as in Proposition 1). In the intermediate scenario we simulate here, there are modest effective benefit cuts for some workers because the Delayed Retirement Credit is only available from a later age. NRA reforms that entail benefit cuts across the board yield much larger positive fiscal effects at the expense of a stronger decrease in individual consumption.

[18] The relative quantitative similarity between total welfare under $\pi = 0$ and $\pi = 1$ is not a generic feature of the theory. Rather, it mainly occurs in this context because the number of individuals retiring exactly at the NRA, the R group in the theory, is relatively large throughout the range of NRAs we consider. This feature of the environment implies that the behavioral welfare effect of a change in the reference point, which mainly matters when $\pi = 0$, and the direct effect, which mainly matters when $\pi = 1$, will be quantitatively similar. With a larger L group, the direct effect, and thus the change in welfare for $\pi = 1$, could be significantly larger.

FIGURE 6: INCREASING THE NORMAL RETIREMENT AGE



Note: The figure shows results from simulations of pension reforms that increase the NRA to ages between 65 and 67 in monthly increments. Simulations are conducted for birth cohort 1946. All effects in Euros per worker, in terms of net present value at age 65. The signs the effects correspond to influence on welfare. Total welfare is the sum of net fiscal effect and change in worker welfare.

there is no reason for the planner to incentivize workers to move away from the NRA and retire later when reference dependence is not judged as a bias. The only rationale to increase the Delayed Retirement Credit slightly above its current level is to correct the inefficiency that arises from the fiscal externality, due to less than actuarially fair pension adjustment. Indeed, the optimal Delayed Retirement Credit we find in Panel (e) of 7.7% p.a. appears to be close previous calculations of actuarially fair adjustment in the German context (Börsch-Supan and Wilke, 2004).

Overall, this simulation illustrates essentially all of the main ideas from our theoretical results. Increasing the NRA, which corresponds to lowering reference points in terms of lifetime leisure, yields robust increases in total welfare. Increasing the Delayed Retirement Credit, which corresponds to an implicit tax on leisure in the loss domain, increases total welfare if reference dependence is judged as a bias. However, increasing the credit beyond its actuarially fair level decreases welfare if reference dependence carries normative weight. Two further aspects of these types of reforms are worth considering. First, if policymakers are mainly concerned about the fiscal sustainability of pension systems, increasing the NRA may be attractive in its own right as this yields sizable positive fiscal effects. Higher late retirement subsidies tend to yield negative fiscal effects, on the other hand. Second, if policymakers are uncertain about the appropriate welfare judgment of reference dependence, NRA increases may appear even more attractive, as their positive welfare effect is robust to the choice of π. The sign and magnitude of the welfare effects of the Delayed Retirement Credit, on the other hand, fully depend on this normative judgment.

4 Extensions of the Theory

4.1 Alternate Formulation of Reference Dependence

In this section, we consider the alternate formulation of reference dependence for $v(x|r)$, from equation (3), which takes the formulation of Tversky and Kahneman (1991) for reference dependence over riskless choice at face value. As discussed above, the main difference between this formulation and the one we use above is the presence of the η_i parameter in the formulation we consider here.

Setup. The η_i parameter makes the individual consume more x by virtue of comparing x to the reference point, in both the gain and the loss domain. We think of η_i as governing the importance of reference dependence itself, while λ_i governs the strength of loss aversion. We first note that we can re-formulate the reference dependence in equation (3) as follows, to make it slightly more comparable to our earlier model:

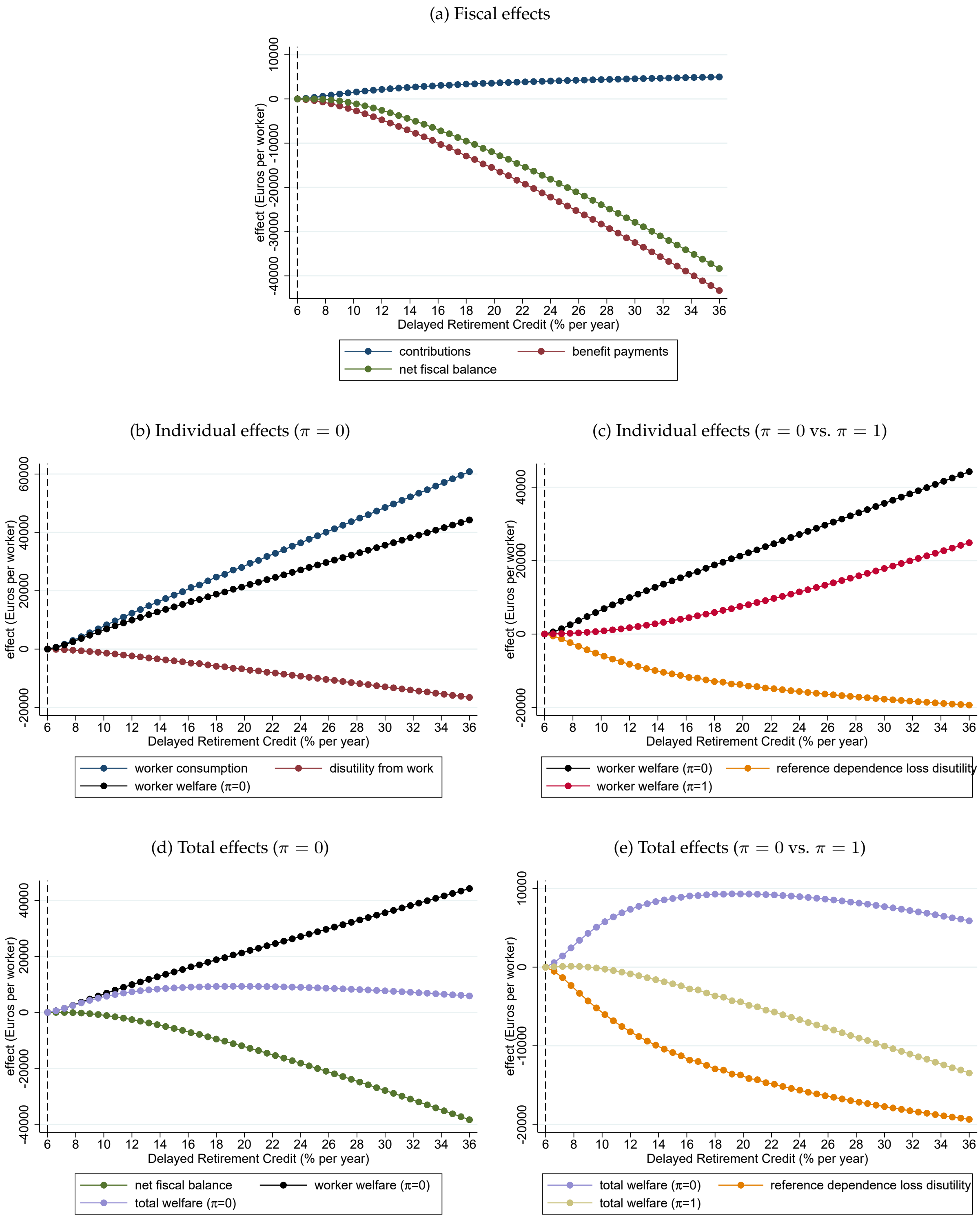
$$\tilde{U}_i(x,y) = \tilde{u}_i(x) + y + \tilde{v}_i(x|r), \tag{21}$$

$$\tilde{v}_i(x|r) = \begin{cases} \eta_i(x-r), & x > r \\ [\eta_i + \Lambda_i](x-r), & x < r, \end{cases} \tag{22}$$

The formulation used here (and in equation 3) is behaviorally equivalent to the model from section 2, with $\tilde{u}_i(x) = u_i(x) - \eta_i x$ and $\Lambda_i = \eta_i(\lambda_i - 1)$ (see Appendix C for a full proof). We can therefore compare the question of how, holding observed behavior fixed, adopting this formulation for welfare instead of the one used in Section 2 affects our normative results.

We consider the question of whether each friction, reference dependence or loss aversion, separately,

FIGURE 7: INCREASING THE DELAYED RETIREMENT CREDIT



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

reflects a behavioral bias or a normative preference, using the parameters $\pi^{RD} \in \{0,1\}$ and $\pi^{LA} \in \{0,1\}$. We therefore use the following specification for welfare:

$$\tilde{U}_i^*(x,y) \quad = \quad \tilde{u}_i(x) + y + \tilde{v}_i^*(x|r), \tag{23}$$

$$\tilde{v}_i^*(x|r) \quad = \quad \begin{cases} \pi^{RD}\eta_i(x-r), & x > r \\ [\pi^{RD}\eta_i + \pi^{LA}\Lambda_i](x-r), & x < r. \end{cases} \tag{24}$$

We do not consider the case where $\pi^{RD} = 0$ and $\pi^{LA} = 1$, because this judgment would imply that reference dependence over gains and losses is a bias, but loss aversion is normative, which does not seem sensible.

Finally, we denote indirect utility by $\tilde{w}_i(p,r)$, and utilitarian social welfare by $\tilde{W}(p,r)$, defined as above.

Results. We next show how adopting this formulation modifies the marginal internality and the first-order social welfare effects of a change in price and the reference point.

Lemma 2. ***Marginal Internalities Under the Tversky-Kahneman (1991) Form.*** *Let $\tilde{m}_i$ be the derivative of $\tilde{U}_i^*(x, z_i - px)$ with respect to x, evaluated at $x_i(p,r)$.*

L2.1. *If $x_i(p,r) > r$, $\tilde{m}_i = -(1-\pi^{RD})\eta_i \equiv \tilde{m}_i^G$.*

L2.2. *If $x_i(p,r) < r$, $\tilde{m}_i = -(1-\pi^{RD})\eta_i - (1-\pi^{LA})\Lambda_i$.*

L2.3. *If $x_i(p,r) = r$,*

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

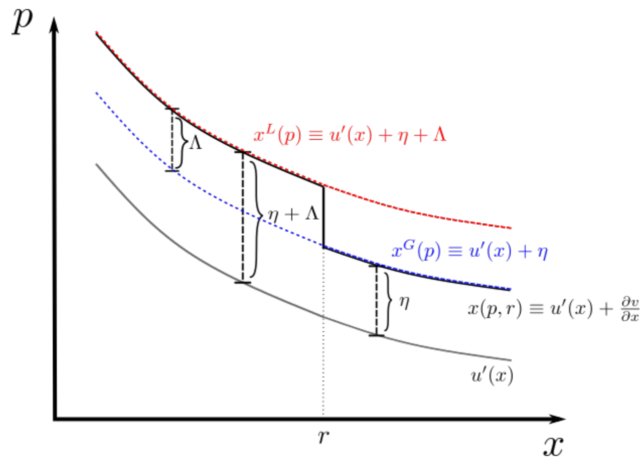
Comparing Lemma 2 to the analogous Lemma 1 helps us understand how adding the friction embodied by η_i changes the model. When $\pi^{RD} = 1$, these marginal internalities are all exactly the same as in our earlier analysis except that π is now denoted π^{LA} – recall that the models are behaviorally isomorphic if $\tilde{u}_i(x) + \eta_i = u_i(x)$. When $\pi^{RD} = 0$, however, reference dependence is generating additional distortions to the choice of x, leading to more over-consumption than in our earlier formulation. The revised marginal internalities and welfare-maximizing demand are plotted in Figure 8. Welfare maximizing demand corresponds to the line depicting $p = u_i'(x)$, when $\pi^{RD} = 0$. When $\pi^{RD} = 1$, welfare-maximizing demand corresponds to either observed demand or $x_i^G(p)$, depending on whether or not $\pi^{LA} = 1$.

Proposition 6. ***Modified First-Order Welfare Effects*** *Consider a change in the reference point that is small enough that the GL group is negligible. Let $\Delta r = r_1 - r_0$, and let $\Delta x_i = x_i(p,r_1) - x_i(p,r_0)$. Holding observed behavior constant, denote the change in welfare that we obtain adopting the original formulation by Δw_i, as in Proposition 2, with $\pi = \pi^{LA}$.*

P6.1. *Let $\Delta\tilde{w}_i = \tilde{w}_i(p,r_1) - \tilde{w}_i(p,r_0)$ be the change in welfare from our modified formulation. We obtain*

$$\Delta\tilde{w}_i \approx \Delta w_i - \pi^{RD}\eta_i\Delta x - (1-\pi^{RD})\Delta r. \tag{25}$$

FIGURE 8: OBSERVED DEMAND, WELFARE-MAXIMIZING DEMAND, AND MARGINAL INTERNALITIES UNDER THE TVERSKY-KAHNEMAN (1991) FORMULATION



Note: This figure depicts observed demand $x(p, r)$, in black, at a given reference point as prices vary. We also plot $u'(x)$, in grey, which coincides with welfare-maximizing demand when $\pi^{RD} = \pi^{LA} = 1$. The marginal internality when $\pi = 0$ in this model is the vertical distance between observed demand and $u'(x)$, which is depicted in both the gain and loss domains. In the loss domain, observed demand coincides with $x^L(p)$ (red). In the gain domain, observed demand coincides with $x^G(p)$ (blue), which is also welfare-maximizing demand when $\pi^{RD} = 1$ and $\pi^{LA} = 0$. In this case there is no marginal internality in the gain domain, and the vertical distance between observed demand and $x^G(p)$ is the marginal internality in the loss domain. When $\pi^{RD} = \pi^{LA} = 1$, observed demand is welfare-maximizing and there are no internalities.

P6.2. *Considering the change in social welfare using our original and modified formulations analogously, we obtain*

$$\Delta\tilde{W} \approx \Delta W - E[\pi^{RD}\eta_i\Delta r] - E[(1-\pi^{RD})\eta_i\Delta r \mid i \in R]P[i \in R]. \tag{26}$$

P6.3. *Now consider a change in price. Let* $\Delta x_i \approx \frac{\partial x_i^L(p)}{\partial p}\Delta p$ *for* $i \in LL, RL$, $\Delta x_i \approx \frac{\partial x_i^G(p)}{\partial p}\Delta p$ *for* $i \in GG, GR$, *and* $\Delta x = 0$ *for* $i \in RR$. *Using analogous notation, differencing welfare across varying prices, we have:*

$$\Delta\tilde{w}_i \approx \Delta w_i - (1-\pi^{RD})\eta_i\Delta x_i. \tag{27}$$

P6.4. *Using similar notation again, the first-order social welfare effect of a change in price is*

$$\Delta\tilde{W} \approx \Delta W - E[(1-\pi^{RD})\eta_i\Delta x_i], \tag{28}$$

Proposition 6 shows that introducing the potential additional distortion to demand represented by the η_i parameter in this formulation of reference dependence has two potential effects on our welfare calculations. We illustrate welfare effect for change in reference point for the five cases, as in Figure 2, in Appendix Figure A1. Likewise, Appendix Figure A2 shows the modifications to Figure 3.

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

That the two formulations we have considered so far are behaviorally indistinguishable but carry differing implications for welfare raises interesting questions for future empirical research. In general, one might obtain specialized choice data, beyond observed demand at one or more reference points, to distinguish between the models. For example, the two models carry different predictions for what would happen if we can observe a situation where the reference dependence were eliminated by some intervention. Alternatively, imagine that we solicit individuals' willingness to pay to change the reference point.[19] The different formulations of reference dependence we consider make detailed predictions about how individuals would respond to either of these interventions. Additionally, with either of these novel designs, one could potentially identify the parameter η.

4.2 Multi-Dimensional Reference Dependence

Behavior. Thus far, we have considered reference dependence along a single dimension of the menu space. In some contexts, individuals apparently exhibit reference dependence in more than one dimension.

In this section, we consider welfare in a two-dimensional model of reference dependence.[20] For instance, in our empirical application, we could interpret the two dimensions as representing reference dependence over consumption and leisure as in Crawford and Meng (2011). We lay out a framework for normative analysis in this model and discuss how empirical analysis can discipline its use.

4.2.1 A Two Dimensional Model

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

$$\begin{gathered}\max_{x,y} u_i(x) + v_i(x|r) + y + w_i(y|s) \\ \text{subject to } px + y = z_i\end{gathered} \tag{29}$$

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

$$w_i(y|s) = \begin{cases} 0, & y > s \\ \Gamma_i(y-s) & y \leq s. \end{cases} \tag{30}$$

[19]The type of meta-willingness-to-pay design we have in mind resembles that of Allcott and Kessler (2019), who studied it for a different intervention.

[20]Extending this analysis to arbitrary dimensionality of the goods space is straightforward, though as we discuss below, it can be difficult to identify a multi-dimensional model empirically.

For simplicity, we do not include potential distortions in the gain domain, like η_i from the previous section, when we specify reference dependence over x and y.[21] Also for simplicity, we restrict our attention to the situation where the two-dimensional reference point (r, s) is on the budget constraint: $pr + s = z_i$.[22] With this restriction $x > r \iff y < s$. As in equations (4) and (5), the first order conditions for $x > r$ and $x < r$ are now given by

$$\frac{u_i'(x_i^G(p))}{1+\Gamma_i} = p, \tag{31}$$

$$u_i'(x_i^L(p)) + \Lambda_i = p \tag{32}$$

Behavior is given by equation (6) with the new potential demand curves $x^L(p)$ and $x^G(p)$ implied by equations (31) and (32).

Welfare. As in equation (7), we specify welfare as:

$$U_i^*(x, y) = u_i(x) + \pi v_i(x|r) + y + \pi w_i(y|s). \tag{33}$$

We assume that in the presence of loss aversion over both x and y, the planner makes the same judgment over whether loss aversion is normative, encoded by $\pi \in \{0, 1\}$.[23]

Lemma 3. ***The Marginal Internality in the 2-D Model.*** *Let m_i be the derivative of $U_i^*(x, z_i - px)$ from equation (33) with respect to x, evaluated at $x_i(p, r)$.*[24]

L3.1. *If $x_i(p, r) > r$, $m_i(p, r) = (1 - \pi)\Gamma_i p$.*

L3.2. *If $x_i(p, r) < r$, $m_i(p, r) = -(1 - \pi)\Lambda_i \equiv m_i^L$*

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

- *$m_i(p, r)$ is undefined when $\pi = 1$.*
- *$m_i(p, r) = u_i'(r) - p$ when $\pi = 0$, with $-\Lambda_i \leq m_i \leq \Gamma_i p$*

Note that when $\pi = 0$, the marginal internality is positive in the gain domain, unlike before, while it continues to be negative in the loss domain. The individual under-consumes x out of loss aversion over y when $x > r$ and over-consumes x out of loss aversion over x when $x < r$. Figure 9 illustrates demand in this model. We plot demand in the gain and loss domain, x^L and x^G according to equations (31) and (32). The main difference in the two-dimensional model is that demand with no reference dependence, pinned down by $p = u'(x)$, now falls *between* observed demand in the gain and loss domains. As $p = u'(x)$ describes welfare-maximizing demand when $\pi = 0$, the vertical distance between this demand curve and observed demand equals the marginal internality when $\pi = 0$, which we know from Lemma 3 is now positive in the gain domain and negative in the loss domain. When $\pi = 1$, observed demand is welfare-maximizing and there is no marginal internality.

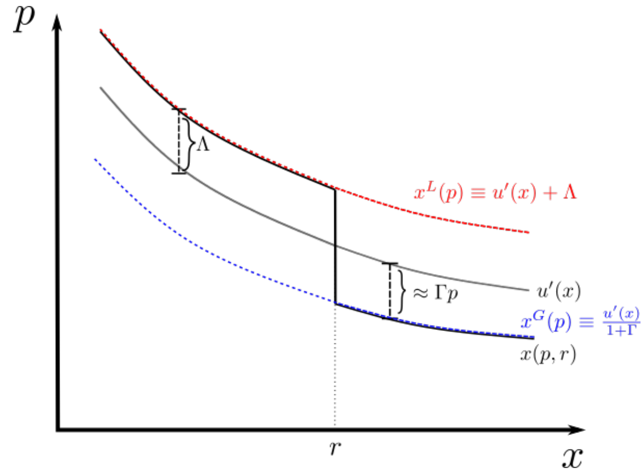
[21] Including an η_i-like term for both dimensions is a straightforward extension of this model, but such a model is very difficult to identify empirically without some methodological progress. Identifying η_i for a single dimension is difficult due to the issues discussed above; separately identifying such a parameter for two different dimensions would be even more challenging.

[22] Implicitly this implies that holding r fixed for everyone, s is heterogeneous across individuals with different z_i; our welfare effects below account for this.

[23] We use the same π in both dimensions. Relaxing this assumption is straightforward, but we have difficulty imagining why one might judge that loss aversion is normative in one dimension and not in another.

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

FIGURE 9: OBSERVED DEMAND, WELFARE-MAXIMIZING DEMAND, AND MARGINAL INTERNALITIES UNDER 2-DIMENSIONAL REFERENCE DEPENDENCE



Note: This figure depicts observed demand $x(p, r)$, in black, at a given reference point as prices vary. We also plot $u'(x)$, in grey, which coincides with welfare-maximizing demand when $\pi = 1$. The marginal internality when $\pi = 0$ in this model is the vertical distance between observed demand and $u'(x)$, which is depicted in both the gain and loss domains. In the loss domain, observed demand coincides with $x^L(p)$ (red). In the gain domain, observed demand coincides with $x^G(p)$ (blue). In this model, when $\pi = 0$, the marginal internality is therefore positive in the gain domain and negative in the loss domain.

4.2.2 Disciplining Dimensionality Empirically

We can infer from the similarities between Figures 1 and 9 that separately identifying Λ_i and Γ_i will be challenging using within-individual demand curves. However, it turns out that the character of the bunching around the reference point, which comes from between-individual comparisons of demand, can be strongly suggestive of the relative importance of reference dependence in a given dimension.

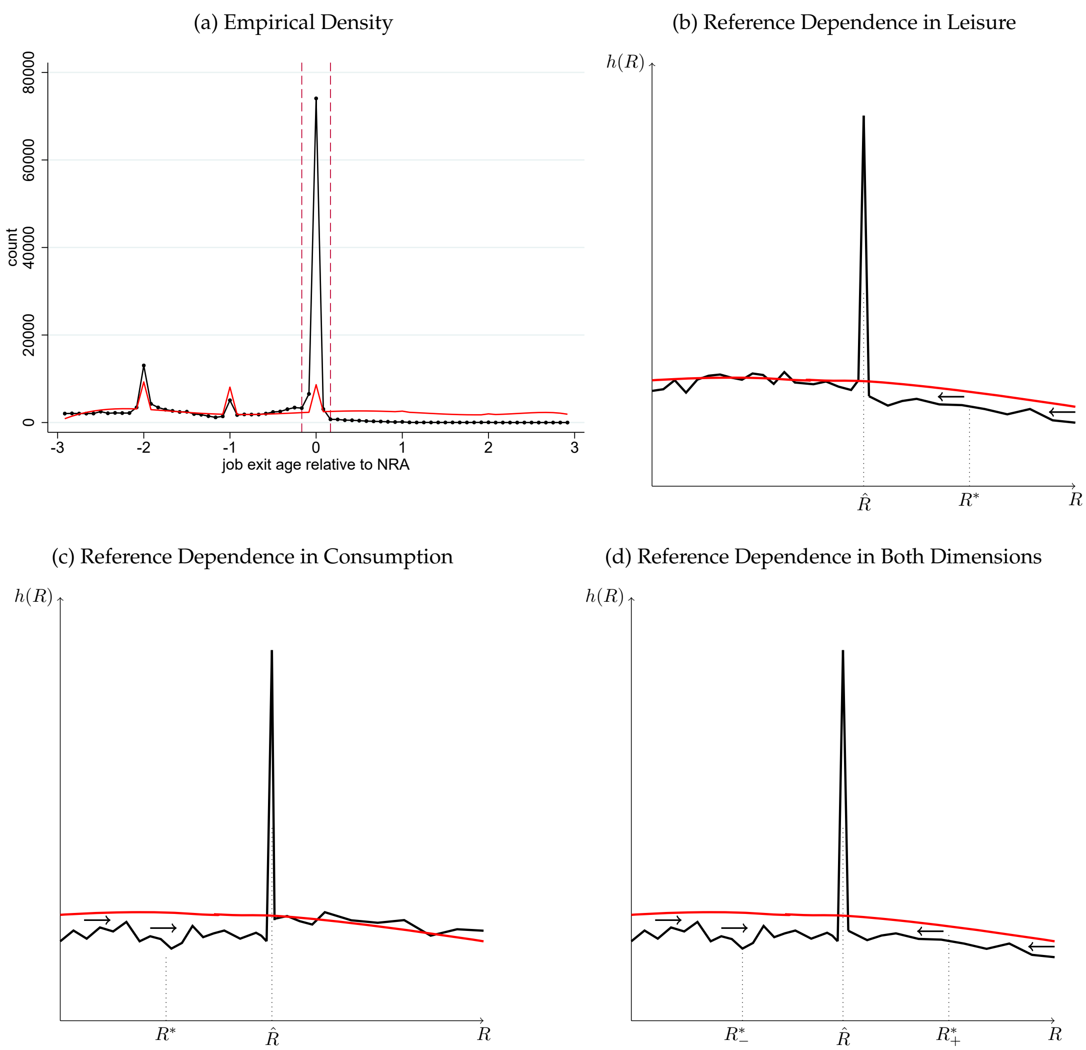
Specifically, holding p fixed, suppose we examine closely the distribution of individual choices around the reference point, $x_i(p, r)$. Examining this distribution is similar to the bunching analysis from our empirical application above. Figure 10 illustrates the patterns of bunching in retirement ages we would observe under-two dimensional reference in three cases: when $\Gamma_i = 0$, when $\Lambda_i = 0$, and when both Γ_i and Λ_i are nonzero.[25] Recall that when we plot the distribution of retirement ages for this purpose, the domain of ages greater than the NRA is the loss domain.

We note in Figure 10 that in the German pension context, bunching around the reference point of interest appears to be driven by individuals who would make a choice loss domain over leisure in the counterfactual with no bunching (and thus no influence of the reference point). This pattern suggests that reference dependence in the domain of the other, numeraire good - consumption in the application - is negligible, so we implicitly assumed $\Gamma_i = 0$ in our empirical application above.[26] Similar bunching originating from the loss domain is also found in the distribution of tax due at filing (Rees-Jones, 2018) and in the analysis of marathon-runner finishing times (Allen et al., 2017). In fact, we know of no situations where bunching

[25]In the third case, the relative dominance of one or the other dimension of reference dependence determines the degree to which bunching comes for the left or right and the exact levels of the distribution just to the right and left of the reference point.

[26]This test is ultimately only suggestive, as it could be the case that reference dependence is significantly stronger in the leisure dimension than in the consumption dimension. Given that we cannot identify a model with reference dependence in two dimensions with the available data, we maintain that considering reference dependence only over leisure is most appropriate given what we observe. Still, we fully acknowledge this limitation.

FIGURE 10: BUNCHING AND THE DIMENSIONS OF REFERENCE DEPENDENCE



Note: This figure illustrates the patterns of bunching around the reference point we should expect for different possibilities nested by our two-dimensional model of reference dependence. To emphasize the relationship to our empirical application in section 3, we plot the hypothetical distributions of retirement ages (analogous to $-x$ in the theory) and we call the two goods leisure (analogous to x in the theory) and consumption (analogous to y). The black line denotes the observed density of retirement ages and the red line shows the counterfactual distribution, representing the distribution of choices in the absence of reference dependence. We observe that whether bunching comes from the gain or loss domain depends on whether reference dependence is primarily over leisure or consumption, which is determined by the values of Λ and Γ in the model. In panel (a), we find that the empirical evidence is consistent with reference dependence over leisure - panel (b) - which supports our use of that form of reference dependence in Section 3.

attributed to reference dependence appears to come from both the gain and the loss domain. Nevertheless, in some applications, multi-dimensional reference dependence does seem reasonable. We ultimately defer to future research the difficult question of when and why reference dependence appears predominantly in one dimension or in multiple dimensions.

4.2.3 Main Social Welfare Effects with Two Dimensions

Proposition 7. ***First-Order Social Welfare Effects in the 2-D Model*** *Starting from any given price and an initial reference point, define groups G, L and R based on $x_i(p, r)$.*

P7.1. *The effect of a small change in the reference point of Δr on social welfare in this model is approximately*

$$\begin{aligned}\Delta W \approx & \Delta r \pi \left\{ E\left[\Gamma_i p \mid i \in G\right] P\left[i \in G\right] - E\left[\Lambda_i \mid i \in L\right] P\left[i \in L\right] \right\} \\ & - \Delta r E\left[p - u_i'(r) \mid i \in R\right] P\left[i \in R\right]. \end{aligned} \tag{34}$$

P7.2. *The effect of a small change in price, Δp, on social welfare in this model is approximately*[27]

$$\begin{aligned}\Delta W \approx & \Delta p \left\{ E\left[(1-\pi)\Gamma_i p \frac{\partial x_i^G}{\partial p} \;\middle|\; i \in G \right] P[i \in G] - E\left[(1-\pi)\Lambda_i \frac{\partial x_i^L}{\partial p} \;\middle|\; i \in L \right] P[i \in L] \right\} \\ & - E[x_i(p_0, r_0)]\Delta p, \end{aligned} \tag{35}$$

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

Proposition 7 summarizes our main social welfare effects in the two-dimensional model. The Proposition entirely nests Propositions 3 and 5 above when $\Gamma_i = 0$ for all i, i.e. when there is no loss aversion over y. When $\Gamma_i > 0$, we observe an additional direct effect of changing the reference point for individuals consuming in the loss domain over y.

Proposition 7.1 considers changes in the reference point. We also depict the welfare effects of changes in the reference point visually in Figure A3. When the reference point is a point on the budget constraint $s = z - pr$, decreasing r must increase s, which is the policy change we consider in equation (34). The direct effect is therefore positive for individuals with $x > r$. Additionally, we should now expect that there are some individuals in the R group for whom $p > u_i'(r)$ and some for whom $p < u_i'(r)$, so the third term in equation (34) is ambiguously signed. These factors imply that in the two dimensional model, decreasing the reference point along a single dimension – i.e. decreasing r or decreasing s in isolation – would be a robust Pareto improvement, but decreasing the reference point r *along the budget constraint* is not generally a robust social welfare improvement. We can infer from equation (34) that such a decrease in the reference point r would be welfare-improving only if 1) Λ_i is sufficiently large compared to $p\Gamma_i$ on average, 2) there are more individuals in the loss domain compared to the gain domain, and 3) individuals consuming at the reference point tend to be over-consuming rather than under-consuming relative to marginal utility u'.[28]

Proposition 7.2 and Figure A4 consider price changes in the two-dimensional model. For a price change we continue to observe a behavioral welfare effect valued by the marginal internality m_i, and a direct effect.

[27] In this approximation we do not allow the reference point over y, s, to change when the price changes. Doing so would introduce a direct welfare effect of a change in that reference point.

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

The main difference in the two-dimensional model is that the marginal internality is positive for $x > r$, which implies that decreasing consumption in response to a change in price decreases welfare for $i \in G$.

Our treatment of two-dimensional reference dependence here is intended mainly to sketch a path forward for future work on this topic. Even from our simple results, it is evident that attempts to study welfare under multi-dimensional reference dependence can be quite complicated. First, we can see from our results that the application a model like this one to concrete settings will require empirically separating the effect of reference dependence along different dimensions, or further restrictions on the relative strength of reference dependence in a given dimension. One intuitive and influential restriction was proposed by Köszegi and Rabin (2006), but whether and when this restriction is empirically justified is less clear to us and adopting it would impose substantial structure on welfare. Second, we have not included the friction from the introduction of the η_i parameter in the previous section to the two-dimensional model. Doing so is technically straightforward, but, as η_i is difficult to identify even in one dimension, it would be difficult to take such a model to data. Third, one might wish to consider more than two dimensions, for instance to study multi-attribute reference dependence and brand choice (Hardie et al., 1993). Finally, we assumed that the reference point must be on the budget constraint because we found this restriction simple and intuitive, but it may not be appropriate in all contexts. In short, a large number of extensions are theoretically feasible, but whether the resulting models will be empirically useful is far less clear.

4.3 Goals

The only potential behavioral friction we considered so far is reference dependence. Part of the literature on reference dependence considers the possibility that reference points may serve as goals, in order to overcome another bias, such as present bias (Koch and Nafziger, 2011, 2016). Such a model contains two new elements compared to what we have done so far: an additional behavioral bias, and the possibility that individuals set their own reference points with some degree of sophistication about those biases.[29]

To extend our theory along these lines, we first suppose that decision utility governing the choice of x is the same as in our original model, i.e. equations (1) and (2). Because reference dependence induces individuals to consume more of a given good in order to avoid losses, using reference points to overcome biases is useful when biases lead to under-consumption of some good ordinarily.[30] As such, the new component of our theory here is an additional positive internality from consuming good x, which we model in a reduced form fashion as follows:

$$U_i^*(x, y) = u_i(x) + y + b_i(x) + \pi v_i(x|r). \tag{36}$$

The new bias term $b_i(x)$ captures the extent to which the individual under-values x when making a decision. We assume $b_i'(x) > 0$, i.e. a positive marginal internality from $b_i(x)$, and $b_i''(x) \leq 0$, which ensures an interior solution for the welfare-maximizing choice of x. As a benchmark, we assume individuals are *fully sophisticated* about their biases. That is, the individual is perfectly aware of the bias $b(x)$ when they set a reference point. We discuss what happens if we relax this assumption below.

We next make a simplifying assumption, which is that the bias $b_i(x)$ is not so large that it cannot be overcome by setting a reference point. Formally, let $x_i^* = \arg\max u_i(x) + b_i(x) + z - px$. We assume that

[29] This notion of sophistication is closely related to "planner-doer" models Fudenberg and Levine (2006). Intuitively, the planner sets r with knowledge of what the doer will choose, and then the doer makes choices as in our basic model.

[30] If the individual could set reference points or goals for other goods, then a tendency to over-consume one good (e.g. leisure due to present bias and the up-front costs of work) could potentially be overcome by setting a goal/reference point for the residual good (e.g. a reference point over consumption, or an earnings target). We abstract away from this case here, but our results can be straightforwardly applied to this type of situation too.

$b_i'(x_i^*) \leq \Lambda_i$.[31]

In this setting, it is straightforward to show that a fully sophisticated individual would set an optimal reference point at $r_i = x_i^*$. Because $u'(x^*) + b'(x^*) = p$ and $b_i'(x_i^*) \leq \Lambda_i$, we know that the individual subsequently chooses x at the reference point, i.e. in the R case from above. Appendix Figure A5 illustrates the choice of r in this case. In other words, sophisticated individuals set a goal r to overcome their biases and they meet this goal exactly.

How does this result change our earlier thinking? Unsurprisingly, letting individuals choose r optimally implies that policies that aim to reduce r will no longer improve welfare. More specifically, inducing a marginal increases or decrease in r in this case has no first-order welfare effects due to the envelope theorem, and a second-order loss due to the optimality of r_i. Next, to understand price changes, note that the individual always ends up in the R case, where a marginal change in price has no effect on behavior, only a first-order negative direct effect on welfare. Another interesting implication of this line of reasoning is that if individuals self-regulate their own biases by setting goals, there is no need to correct the bias in $b(x)$ by setting a corrective tax. Note also that all of the above obtains regardless of π: in this situation, the individual never incurs a loss, so whether loss aversion is normative becomes irrelevant.

From the logic we have now laid out, it is straightforward to infer what would happen if we relaxed the assumption of full sophistication about biases. An individual who underestimates their bias, or who neglects it entirely, would set a goal r that is too low. In this case, inducing the individual to set a higher reference point would have a first-order, positive impact on welfare. Similarly, an individual who overestimates their future bias would set an over-optimistic goal r, and could be made better off by setting a reduced reference point, especially when $\pi = 1$ and the losses incurred by failing to meet one's goal generate a negative payoff. Obviously, people sometimes do fail to meet their goals, incurring potentially painful losses, and sometimes they significantly exceed them. Whether policymakers would have enough information to correct these types of mistakes in the choice of goal reference points is less clear, but estimates of individuals perceptions of their own biases compared to estimates of their actual biases could inform this question.

This model is just one intuitive model in which individuals deliberately choose their own reference points. Thinking more broadly, the model we laid out in Section 2 implies that if given the ability to manipulate their reference points, individuals would choose low levels of r, low enough to avoid ever incurring losses. Insofar as individuals appear to deliberately set higher reference points, by revealed preference they must have some concern that is not present in the model in Section 2. Once we account for these additional concerns, provided that individuals choose their reference points fully optimally, the envelope theorem bites, implying that changing reference points will not have first-order welfare effects. Moreover, all of this will be true for welfare in other models that endow individuals with other motives to choose high reference points, such as models of anticipatory utility (Sarver, 2012; Thakral and Tô, 2020a).

[31] Without this assumption, the individual would choose a reference point in order to induce the highest consumption of x possible without incurring a loss. In other words they would set r such that $u_i'(r) + \Lambda_i = p$, so that $r = x_i^L(p)$. Another related complication is the question of whether the individual, when setting r, cares about the reference dependent payoff in $v_i(x|r)$. Under the simplifying assumption we make on the size of the bias here, this turns out not to matter for the choice of r: the individual would choose the same r if they only sought to maximize $u_i(x) + b_i(x) + y$ rather than $u_i(x) + b_i(x) + y + v_i(x|r)$. If our simplifying assumption does not obtain, so $b_i'(x_i^*) > \Lambda$, the individual choosing r to maximize $u_i(x) + b_i(x) + y$ would be indifferent between any r such that they made a choice of x in the loss domain.

4.4 Further Theoretical Extensions

In this section, we briefly discuss three further complications that could be added to our model, which may be useful in some applied settings. We defer fully characterizing welfare in these models to future work. Our subjective view is that these complications are best explored in applied contexts where empirical evidence and features of the environment can discipline the structure one imposes on the model.

Diminishing Sensitivity. As discussed above, we have so far ignored diminishing sensitivity, which, compared to equation (3), would require that $v'' < 0$ for $x > r$ and $v'' > 0$ for $x < r$.

Adding diminishing sensitivity is straightforward, but there is limited evidence of diminishing sensitivity for decision-making under certainty. Moreover, with diminishing sensitivity, it becomes difficult to empirically distinguish curvature of intrinsic utility over x, which we denoted $u''(x)$, and curvature over reference dependent utility $v''(x)$ in the gain domain.

Allowing for diminishing sensitivity is unlikely to change much of the qualitative intuition above, about the direct and behavioral effects of a change in the price or the reference point. For example, the presence and sign of the marginal internality will be unaffected, and the result that lower reference points tend to improve welfare will obtain under diminishing sensitivity. However, the presence of v'' would imply that the various demand curves in Figures 1 through 3 are no longer parallel, which implies that the size of various welfare effects will be quantitatively different. Insofar as we only consider choices nearby the reference point, such differences are negligible as in this region the piece-wise linear formulation we use is approximately accurate.

Risk and Uncertainty. We have here considered the case of reference dependence under certainty. Reference dependence under uncertainty is the subject of a rich theoretical and experimental literature.

There are two challenges in adapting the type of welfare analysis we consider here to a model with uncertainty. First, the question of exactly how to specify welfare in such a model with uncertainty can be tricky. Many applications of welfare economics under uncertainty use certainty equivalence welfare metrics (Einav et al., 2010). With reference dependence, at least under $\pi = 1$, state dependence in the model makes certainty equivalence a poor welfare metric, so a generalization of equivalent variation that allows for uncertainty and state dependence would be more appropriate.

Second, as discussed in Section 2.2.1, there is much more of a debate on the origins of reference points for the stochastic case. The mixed empirical evidence on the origins of reference points and the wide variety of models one might use makes it more difficult to choose a formulation of reference-dependent preferences and it may not be clear which policy changes might induce a shift in reference points. Researchers often pose that reference points are based on expectations (e.g. Köszegi and Rabin (2006)). In this case, changing beliefs would change the reference point, but changing beliefs can also influence welfare and behavior in other ways, which complicates the analysis.

Narrow Bracketing. An important component of prospect theory as laid out by Kahneman and Tversky (1979) is the bracketing of payoffs, which is closely related to the concept of mental accounting (Thaler, 1985). What we have considered above is essentially "broad bracketing." The agents evaluate all purchases of good x, with a reference point over total x consumed. In the labor-leisure model from our application, there is a single reference point for lifetime leisure.

In other contexts, individuals seem to adopt *narrower bracketing,* where what we would ordinarily think of as the same option is partitioned into component parts and evaluated separately, with a reference point for each. For example, narrow bracketing of assets in a financial portfolio, through which individuals receive a jolt of reference-dependent utility each time they sell a specific stock, appears to be an important driver of the disposition to hold stocks that depreciate and sell those that appreciate (Barberis and Xiong, 2012; Imas, 2016). Narrow bracketing implies individuals receive a payoff based on whether a specific stock has gained or lost value (a zero reference point) instead of receiving payoffs based only on the value of their entire portfolio. Modelling welfare in the presence of narrow bracketing would require a normative judgment over not only whether reference-dependent payoffs deserve normative weight, but also the question of whether an individual should be bracketing at all (see e.g. Koch and Nafziger (2016)).

5 Conclusion

In this paper, we provide a first attempt at characterizing the welfare economics of reference dependence. Our most robust finding is that lowering reference points tends to improve welfare, even though different views of welfare provide different answers as to why this is the case. Other welfare effects, such as the welfare effects of prices or taxes, are more inherently ambiguous because evaluating them requires taking a stand on whether some individuals are over-consuming out of loss aversion.

Our empirical application highlights the real-world policy relevance of these results. Reference-dependent behavior has been documented in a number of empirical contexts, raising important questions of how reference dependence may affect the welfare consequences of different policies. In the context of retirement, we show that increasing the Normal Retirement Age is welfare-improving when it serves as a reference point in the labor supply/leisure dimension. The welfare effects of subsidies for later retirement, on the other hand, are more ambiguous and depend on normative judgments regarding reference dependence.

Taking our results at face value would suggest that lowering reference points to extreme degrees, or, in our empirical application, raising the Normal Retirement Age to an extremely high level, may be optimal. This notion should be taken with a hefty grain of salt. In practice, we can imagine several ways in which this admittedly extreme result may be disciplined. First, it may be that shifting statutory retirement ages, or other policies influencing reference points, to extreme levels would cause individuals to stop using the policy as a reference point. Relatedly, individuals may be insensitive to extreme reference points due to some form of self-regulation or bounded rationality, and the costs of self-regulation might themselves carry normative importance (Goldin and Reck, 2018). Additionally, in some settings decision-makers may deliberately exercise control over their own reference points due to concerns that are not present in the simplest model we considered; we showed that this can lead them to prefer higher reference points. These concerns could include utility from anticipation of future gains (Sarver, 2012; Thakral and Tô, 2020a), or the motive to over-come self-control problems by setting goals (Koch and Nafziger, 2011). Thus, we believe our results are likely to be useful in applications considering policies that shift reference points locally, in the absence of countervailing concerns like self-control problems or anticipation. Understanding larger, global changes requires further research, as does analysis of whether and when individuals deliberately control their own reference points. Finally, reference points set by public policy are often linked to other policies in practice, and it may not always be realistic to influence reference points in isolation from other factors. For instance, pension reforms that increase the NRA often feature large pension benefit cuts due to an institutional linkage of benefit levels to the NRA.

More broadly, our results here demonstrate that embracing normative ambiguity can provide a way forward for difficult problems at the core of behavioral economics. The question of whether behavioral phenomena arise due to behavioral biases or non-standard normative preferences has so far hindered applications of behavioral economics to welfare in several domains. Nevertheless, policy interest in behavioral economics has grown extremely rapidly in recent years, and, as in other policy settings, careful analysis of the welfare effect of policy changes can inform and discipline the policy debate. Embracing normative ambiguity can illuminate the path forward because it lets us separate questions that can be empirically analyzed, such as the influence of a change in reference point or prices on behavior, from normative judgments. This distinction between normative judgments and empirically estimable effects actually has a long tradition in public economics when it comes to questions of equity and efficiency (Mirrlees, 1971; Saez, 2001).

Finally, our work here demonstrates that the wide variety of models of reference dependence does pose a significant challenge for welfare analysis. This difficulty ultimately reflects the broader problem of a lack of parsimony in behavioral economics. We take a few lessons away from confronting this challenge. First, our attempts to make the model empirically implementable using standard formulations of reference dependence led to a number of new testable hypotheses. Specifically, alternative formulations of reference-dependent preferences make detailed predictions about 1) how reducing or eliminating reference dependence affects behavior, and 2) the willingness to pay to change a reference point. Testing these would be would be informative for welfare, and interesting in its own right. Second, our work on extensions of the model leaves room for much future research. Our experience suggests that further developing the normative aspects of the theory of reference dependence will not only inform optimal policy, but it will also generate more interesting empirical questions.

References

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

Allcott, H. and Kessler, J. B. (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics*, 11(1):236–76.

Allcott, H., Lockwood, B. B., and Taubinsky, D. (2019). Regressive sin taxes, with an application to the optimal soda tax. *Quarterly Journal of Economics*, 23(3):1557–1626.

Allcott, H. and Taubinsky, D. (2015). Evaluating behaviorally motivated policy: Experimental evidence from the lightbulb market. *American Economic Review*, 105(8):2501–38.

Allen, E. J., Dechow, P. M., Pope, D. G., and Wu, G. (2017). Reference-Dependent Preferences: Evidence from Marathon Runners. *Management Science*, 63(6):1657–1672.

Barberis, N. and Xiong, W. (2012). Realization utility. *Journal of Financial Economics*, 104(2):251–271.

Barseghyan, L., Molinari, F., O'Donoghue, T., and Teitelbaum, J. C. (2013). The nature of risk preferences: Evidence from insurance choices. *American Economic Review*, 103(6):2499–2529.

Behaghel, L. and Blau, D. M. (2012). Framing social security reform: Behavioral responses to changes in the full retirement age. *American Economic Journal: Economic Policy*, 4(4):41–67.

Bernheim, B. D., Fradkin, A., and Popov, I. (2015). The welfare economics of default options in 401(k) plans. *American Economic Review*, 105(9):2798–2837.

Bernheim, B. D. and Rangel, A. (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124(1):51–104.

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

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

Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. (1997). Labor supply of new york city cabdrivers: One day at a time. *Quarterly Journal of Economics*, 112(2):407–441.

Crawford, V. P. and Meng, J. (2011). New york city cab drivers' labor supply revisited: Reference-dependent preferences with rational-expectations targets for hours and income. *American Economic Review*, 101(5):1912–32.

Cribb, J., Emmerson, C., and Tetlow, G. (2016). Signals matter? large retirement responses to limited financial incentives. *Labour Economics*, 42:203–212.

De Martino, B., Camerer, C. F., and Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences*, 107(8):3788–3792.

DellaVigna, S., Lindner, A., Reizer, B., and Schmieder, J. F. (2017). Reference-dependent job search: Evidence from hungary. *Quarterly Journal of Economics*, 132(4):1969–2018.

DellaVigna, S. and Malmendier, U. (2006). Paying not to go to the gym. *American Economic Review*, 96(3):694–719.

Einav, L., Finkelstein, A., and Cullen, M. R. (2010). Estimating welfare in insurance markets using variation in prices. *Quarterly Journal of Economics*, 125(3):877–921.

Ericson, K. M. and Fuster, A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *Quarterly Journal of Economics*, 126(4):1879–1907.

Fehr, E. and Goette, L. (2007). Do workers work more if wages are high? evidence from a randomized field experiment. *American Economic Review*, 97(1):298–317.

Forschungsdatenzentrum der Rentenversicherung (FDZ-RV) (2015). Versichertenrentenzugang 1992 – 2014. Research Data Center of the German State Pension Fund (last accessed 2020-08-14).

Fudenberg, D. and Levine, D. K. (2006). A dual-self model of impulse control. *The American Economic Review*, pages 1449–1476.

Gali, J. (1994). Keeping up with the joneses: Consumption externalities, portfolio choice, and asset prices. *Journal of Money, Credit and Banking*, 26(1):1–8.

Gneezy, U., Goette, L., Sprenger, C., and Zimmermann, F. (2017). The limits of expectations-based reference dependence. *Journal of the European Economic Association*, 15(4):861–876.

Goette, L., Harms, A., and Sprenger, C. (2020). Randomizing endowments: an experimental study of rational expectations and reference-dependent preferences. *American Economic Journal: Microeconomics, forthcoming*.

Goldin, J. and Reck, D. (2018). Rationalizations and mistakes: optimal policy with normative ambiguity. In *AEA Papers and Proceedings*, volume 108, pages 98–102.

Goldin, J. and Reck, D. (2020). Optimal defaults with normative ambiguity. *Review of Economics and Statistics*.

Hardie, B. G., Johnson, E. J., and Fader, P. S. (1993). Modeling loss aversion and reference dependence effects on brand choice. *Marketing science*, 12(4):378–394.

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

Homonoff, T. A. (2018). Can small incentives have large effects? the impact of taxes versus bonuses on disposable bag use. *American Economic Journal: Economic Policy*, 10(4):177–210.

Imas, A. (2016). The realization effect: Risk-taking after realized versus paper losses. *American Economic Review*, 106(8):2086–2109.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, pages 263–291.

Kahneman, D., Wakker, P. P., and Sarin, R. (1997). Back to bentham? explorations of experienced utility. *Quarterly Journal of Economics*, 112(2):375–406.

Kermer, D. A., Driver-Linn, E., Wilson, T. D., and Gilbert, D. T. (2006). Loss aversion is an affective forecasting error. *Psychological science*, 17(8):649–653.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8:435–464.

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

Koch, A. K. and Nafziger, J. (2016). Goals and bracketing under mental accounting. *Journal of Economic Theory*, 162:305–351.

Köszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics*, 121(4).

Kőszegi, B. and Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073.

Manoli, D. S. and Weber, A. (2016). The effects of the early retirement age on retirement decisions. Technical report, NBER Working Paper.

Masatlioglu, Y. and Raymond, C. (2016). A behavioral analysis of stochastic reference dependence. *American Economic Review*, 106(9):2760–82.

Mastrobuoni, G. (2009). Labor supply effects of the recent social security benefit cuts: Empirical estimates using cohort discontinuities. *Journal of public Economics*, 93(11-12):1224–1233.

Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. *The Review of Economic Studies*, 38(2):175–208.

Mullainathan, S., Schwartzstein, J., and Congdon, W. J. (2012). A reduced-form approach to behavioral public finance. *Annu. Rev. Econ.*, 4(1):511–540.

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

OECD (2019). Pensions at a glance 2019. Oecd database.

Rees-Jones, A. (2018). Quantifying loss-averse tax manipulation. *The Review of Economic Studies*, 85(2):1251–1278.

Rick, S. (2011). Losses, gains, and brains: Neuroeconomics can help to answer open questions about loss aversion. *Journal of Consumer Psychology*, 21(4):453–463.

Rosch, E. (1975). Cognitive reference points. *Cognitive psychology*, 7(4):532–547.

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., Davison, C., Demić, E., Esteban-Serna, C., Friedemann, M., et al. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, pages 1–12.

Saez, E. (2001). Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, 68(1):205–229.

Saez, E. and Stantcheva, S. (2016). Generalized social marginal welfare weights for optimal tax theory. *American Economic Review*, 106(1):24–45.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71.

Sarver, T. (2012). Optimal reference points and anticipation. Working paper.

Seibold, A. (2020). Reference points for retirement behavior: Evidence from german pension discontinuities. *forthcoming, American Economic Review*.

Sokol-Hessner, P., Camerer, C. F., and Phelps, E. A. (2013). Emotion regulation reduces loss aversion and decreases amygdala responses to losses. *Social cognitive and affective neuroscience*, 8(3):341–350.

Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F., and Phelps, E. A. (2009). Thinking like a trader selectively reduces individuals' loss aversion. *Proceedings of the National Academy of Sciences*, 106(13):5035–5040.

Sokol-Hessner, P. and Rutledge, R. B. (2019). The psychological and neural basis of loss aversion. *Current Directions in Psychological Science*, 28(1):20–27.

Staubli, S. and Zweimüller, J. (2013). Does raising the early retirement age increase employment of older workers? *Journal of public economics*, 108:17–32.

Thakral, N. and Tô, L. T. (2020a). Anticipation and consumption. Working paper.

Thakral, N. and Tô, L. T. (2020b). Daily labor supply and adaptive reference points. *American Economic Review*.

Thaler, R. (1985). Mental accounting and consumer choice. *Marketing science*, 4(3):199–214.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*.

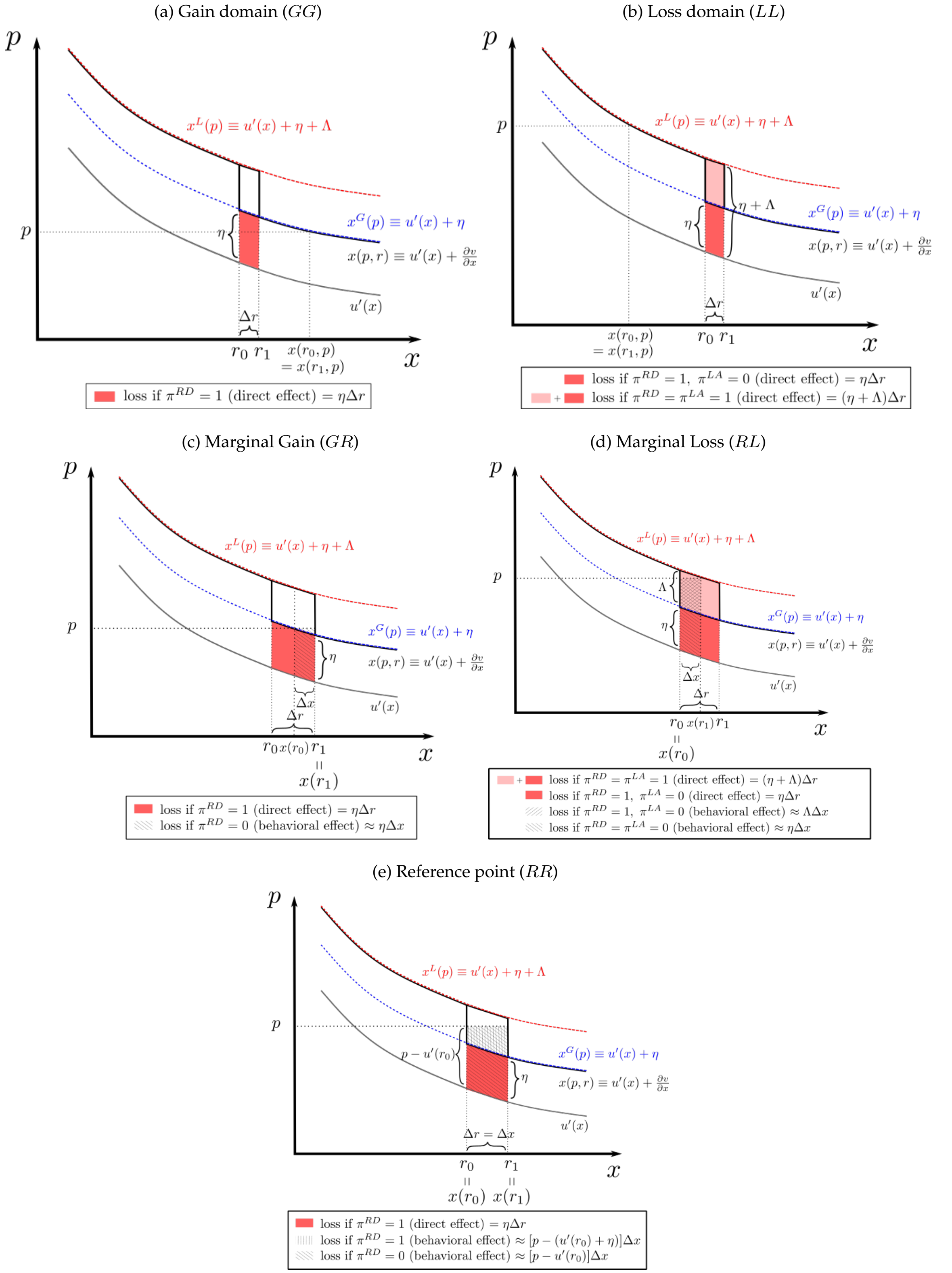
Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics*, 106(4):1039–1061.

Appendix

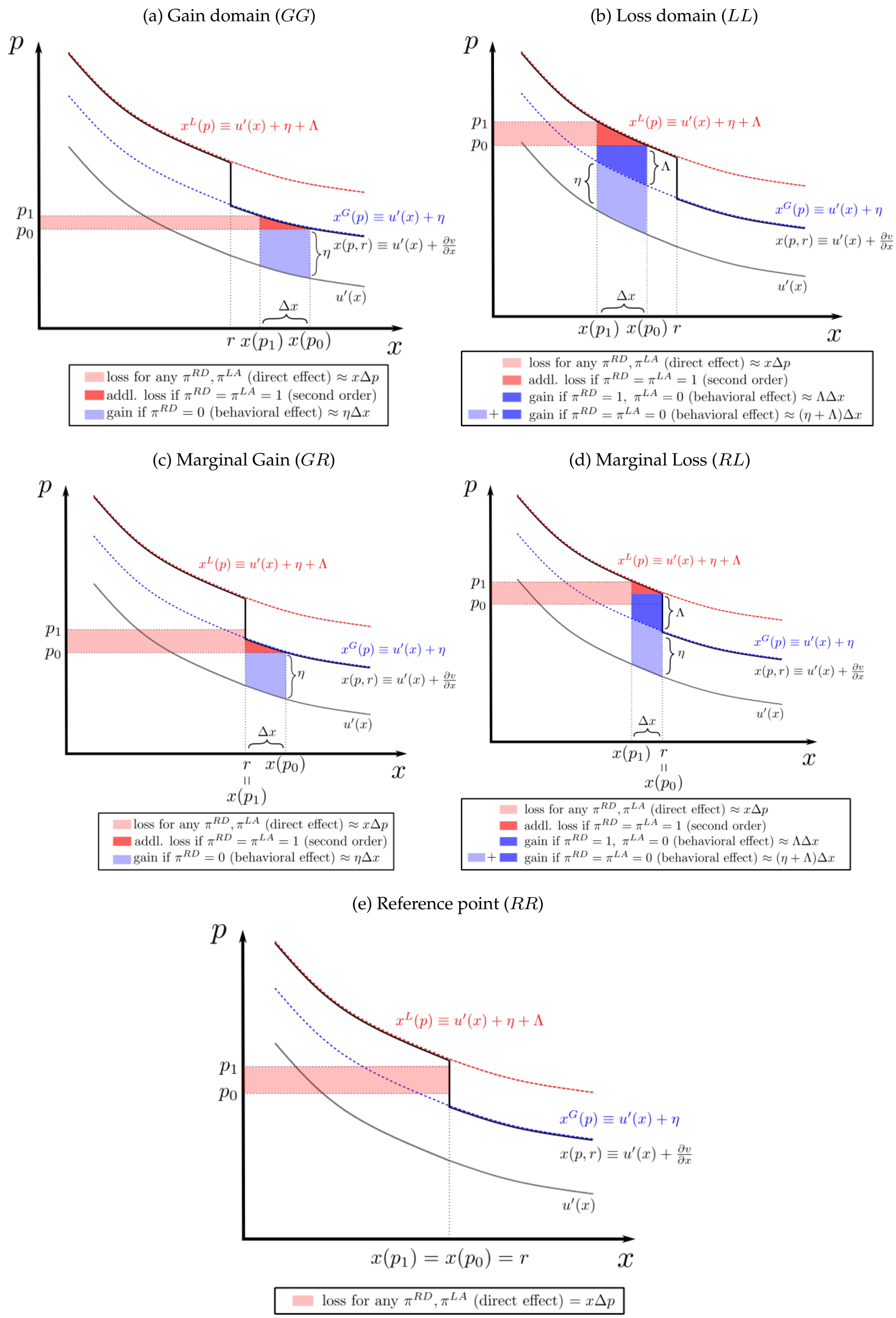
A Additional Figures and Tables

FIGURE A1: REFERENCE POINT WELFARE EFFECTS UNDER THE TVERSKY-KAHNEMAN (1991) FORM



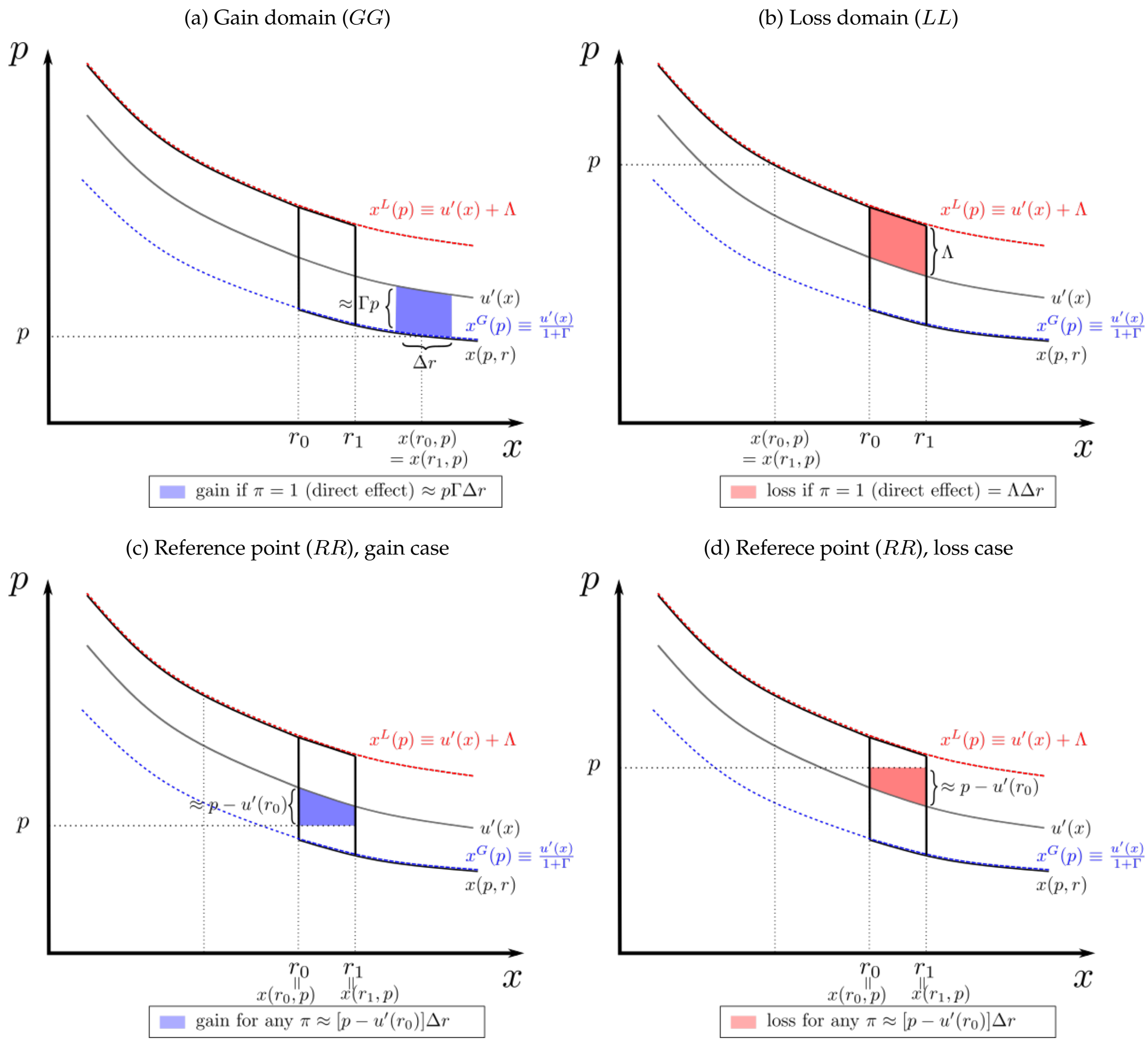
Note: This figure plots the welfare effect of changing the reference point in each domain. Unlike in Figure 2, we adopt the formulation of reference dependence including the η parameter from Tversky and Kahneman (1991) - see Equation: (21). We denote observed demand in black, marginal utility $u'(x)$ in grey, and gain and loss domain demand in blue and red, respectively, as in Figure 8. The direct effect of a change in the reference point is depicted with red shaded regions and the behavioral welfare effects are depicted with hatching. In this case the size of the direct effect can depend on both π^{RD} and π^{LA}. Refer to Section 4.1 for further details.

FIGURE A2: WELFARE EFFECTS OF PRICE CHANGES UNDER THE TVERSKY-KAHNEMAN (1991) FORM



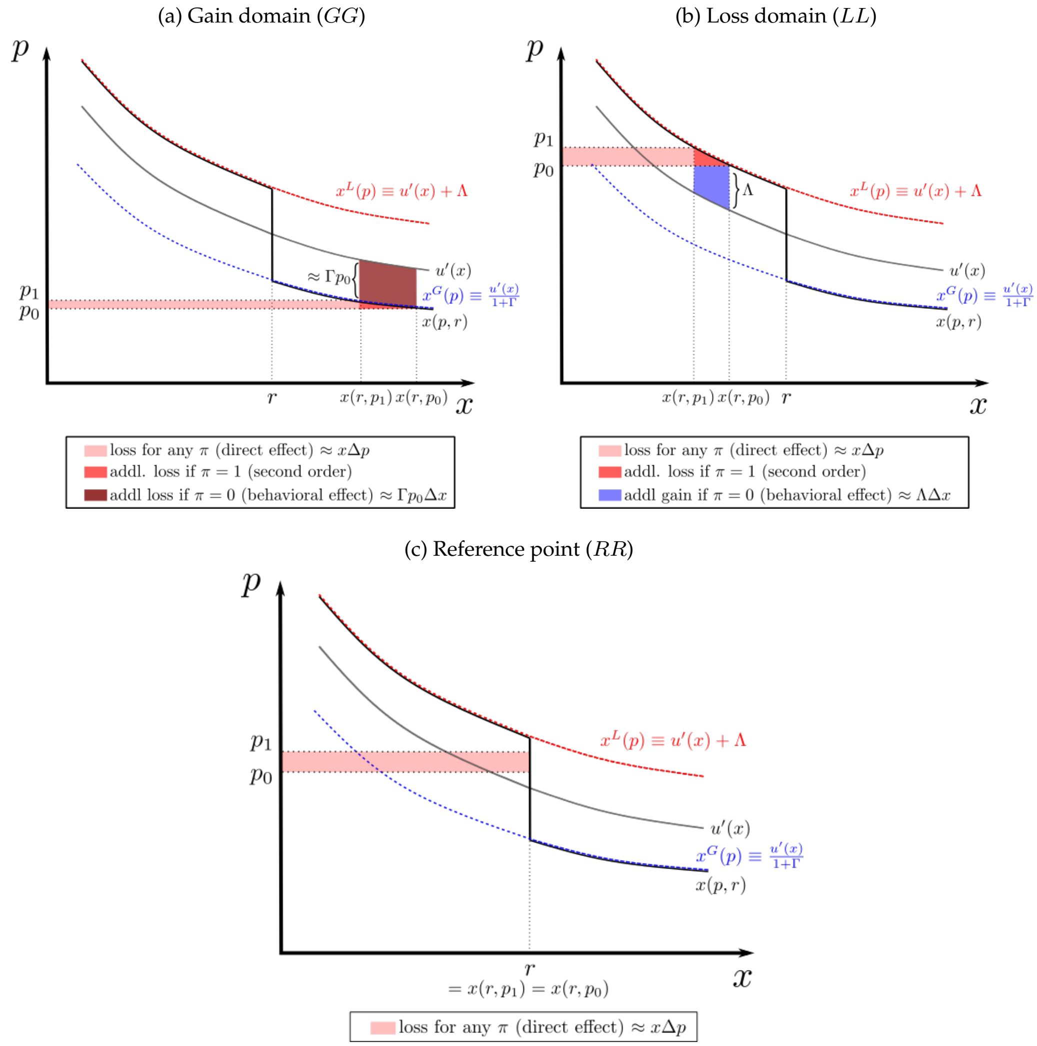
Note: This figure plots the welfare effect of changing the price in each domain. Unlike in Figure 3, we adopt the formulation of reference dependence including the η parameter from Tversky and Kahneman (1991) - see Equation: (21). We denote observed demand in black, marginal utility $u'(x)$ in grey, and gain and loss domain demand in blue and red, respectively, as in Figure 8. The direct (mechanical) negative effect of the price change is depicted with red shaded regions, and the positive behavioral effect is plotted in blue. In this case the size of the behavioral welfare effect depends on both π^{RD} and π^{LA}. Refer to Section 4.1 for further details.

FIGURE A3: REFERENCE POINT WELFARE EFFECTS UNDER 2-DIMENSIONAL REFERENCE DEPENDENCE



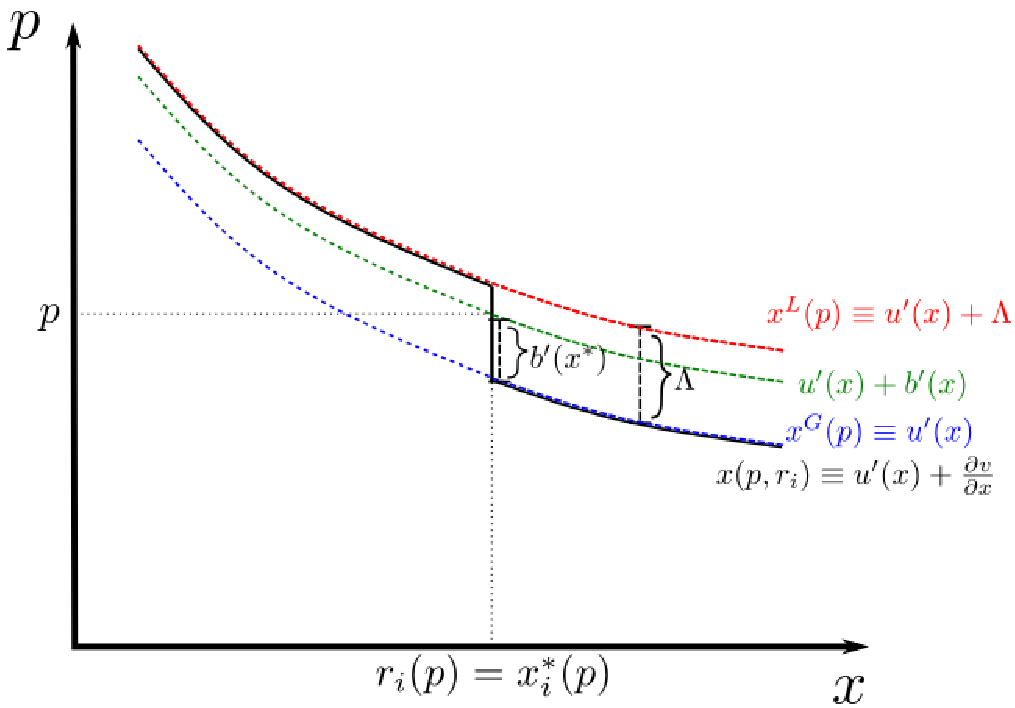
Note: This figure plots the welfare effect of changing the reference point in each domain. Unlike in Figure 2, we assume reference dependence over both x and the background good y. We include only those cases that are relevant for first-order social welfare for simplicity. We also depict the RR case in two situations: those in which the individual experiences a gain or loss. We denote observed demand in black, marginal utility $u'(x)$ in grey, and gain and loss domain demand in blue and red, respectively, as in Figure 9. Gains are depicted in blue shaded regions and losses in red. Because the size of the direct effect on the G group in equation 34 depends on the price, we can no longer illustrate the direct effect of a change in the reference point in the figures like Figure 2 or A1, or Panel (b) of this figure. We depict this quantity slightly differently in Panel (a) for this reason. Refer to Section 4.2for further details.

FIGURE A4: WELFARE EFFECTS OF PRICE CHANGES UNDER 2-DIMENSIONAL REFERENCE DEPENDENCE



Note: This figure plots the welfare effect of changing the price in each domain. Unlike in Figure 2, we assume reference dependence over both x and the background good y. We include only those cases that are relevant for first-order social welfare for simplicity. We denote observed demand in black, marginal utility $u'(x)$ in grey, and gain and loss domain demand in blue and red, respectively, as in Figure 9. Gains are depicted in blue shaded regions and losses in red. The only difference with Figure 3 is the presence of an additional loss from the behavioral effect in the gain domain, depicted in dark red in panel (a). Refer to Section 4.2 for further details.

FIGURE A5: THE OPTIMAL REFERENCE POINT UNDER GOAL-SETTING



Note: This figure illustrates the optimal choice of reference point given an additional bias unrelated to reference dependence, $b(x)$, and reference dependent preferences over x, for the model in Section 4.3. Under the assumption that $b'(x^*) < \Lambda$, which can be seen in the figure, the individual sets a reference point r_i to completely correct the bias and makes a choice in the R domain. Analyzing welfare building on the result illustrated here leads straightforwardly to the results described in the text of section 4.3.

Table A1: Bunching and Parameter Estimates

Panel A: Bunching Estimates			
	(1)	(2)	(3)
	Excess mass	Kink size	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)

Note: 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 (2020) 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 (20), using the bunching estimates across thresholds summarized in Panel A. See the main text for more details of the estimation.

B Proofs

Lemma 1. The Marginal Internality

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

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

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

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

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

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

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

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

Proposition 1. The Desirability of Low Reference Points

- **P 1.1.**

  Case 1: $i \in GG$

  $$w_i(p, r_1) - w_i(p, r_0) = 0 \;\Rightarrow\; \Delta w_i(\pi) \leq 0$$

  Case 2: $i \in GR$

  $$w_i(p, r_1) - w_i(p, r_0) = u_i(r_1) - u_i(x_i^G) - p(r_1 - x_i^G)$$

  Holding p fixed, since the individual's consumption switches from the G domain to the R domain as r becomes r_1, we necessarily have $x_i^G < r_1$.
  Since u_i is strictly concave, $u_i(r_1) - u_i(x_i^G) < u_i'(x_i^G)(r_1 - x_i^G)$.
  Following the first-order condition of the individual's welfare-maximizing program, $p = u_i'(x_i^G)$. Then, $u_i(r_1) - u_i(x_i^G) < p(r_1 - x_i^G)$, which implies $\Delta w_i(\pi) < 0$.

  Case 3: $i \in GL$

  $$w_i(p, r_1) - w_i(p, r_0) = u_i(x_i^L) - u_i(x_i^G) - p(x_i^L - x_i^G) + \pi\Lambda_i(x_i^L - r_1)$$

  Holding p fixed, since the individual's consumption switches from the G domain to the L domain as r becomes r_1, we necessarily have $r_1 > x_i^L > x_i^G > r_0$.
  Since u_i is strictly concave, $u_i(x_i^L) - u_i(x_i^G) < u_i'(x_i^G)(x_i^L - x_i^G)$.
  Following the first-order condition of the individual's welfare-maximizing program, $p = u_i'(x_i^G)$. Then,

  $$u_i(x_i^L) - u_i(x_i^G) < p(x_i^L - x_i^G) \;\Rightarrow\; u_i(x_i^L) - u_i(x_i^G) - p(x_i^L - x_i^G) + \pi\Lambda_i(x_i^L - r_1) < \pi\Lambda_i(x_i^L - r_1) \leq 0$$

Then, for all π, $\Delta w_i(\pi) < 0$
Case 4: $i \in RR$

$$w_i(p, r_1) - w_i(p, r_0) = u_i(r_1) - u_i(r_0) - p(r_1 - r_0)$$

Holding p fixed, since the individual's consumption remains in the R domain as r becomes r_1, we necessarily have $u_i'(r_0) \leq p \leq u_i'(r_1) + \Lambda$.
Since u_i is strictly concave,

$$u_i(r_1) - u_i(r_0) < u_i'(r_0)(r_1 - r_0) \;\Rightarrow\; u_i'(r_0) > \frac{u_i(r_1) - u_i(r_0)}{r_1 - r_0} \;\Rightarrow\; p > \frac{u_i(r_1) - u_i(r_0)}{r_1 - r_0} \text{ as } p \geq u_i'(r_0)$$

This implies that $u_i(r_1) - u_i(r_0) - p(r_1 - r_0) < 0$, thus $\Delta w_i(\pi) < 0$.
Case 5: $i \in RL$

$$w_i(p, r_1) - w_i(p, r_0) = u_i(x_i^L) - u_i(r_0) - p(x_i^L - r_0) + \pi\Lambda_i(x_i^L - r_1)$$

Holding p fixed, since the individual's consumption switches from the R domain to the L domain as r becomes r_1, we necessarily have $r_1 > x_i^L > r_0$.
Moreover, this setting imposes the two following conditions on p :

$$u_i'(r_0) + \Lambda_i \geq p \geq u_i'(r_1) + \Lambda_i \quad \text{that defines the domain } RL$$

$$u_i'(r_0) + \Lambda_i \geq p \geq u_i'(r_0) \;\; \text{that defines the domain such that the individual is in the } R \text{ domain at } r = r_0$$

Since u_i is strictly concave, $u_i(x_i^L) - u_i(r_0) < u_i'(r_0)(x_i^L - r_0)$.
Then, as $p \geq u_i'(r_0)$ by second condition it must be that

$$u_i(x_i^L) - u_i(r_0) < p(x_i^L - r_0) \;\Rightarrow\; u_i(x_i^L) - u_i(r_0) - p(x_i^L - r_0) + \pi\Lambda_i(x_i^L - r_1) < \pi\Lambda_i(x_i^L - r_1) \leq 0$$

Then, for all π, $\Delta w_i(\pi) < 0$
Case 6: $i \in LL$

$$w_i(p, r_1) - w_i(p, r_0) = -\pi\Lambda_i(r_1 - r_0) \quad \Rightarrow \quad \Delta w_i(\pi) \leq 0$$

- **P 1.2.** $\Delta w_i(1) = 0$ for $i \in GG$ and $\Delta w_i(1) < 0$ for $i \notin GG$. Then if $P[i \in GG] < 1$, there exists at least one individual, the one who does not belong to the GG group, who strictly loses from the increase of r and all others who are left as well off. Consequently, social welfare strictly decreases after the change in r.

$$\Delta W(1) = \int_i \Delta w_i(1)di = \int_{i\in GG} \Delta w_i(1)di + \int_{i\notin GG} \Delta w_i(1)di = \int_{i\notin GG} \Delta w_i(1)di < 0$$

  If $P[i \in GG] < 1$, r_0 Pareto dominates r_1 when $\pi = 1$ as all individuals who belong to the GG group are indifferent and all others lose in welfare.

- **P 1.3.** $\Delta w_i(0) = 0$ for $i \in GG \cup LL$ and $\Delta w_i(1) < 0$ otherwise. Then if $P[i \in GG] + P[i \in LL] < 1$, there exists at least one individual, the one who is not GG or LL, who strictly loses from the increase of r and all others who are left as well off. Consequently, social welfare strictly decreases after the change in r.

$$\Delta W(1) = \int_i \Delta w_i(1)di = \int_{i\in GG\cup LL} \Delta w_i(1)di + \int_{i\notin GG\cup LL} \Delta w_i(1)di = \int_{i\notin GG\cup LL} \Delta w_i(1)di < 0$$

If $P[i \in GG] + P[i \in LL] < 1$, r_0 Pareto dominates r_1 when $\pi = 1$ as all individuals GG or LL are indifferent and all others lose in welfare.

Proposition 2. First-Order Individual Welfare Effect of a Change in the Reference Point.
Recall that $w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r) + \pi v(x_i(p,r)|r)$.
Case 1: $i \in GG : w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r)$
By first-order Taylor series approximation :

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

And

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

As the demand x^G does not depend on r (according to FOC^G), the derivative $\frac{\partial x_i^G(p)}{\partial r}$ is null. Then, $\Delta w_i = 0$.
Case 2: $i \in GR : w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r)$
By first-order Taylor series approximation :

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

And

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

According to FOC^G, $u_i'(x_i^G) = p$. Then, $\Delta w_i = 0$.
Case 3: $i \in RR : w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r) + \pi\Lambda_i(x_i(p,r) - r)$
The kink at $x = r$ results in an indeterminate form for first-order approximation. However, independently of the change in r, the individual remains with a consumption level equal to his reference point in both situations. Then, there is no direct effect here and we only consider the behavioral effect of the change.
Then, $\Delta w_i = \Delta r(u'(r_0) - p)$.
Case 4: $i \in RL : w_i(p,r) = u(x_i(p,r)) + z_i - px_i(p,r) + \pi\Lambda_i(x_i(p,r) - r)$
By first-order Taylor series approximation :

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

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

And

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

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

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

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

And

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

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

Proposition 3. The First-Order Social Welfare Effect of a Change in the Reference Point.
We denote for all i, f_p and f_{Λ_i} the probability density functions respectively of $p|\Lambda_i$ and Λ_i. We also sometimes abbreviate for $k = G, L, R$,
$U_i^k = U_i^*(x_i^k, z_i - px_i^k)$.

$$\begin{aligned}
W = \int_0^{+\infty} \Bigg[&\int_{u_i'(r)+\Lambda_i}^{+\infty} U_i^*(x_i^L, z_i - px_i^L) f_p(p)dp + \int_{u_i'(r)}^{u_i'(r)+\Lambda_i} U_i^*(r, z_i - pr) f_p(p)dp \\
&+ \int_0^{u_i'(r)} U_i^*(x_i^G, z_i - px_i^G) f_p(p)dp \Bigg] f_{\Lambda_i}(\Lambda_i)d\Lambda_i
\end{aligned}$$

Applying Leibniz rule,

$$\begin{aligned}
\frac{\partial W}{\partial r} = &\int_0^{+\infty} \left[\int_{u_i'(r)+\Lambda_i}^{+\infty} \frac{\partial U_i^L}{\partial r} f_p(p)dp + \int_{u_i'(r)}^{u_i'(r)+\Lambda_i} \frac{\partial U_i^R}{\partial r} f_p(p)dp + \int_0^{u_i'(r)} \frac{\partial U_i^G}{\partial r} f_p(p)dp \right] f_{\Lambda_i}(\Lambda_i)d\Lambda_i \\
&+ \int_0^{+\infty} \Big[-u_i"(r)U_i^L|_{p=u_i'(r)+\Lambda_i} f_p(u_i'(r)+\Lambda_i) + u_i"(r)U_i^R|_{p=u_i'(r)+\Lambda_i} f_p(u_i'(r)+\Lambda_i) \\
&\quad -u_i"(r)U_i^R|_{p=u_i'(r)} f_p(u_i'(r)) + u_i"(r)U_i^G|_{p=u_i'(r)} f_p(u_i'(r)) \Big] f_{\Lambda_i}(\Lambda_i)d\Lambda_i
\end{aligned}$$

Focus on the second part of the expression. For all i,

$$
\begin{aligned}
&-u_i"(r)U_i^L|_{p=u_i'(r)+\Lambda_i}f_P(u_i'(r)+\Lambda_i)+u_i"(r)U_i^R|_{p=u_i'(r)+\Lambda_i}f_P(u_i'(r)+\Lambda_i)-u_i"(r)U_i^R|_{p=u_i'(r)}f_P(u_i'(r)) \\
&\qquad\qquad\qquad + u_i"(r)U_i^G|_{p=u_i'(r)}f_P(u_i'(r)) \\
&= u_i"(r)\left[f_P(u_i'(r)+\Lambda_i)(U_i^R|_{p=u_i'(r)+\Lambda_i}-U_i^L|_{p=u_i'(r)+\Lambda_i})+f_P(u_i'(r))(U_i^G|_{p=u_i'(r)}-U_i^R|_{p=u_i'(r)})\right]
\end{aligned}
$$

For all i, $p=u_i'(r)$ corresponds to the situation where the individual i is at the threshold between the G and R groups. At this point, the individual's utility is the same regardless of whether $x>r$ or $x\leq r$. Formally,

$$
\begin{aligned}
U_i^*(r,z_i-pr)|_{u_i'(r)=p} &= u_i(r)+z_i-u_i'(r)r \\
U_i^*(x_i^G,z_i-px_i^G)|_{u_i'(r)=p} &= u_i(x_i^G)+z_i-u_i'(x_i^G)x_i^G
\end{aligned}
$$

Since $r=x_i^G$ for i such that $p=u_i'(r)$, then $U_i^*(r,z_i-pr)|_{u_i'(r)=p}-U_i^*(x_i^G,z_i-px_i^G)|_{u_i'(r)=p}=0$.
Similarly for the R and L groups, $U_i^*(x_i^L,z_i-px_i^L)|_{p=u_i'(r)+\Lambda_i}-U_i^*(r,z_i-pr)|_{p=u_i'(r)+\Lambda_i}=0$.
Consequently,

$$
\begin{aligned}
\frac{\partial W}{\partial r} &= \int_0^{+\infty}\left[\int_{u_i'(r)+\Lambda_i}^{+\infty}\frac{\partial U_i^L}{\partial r}f_P(p)dp+\int_{u_i'(r)}^{u_i'(r)+\Lambda_i}\frac{\partial U_i^R}{\partial r}f_P(p)dp+\int_0^{u_i'(r)}\frac{\partial U_i^G}{\partial r}f_P(p)dp\right]f_{\Lambda_i}(\Lambda_i)d\Lambda_i \\
&= \int_0^{+\infty}\left[\int_{u_i'(r)+\Lambda_i}^{+\infty}\frac{\partial U_i^L}{\partial r}f_P(p)dp+\int_{u_i'(r)}^{u_i'(r)+\Lambda_i}\frac{\partial U_i^R}{\partial r}f_P(p)dp\right]f_{\Lambda_i}(\Lambda_i)d\Lambda_i \\
&= \int_0^{+\infty}\left[\int_{u_i'(r)+\Lambda_i}^{+\infty}-\pi\Lambda_i f_P(p)dp+\int_{u_i'(r)}^{u_i'(r)+\Lambda_i}(u_i'(r)-p)f_P(p)dp\right]f_{\Lambda_i}(\Lambda_i)d\Lambda_i \\
&= \int_0^{+\infty}\left\{E\left[-\pi\Lambda_i|p>u_i'(r)+\Lambda_i,\Lambda_i\right]\right\}P(p>u_i'(r)+\Lambda_i|\Lambda_i)(f_{\Lambda_i}(\Lambda_i)d\Lambda_i \\
&\quad -\int_0^{+\infty}\left\{E\left[p-u_i'(r)|u_i'(r)>p>u_i'(r)+\Lambda_i,\Lambda_i\right]\right\}P(u_i'(r)>p>u_i'(r)+\Lambda_i|\Lambda_i)f_{\Lambda_i}(\Lambda_i)d\Lambda_i
\end{aligned}
$$

By law of iterated expectations,

$$
\frac{\partial W}{\partial r}=-\pi E\left[\Lambda_i|i\in L\right]P\left[i\in L\right]-E\left[p-u_i'(r)|i\in R\right]P\left[i\in R\right]
$$

Which yields the result by first-order Taylor series approximation.

Proposition 4. First-Order Welfare Effect of a Change in Price.
Case 1: $i\in GG,GR$

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

And

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

As

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

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

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

Case 2: $i \in LL, RL$

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

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

And

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

As

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

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

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

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

Case 3: $i \in RR$

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

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

Proposition 5. The First-Order Social Welfare Effect of a Price Change.

We denote for all i, f_u and f_{Λ_i} the probability density functions respectively of $u_i'(r)|\Lambda_i$ and Λ_i. We also sometimes abbreviate for $k = G, L, R$, $U_i^k = U_i^*(x_i^k, z_i - px_i^k)$.

$$W = \int_0^{+\infty} \left[\int_{-\infty}^{p-\Lambda_i} U_i^*(x_i^L, z_i - px_i^L) f_u(u_i'(r)) du_i'(r) + \int_{p-\Lambda_i}^{p} U_i^*(r, z_i - pr) f_u(u_i'(r)) du_i'(r) \right.$$
$$\left. + \int_p^{+\infty} U_i^*(x_i^G, z_i - px_i^G) f_u(u_i'(r)) du_i'(r) \right] f_{\Lambda_i}(\Lambda_i) d\Lambda_i$$

Applying Leibniz rule,

$$\frac{\partial W}{\partial p} = \int_0^{+\infty} \left[\int_{-\infty}^{p-\Lambda_i} \frac{\partial U_i^L}{\partial p} f_u(u_i'(r)) du_i'(r) + \int_{p-\Lambda_i}^{p} \frac{\partial U_i^R}{\partial p} f_u(u_i'(r)) du_i'(r) + \int_p^{+\infty} \frac{\partial U_i^G}{\partial p} f_u(u_i'(r)) du_i'(r) \right] f_{\Lambda_i}(\Lambda_i) d\Lambda_i$$
$$+ \int_0^{+\infty} \left[U_i^L|_{u_i'(r)=p-\Lambda_i} f_u(p-\Lambda_i) + U_i^R|_{u_i'(r)=p} f_u(p) - U_i^R|_{u_i'(r)=p-\Lambda_i} f_u(p-\Lambda_i) - U_i^G|_{u_i'(r)=p} f_u(p) \right] f_{\Lambda_i}(\Lambda_i) d\Lambda_i$$

Focus on the second part of the expression. For all i,

$$U_i^L|_{u_i'(r)=p-\Lambda_i} f_u(p-\Lambda_i) + U_i^R|_{u_i'(r)=p} f_u(p) - U_i^R|_{u_i'(r)=p-\Lambda_i} f_u(p-\Lambda_i) - U_i^G|_{u_i'(r)=p} f_u(p)$$
$$= f_u(p) \left[U_i^R(r, z_i - pr)|_{u_i'(r)=p} - U_i^G(x_i^G, z_i - px_i^G)|_{u_i'(r)=p} \right] + f_u(p-\Lambda_i) \left[U_i^L(x_i^L, z_i - px_i^L)|_{u_i'(r)=p-\Lambda_i} \right.$$
$$\left. - U_i^*(r, z_i - pr)|_{u_i'(r)=p-\Lambda_i} \right]$$

For all i, $p = u_i'(r)$ corresponds to the situation where the individual i is at the threshold between the G and R groups. At this point, the individual's utility is the same regardless of whether $x > r$ or $x \leq r$. Formally,

$$U_i^*(r, z_i - pr)|_{u_i'(r)=p} = u_i(r) + z_i - u_i'(r)r$$
$$U_i^*(x_i^G, z_i - px_i^G)|_{u_i'(r)=p} = u_i(x_i^G) + z_i - u_i'(x_i^G)x_i^G$$

Since $r = x_i^G$ for i such that $p = u_i'(r)$, then $U_i^*(r, z_i - pr)|_{u_i'(r)=p} - U_i^*(x_i^G, z_i - px_i^G)|_{u_i'(r)=p} = 0$.
Similarly for the R and L groups, $U_i^*(x_i^L, z_i - px_i^L)|_{u_i'(r)=p-\Lambda_i} - U_i^*(r, z_i - pr)|_{u_i'(r)=p-\Lambda_i} = 0$.
Consequently,

$$\int_0^{+\infty} \left[U_i^L|_{u_i'(r)=p-\Lambda_i} f_u(p-\Lambda_i) + U_i^R|_{u_i'(r)=p} f_u(p) - U_i^R|_{u_i'(r)=p-\Lambda_i} f_u(p-\Lambda_i) - U_i^G|_{u_i'(r)=p} f_u(p) \right] f_{\Lambda_i}(\Lambda_i) d\Lambda_i = 0$$

The social welfare effect now writes as :

$$\frac{\partial W}{\partial p} = \int_0^{+\infty} \left[\int_{-\infty}^{p-\Lambda_i} \frac{\partial U_i^L}{\partial p} f_u(u_i'(r)) du_i'(r) + \int_{p-\Lambda_i}^{p} \frac{\partial U_i^R}{\partial p} f_u(u_i'(r)) du_i'(r) + \int_p^{+\infty} \frac{\partial U_i^G}{\partial p} f_u(u_i'(r)) du_i'(r) \right] f_{\Lambda_i}(\Lambda_i) d\Lambda_i$$
$$= \int_0^{+\infty} \left\{ \int_{-\infty}^{p-\Lambda_i} \left[\frac{\partial x_i^L}{\partial p} (u_i'(x_i^L) - p + \pi\Lambda_i) - x_i^L \right] f_u(u_i'(r)) du_i'(r) + \int_{p-\Lambda_i}^{p} -r f_u(u_i'(r)) du_i'(r) \right.$$
$$\left. + \int_p^{+\infty} \left[\frac{\partial x_i^G}{\partial p} (u_i'(x_i^G) - p) - x_i^G \right] f_u(u_i'(r)) du_i'(r) \right\} f_{\Lambda_i}(\Lambda_i) d\Lambda_i$$

By FOC^G, $u_i'(x_i^G) - p = 0$ and by FOC^L, $u_i'(x_i^L) = p - \Lambda_i$

$$
\begin{aligned}
&= \int_0^{+\infty} \left\{ \int_{-\infty}^{p-\Lambda_i} \left[-(1-\pi)\Lambda_i \frac{\partial x_i^L}{\partial p} - x_i^L \right] f_u(u_i'(r)) du_i'(r) - \int_{p-\Lambda_i}^{p} r f_u(u_i'(r)) du_i'(r) \right. \\
&\qquad \left. + \int_p^{+\infty} -x_i^G f_u(u_i'(r)) du_i'(r) \right\} f_{\Lambda_i}(\Lambda_i) d\Lambda_i \\
&= \int_0^{+\infty} \left\{ \int_{-\infty}^{p-\Lambda_i} -(1-\pi)\Lambda_i \frac{\partial x_i^L}{\partial p} f_u(u_i'(r)) du_i'(r) - \int_{-\infty}^{+\infty} x_i f_u(u_i'(r)) du_i'(r) \right\} f_{\Lambda_i}(\Lambda_i) d\Lambda_i \\
&= \int_0^{+\infty} \left\{ E\left[-(1-\pi)\Lambda_i \frac{\partial x_i^L}{\partial p} | u_i'(r) < p - \Lambda_i, \Lambda_i \right] P(u_i'(r) < p - \Lambda_i | \Lambda_i) - E\left[x_i | \Lambda_i \right] \right\} f_{\Lambda_i}(\Lambda_i) d\Lambda_i \\
&= \int_0^{+\infty} E\left[-(1-\pi)\Lambda_i \frac{\partial x_i^L}{\partial p} | i \in L, \Lambda_i \right] P(i \in L | \Lambda_i) f_{\Lambda_i}(\Lambda_i) d\Lambda_i - \int_0^{+\infty} E\left[x_i | \Lambda_i \right] f_{\Lambda_i}(\Lambda_i) d\Lambda_i
\end{aligned}
$$

By law of iterated expectations

$$
\frac{\partial W}{\partial p} = E\left[-(1-\pi)\Lambda_i \frac{\partial x_i^L}{\partial p} | i \in L \right] P\left[i \in L \right] - E\left[x_i \right]
$$

Which yields the proposition result, after a first-order Taylor series approximation.

Corollary 5.1. Corrective Taxes for Reference Dependence.

- $x \geq r$: Denoting x_i^G the solution to the individual's welfare-maximizing program, the tax-comprehensive individual welfare objective function writes

  $$U_i^*(x_i^G, z_i - px_i^G, t_i(x_i^G, r)) = u_i(x_i^G) + z_i - px_i^G - t_i(x_i^G, r)$$

  The efficient individual tax t_i on x solves

  $$
  \begin{aligned}
  \frac{\partial U_i^*}{\partial x_i^G} = 0 &\Rightarrow u_i'(x_i^G) - p - \frac{\partial t_i}{\partial x_i^G} = 0 \\
  &\Rightarrow \frac{\partial t_i}{\partial x_i^G} = 0 \quad \text{by } FOC^G
  \end{aligned}
  $$

  Then, for $x \geq r$, $t_i(x, r) = 0$.

- $x < r$: Denoting x_i^L the solution to the individual's welfare-maximizing program, the tax-comprehensive individual welfare objective function writes

  $$U_i^*(x_i^L, z_i - px_i^L, t_i(x_i^L, r)) = u_i(x_i^L) + z_i - px_i^L + \pi\Lambda_i(x_i^L - r) - t_i(x_i^L, r)$$

The efficient individual tax t_i on x solves

$$\frac{\partial U_i^*}{\partial x_i^L} = 0 \Rightarrow u_i'(x_i^L) - p + \pi\Lambda_i - \frac{\partial t_i}{\partial x_i^L} = 0$$
$$\Rightarrow -(1-\pi)\Lambda_i - \frac{\partial t_i}{\partial x_i^L} = 0 \quad \text{by } FOC^L$$
$$\Rightarrow \frac{\partial t_i}{\partial x_i^L} = -(1-\pi)\Lambda_i$$

Then, for $x < r$, $t_i(x, r) = -(1-\pi)\Lambda_i(x - r)$.

Note that the first-order conditions are different in the K-T formulation model.

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

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

Lemma 2. The Revised Marginal Internality.
L 2.1 $x_i(p, r) > r$

$$\tilde{m_i}^G = \frac{\partial \tilde{U}_i(x, z - px)}{\partial x}|_{x=x^G} = \tilde{u_i}'(x^G) - p + \pi^{RD}\eta_i = -(1-\pi^{RD})\eta_i \quad \text{by } FOC^G$$

L 2.2 $x_i(p, r) < r$

$$\tilde{m_i}^L = \frac{\partial \tilde{U}_i(x, z - px)}{\partial x}|_{x=x^L} = \tilde{u_i}'(x^L) - p + \pi^{RD}\eta_i + \pi^{LA}\Lambda_i = -(1-\pi^{RD})\eta_i - (1-\pi^{LA})\Lambda_i \quad \text{by } FOC^L$$

L 2.3 $x_i(p, r) = r$
Under this model, $\tilde{m_i}$ is undefined for $x = r$ if $\pi^{LA} = 1$. Otherwise,

$$\tilde{m_i} = \tilde{u_i}'(r) - p + \pi^{RD}\eta_i$$

Proposition 6. Modified First-Order Welfare Effects.
P 6.1 $i \in GG$

$$\Delta\tilde{w}_i = \tilde{u}_i(x^G(p)) - \tilde{u}_i(x^G(p)) - p(x^G(p) - x^G(p)) - \pi^{RD}\eta_i(r_1 - r_0) = -\pi^{RD}\eta_i(r_1 - r_0)$$

equal to the desired result as $\Delta x = 0$.

$i \in GR$

$$
\begin{aligned}
\Delta\tilde{w_i} &= \tilde{u_i}(r_1) - \tilde{u_i}(x^G(p)) - p(r_1 - x^G(p)) - \pi^{RD}\eta_i(x^G(p) - r_0)\\
&\quad by\, Taylor\ approximation \Rightarrow \tilde{u_i}(r_1) \approx \tilde{u_i}(x^G(p)) + \tilde{u_i}'(x^G(p))(r_1 - x^G(p))\\
\Rightarrow \Delta\tilde{w_i} &\approx \tilde{u_i}'(x^G(p))(r_1 - x^G(p)) - p(r_1 - x^G(p)) - \pi^{RD}\eta_i(x^G(p) - r_0)\\
&\approx -\eta_i(r_1 - x^G(p)) - \pi^{RD}\eta_i(x^G(p) - r_0) \quad \text{by } FOC^G\\
&\approx -(1-\pi^{RD})\eta_i(r_1 - x^G(p)) - \pi^{RD}\eta_i(r_1 - r_0) \quad \text{by adding and removing } \pi^{RD}\eta_i r_1.\\
&\approx \Delta w_i - \pi^{RD}\eta_i\Delta r - (1-\pi^{RD})\eta_i\Delta x
\end{aligned}
$$

$i \in RR$

$$
\begin{aligned}
\Delta\tilde{w_i} &= \tilde{u_i}(r_1) - \tilde{u_i}(r_0) - p(r_1 - r_0)\\
&\approx (\tilde{u_i}'(r_0) - p)(r_1 - r_0) \quad \text{by first-order Taylor approximation}\\
&\approx (\tilde{u_i}'(r_0) - p - \eta_i)(r_1 - r_0) \quad \text{as } \tilde{u_i}' = \tilde{u_i}' - \eta_i \text{ for } i \in RR
\end{aligned}
$$

As $\Delta x = \Delta r$ in this case, we obtain the desired result.
LL and RL cases are analogous to GG and GR.
P 6.2 Analogously and using the same notations as in Proposition 3 proof,

$$
\begin{aligned}
\tilde{W} = \int_0^{+\infty} \Bigg[&\int_{\tilde{u_i}'(r)+\Lambda_i}^{+\infty} \tilde{U_i}^*(x_i^L, z_i - px_i^L)f_p(p)dp + \int_{\tilde{u_i}'(r)}^{\tilde{u_i}'(r)+\Lambda_i} \tilde{U_i}^*(r, z_i - pr)f_p(p)dp\\
&+ \int_0^{\tilde{u_i}'(r)} \tilde{U_i}^*(x_i^G, z_i - px_i^G)f_p(p)dp\Bigg] f_{\Lambda_i}(\Lambda_i)d\Lambda_i\\
\Rightarrow \frac{\partial\tilde{W}}{\partial r} &= \int_0^{+\infty}\left[\int_{\tilde{u_i}'(r)+\Lambda_i}^{+\infty} \frac{\partial\tilde{U_i^L}}{\partial r} f_p(p)dp + \int_{\tilde{u_i}'(r)}^{\tilde{u_i}'(r)+\Lambda_i} \frac{\partial\tilde{U_i^R}}{\partial r} f_p(p)dp + \int_0^{\tilde{u_i}'(r)} \frac{\partial\tilde{U_i^G}}{\partial r} f_p(p)dp\right] f_{\Lambda_i}(\Lambda_i)d\Lambda_i\\
&= \int_0^{+\infty}\Bigg[\int_{\tilde{u_i}'(r)+\Lambda_i}^{+\infty} -(\pi^{RD}\eta_i + \pi^{LA}\Lambda_i)f_p(p)dp + \int_{\tilde{u_i}'(r)}^{\tilde{u_i}'(r)+\Lambda_i} (\tilde{u_i}'(r) - p)f_p(p)dp\\
&\quad + \int_0^{\tilde{u_i}'(r)} -\pi^{RD}\eta_i f_p(p)dp\Bigg] f_{\Lambda_i}(\Lambda_i)d\Lambda_i\\
&= \int_0^{+\infty}\left\{E\left[-(\pi^{RD}\eta_i + \pi^{LA}\Lambda_i)|p > \tilde{u_i}'(r) + \Lambda_i, \Lambda_i\right]\right\}P(p > \tilde{u_i}'(r) + \Lambda_i|\Lambda_i)(f_{\Lambda_i}(\Lambda_i)d\Lambda_i\\
&\quad - \int_0^{+\infty}\left\{E\left[p - \tilde{u_i}'(r)|\tilde{u_i}'(r) + \Lambda_i > p > \tilde{u_i}'(r), \Lambda_i\right]\right\}P(\tilde{u_i}'(r) + \Lambda_i > p > \tilde{u_i}'(r)|\Lambda_i)f_{\Lambda_i}(\Lambda_i)d\Lambda_i\\
&\quad + \int_0^{+\infty}\left\{E\left[-\pi^{RD}\eta_i|\tilde{u_i}'(r) > p, \Lambda_i\right]\right\}P(\tilde{u_i}'(r) > p|\Lambda_i)f_{\Lambda_i}(\Lambda_i)d\Lambda_i\\
&= \int_0^{+\infty}\Bigg\{E\left[-\pi^{LA}\Lambda_i|i \in L, \Lambda_i\right]P(i \in L|\Lambda_i) - E\left[p - u_i'(r)|i \in R, \Lambda_i\right]P(i \in R|\Lambda_i)\\
&\quad - E\left[\pi^{RD}\eta_i|\Lambda_i\right] - E\left[(1-\pi^{RD})\eta_i|i \in R, \Lambda_i\right]P(i \in R|\Lambda_i)\Bigg\}f_{\Lambda_i}(\Lambda_i)d\Lambda_i
\end{aligned}
$$

Since $\tilde{u_i}'(r) - p = u_i'(r) - p - \eta_i$ from the FOC^R in both models.

By law of iterated expectations,

$$\frac{\partial \tilde{W}}{\partial r} = -\pi^{LA}\Lambda_i E\left[\Lambda_i | i \in L\right] P\left[i \in L\right] - E\left[p - \tilde{u_i}'(r) | i \in R\right] P\left[i \in R\right] - E\left[\pi^{RD}\eta_i\right]$$
$$- E\left[(1-\pi^{RD})\eta_i | i \in R\right] P(i \in R)$$

Then, as π from the first model is now π^{LA},

$$\Rightarrow \frac{\partial \tilde{W}}{\partial r} = \frac{\partial W}{\partial r} - E[\pi^{RD}\eta_i] - E[(1-\pi^{RD})\eta_i | i \in R] P[i \in R]$$

Which yields the result after two first-order Taylor approximations.

P 6.3 The result proven in Proposition 4 holds for the new model:

$$\tilde{w_i}(p_1, r) - \tilde{w_i}(p_0, r) \approx \tilde{m_i}(\hat{p}, r)\Delta x - x(\hat{p}, r)\Delta p$$

Then the difference between the two variations in indirect utility is all due to the combination of the marginal internality and the behavioral effect :

$$\Delta \tilde{w_i} - \Delta w_i \approx (\tilde{m_i}(p, r) - m_i(p, r))\Delta x_i$$

For $i \notin RR$, $\Delta x_i \neq 0$ and the difference in marginal internalities is

$$\tilde{m_i}(p, r) - m_i(p, r) = -(1 - \pi^{RD})\eta_i$$

For $i \in RR$, $\Delta x = 0$, then the result stands independently of the marginal internalities.

P 6.4 Analogously and using similar notations as in Proposition 5 proof,

$$\tilde{W} = \int_0^{+\infty} \left[\int_{-\infty}^{p-\Lambda_i} \tilde{U_i}^*(x_i^L, z_i - p x_i^L) f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r) + \int_{p-\Lambda_i}^{p} \tilde{U_i}^*(r, z_i - pr) f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r)\right.$$
$$\left. + \int_p^{+\infty} \tilde{U_i}^*(x_i^G, z_i - p x_i^G) f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r)\right] f_{\Lambda_i}(\Lambda_i) d\Lambda_i$$

$$\frac{\partial W}{\partial p} = \int_0^{+\infty} \left[\int_{-\infty}^{p-\Lambda_i} \frac{\partial \tilde{U_i^L}}{\partial p} f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r) + \int_{p-\Lambda_i}^{p} \frac{\partial \tilde{U_i^R}}{\partial p} f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r)\right.$$
$$\left. + \int_p^{+\infty} \frac{\partial \tilde{U_i^G}}{\partial p} f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r)\right] f_{\Lambda_i}(\Lambda_i) d\Lambda_i$$
$$= \int_0^{+\infty} \left[\int_{-\infty}^{p-\Lambda_i} \left[\frac{\partial x_i^L}{\partial p}(\tilde{u_i}'(x_i^L) - p + \pi^{RD}\eta_i + \pi^{LA}\Lambda_i) - x_i^L\right] f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r) + \int_{p-\Lambda_i}^{p} -r f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r)\right.$$
$$\left. + \int_p^{+\infty} \left[\frac{\partial x_i^G}{\partial p}(\tilde{u_i}'(x_i^G) - p + \pi^{RD}\eta_i) - x_i^G\right] f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r)\right] f_{\Lambda_i}(\Lambda_i) d\Lambda_i$$

By FOC^G and FOC^L,

$$= \int_0^{+\infty} \left[\int_{-\infty}^{p-\Lambda_i} \left[\frac{\partial x_i^L}{\partial p} - ((1-\pi^{RD})\eta_i + (1-\pi^{LA})\Lambda_i) - x_i^L \right] f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r) + \int_{p-\Lambda_i}^{p} -r f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r) \right.$$
$$\left. + \int_p^{+\infty} \left[-\frac{\partial x_i^G}{\partial p}(1-\pi^{RD})\eta_i - x_i^G \right] f_{\tilde{u}}(\tilde{u_i}'(r)) d\tilde{u_i}'(r) \right] f_{\Lambda_i}(\Lambda_i) d\Lambda_i$$
$$= \int_0^{+\infty} \left\{ E\left[-((1-\pi^{RD})\eta_i + (1-\pi^{LA})\Lambda_i) \frac{\partial x_i^L}{\partial p} | \tilde{u_i}'(r) < p - \Lambda_i, \Lambda_i \right] P(\tilde{u_i}'(r) < p - \Lambda_i | \Lambda_i) - E\left[x_i | \Lambda_i \right] \right.$$
$$\left. - E\left[(1-\pi^{RD})\eta_i \frac{\partial x_i^G}{\partial p} | \tilde{u_i}'(r) > p, \Lambda_i \right] P(\tilde{u_i}'(r) > p | \Lambda_i) \right\} f_{\Lambda_i}(\Lambda_i) d\Lambda_i$$

By law of iterated expectations,

$$\frac{\partial \tilde{W}}{\partial p} = E\left[-(1-\pi^{LA})\Lambda_i \frac{\partial x_i^L}{\partial p} | \tilde{u_i}'(r) < p - \Lambda_i \right] P(\tilde{u_i}'(r) < p - \Lambda_i) - E\left[(1-\pi^{RD})\eta_i) \frac{\partial x_i^G}{\partial p} | \tilde{u_i}'(r) > p \right] P(\tilde{u_i}'(r) > p)$$
$$- E\left[x_i \right] - E\left[(1-\pi^{RD})\eta_i) \frac{\partial x_i^L}{\partial p} | \tilde{u_i}'(r) < p - \Lambda_i \right] P(\tilde{u_i}'(r) < p - \Lambda_i)$$
$$= \frac{\partial W}{\partial p} - E\left[(1-\pi^{RD})\eta_i \frac{\partial x_i}{\partial p} \right] \quad \text{as } \frac{\partial x_i}{\partial p} = 0 \text{ in the } R \text{ group.}$$

Which yields the proposition result after two first-order Taylor approximations.

Proposition 7. First-Order Social Welfare Effects in the 2-D Model.
In this model, welfare is given by:

$$w_i(p,r) = u_i(x(p,r)) + z_i - p_x + \pi \left[1\{x(p,r) < r\} \Lambda_i (x(p,r) - r) + 11\{x(p,r) > r\} \Gamma_i p(z_i - px(p,r) - s(r)) \right] \tag{37}$$

Note that we presume $s(r) = z_i - pr$ but we disregard $\partial s / \partial p$, so that a change in p does not effect the reference point s and cause a direct effect. Taking derivatives of this expression with respect to r or p for the G, R, L, substituting for $u_i'(x)$ using the FOCs for the G and L case from equations (31) and (32), and integrating over the three first-order groups for social welfare, we obtain the desired results.

C The Behavioral Equivalence of Alternative Formulations

In the main text, we considered two formulations of reference dependence. A key fact for our analysis was that these models are behaviorally indistinguishable using observed choices (i.e. demand at given prices and reference points), but that they carry somewhat different implications for welfare. In this Appendix, we formalize the sense in which the models are behaviorally indistinguishable. Crucially, the main reason that these models are indistinguishable is that we assume that choices of x given a reference point (and price and endowment), but we do not observe choices or revealed preferences over chosen options and reference points, i.e. (x, r), jointly. This assumption is consistent with what is observed in typical applications of models of reference-dependent preferences, but it might be relaxed in more stylized experiments. We note that a similar equivalence, implying that the η_i parameter is typically unidentified, is shown for the stochastic case in Barseghyan et al. (2013).

In Section 2 of the main text, we mainly analyzed the following model of behavior, which we will here call *Model 1*:

$$x_i(p,r,z) = \arg\max_x u_i(x) + z - px + 1\{x < r\}\Lambda_i(x - r), \tag{38}$$

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

In Section 4.1 and very briefly earlier on in equation (3), we considered an alternative model in line with the formulation of reference dependence proposed by Tversky and Kahneman (1991), which we here call *Model 2*.

$$x_i(p,r,z) = \arg\max_x \tilde{u}_i(x) + z - px + \begin{cases} \eta_i(x-r) & x > r \\ \eta_i\lambda_i(x-r) & x \le r, \end{cases} \tag{39}$$

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

Consider a demand function $x(p,r,z)$, which describes the choice of x the consumer makes for any (p,r,z). We say $x(p,r,z)$ is *rationalizable* with either model if there are utility functions and parameters such that the optimization problem the model describes generates the observed behavior for any (p,r,z). That is, $x(p,r,z)$ is rationalizable by model 1 if and only if there is a utility function $u(x)$ with $u' > 0, u'' < 0$ and a parameter $\Lambda_i > 0$ such that for any (p,r,z) equation (38) obtains. We say $x(p,r,z)$ is rationalizable by model 2 under similar conditions.

We make one more modest technical assumption for our desired result to obtain, which is that the domain of good x is compact. In Model 1, this ensures that $u'(x)$ has a strictly positive minimum for all values of x, which we denote $\epsilon \equiv \min u'(x)$. The assumption ensures $\epsilon > 0$ exists. Why we need this assumption will become clear in the proof of the result below.

Proposition 8. ***Behavioral equivalence of Model 1 and Model 2.*** *A demand function $x_i(p,r,z)$ is rationalizable by Model 1 if and only if it is rationalizable by Model 2.*

Corollary 8.1. ***A behavioral isomorphism.*** *If $x_i(p,r,z)$ is rationalizable by model 1 with utility $u_i(x)$ and parameter Λ_i and rationalizable by model 2 with utility $\tilde{u}_i(x)$ and parameters η_i, λ_i, then we must have*

$$u_i(x) = \tilde{u}_i(x) + \eta_i x. \tag{40}$$

$$\Lambda_i = \eta_i(\lambda_i - 1). \tag{41}$$

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

Set any η_i such that $0 < \eta_i < \epsilon$.[32] Specify $\tilde{u}_i$ according to equation (40), i.e. $\tilde{u}_i = u_i(x) - \eta_i x$. Specify λ_i according to equation (41), i.e. $\lambda_i = \frac{\Lambda_i + \eta_i}{\eta_i}$.

Because $u' > \eta_i$ for any x by construction, we know that $\tilde{u}_i' = u_i' - \eta_i >= u_i' - \varepsilon > 0$, and $u'' < 0 \implies \tilde{u}_i'' < 0$. Further, by construction $\eta_i > 0$ and $\lambda_i > 1$. With the necessary restrictions satisfied, we only need to show that with these specifications, the optimization problem in equation (38) is equivalent to the optimization problem in (39). As we have guaranteed equations (40) and (41) hold, we can re-express the optimization problem in model 1 as:

$$x_i(p,r,z) = \arg\max_x \tilde{u}_i(x) + \eta_i x + z - px + 1\{x < r\}\eta_i(\lambda_i - 1)(x - r), \tag{42}$$

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

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

For the converse, suppose that $x_i(p, r, z)$ is rationalizable by model 2 with utility function $\tilde{u}_i(x)$ and parameters $\eta_i > 0$, and $\lambda_i > 1$. Specify $u_i(x)$ using equation (40) and set Λ using (41). Checking the restrictions, we know that $\tilde{u}'_i > 0$ $\eta_>0$, implying that $u'_i = \tilde{u}'_i + \eta > 0$, and $u''_i = \tilde{u}''_i < 0$. And we know that $\Lambda_i > 0$ by $\eta_i > 0$ and $\lambda_i > 1$. We can re-express the optimization problem in Model 2 as

$$x_i(p, r, z) = \arg\max_x \tilde{u}_i(x) + \eta_i x + z - px + 1x > r\eta_i(\lambda - 1)(x - r) - \eta r. \quad (43)$$

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

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

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

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

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

$$u(x, y, r, f_1) = u_i(x) + y + v_i(x|r), \quad (44)$$

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

In order to map our analysis into the Bernheim-Rangel framework, we need to consider a hypothetical choice situation in which reference dependence were eliminated. If we wish to consider the possibility that reference dependence may be a bias, what preferences would be revealed by choices in an unbiased state? We represent choices made in a no-reference-dependence state by encoding a frame f_0. Choices under f_0

maximize

$$u_i(x, y, r, f_0) = u_i(x) + y. \tag{45}$$

Obviously, choices under f_0 are difficult to directly observe in positive empirical analysis, but the application of the Bernheim-Rangel framework does not require that all relevant parts of the choice correspondence are empirically observable. Choices under f_0 could potentially be observed by eliminating the effect of the reference point through some experimental intervention, or by inducing individuals to use an arbitrarily low reference point (recall that $v_i = 0$ in the gain domain).

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

$$u_i(x, y, r, f) = u_i(x) + y + f * v_i(x|r), \tag{46}$$

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

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

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

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

If we interpret our results within the Bernheim-Rangel framework, the content of the results is mainly to show how the these alternative judgments about the welfare-relevant domain influence welfare and optimal policy considerations. Under (J1) or (J2), there is a single utility function (either equation (44) or equation (45)) that ranks all options in the menu space (i.e. all combinations of (x, y, r)). Under (J3), however, we obtain only an incomplete ranking. Recall that we used the word "robust" above to describe situations where whether one situation or the other was better for welfare did not depend on π. Our results map into the Bernheim-Rangel framework as follows:

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

As discussed in the main text, we find that decreases in the reference point tend to improve welfare for either value of π. Through the lens of the Bernheim-Rangel framework, this suggests that even if we include choices under f_1 and f_0 in the welfare relevant domain (J3) and use the revealed preference criterion proposed by Bernheim and Rangel, we would conclude that individuals prefer lower reference points.

Note that because $v_i = 0$ in the gain domain, we find that holding the reference point fixed, equations (44) and (45) express the same preferences in the gain domain. Moreover the reference point has no direct on welfare in the gain domain for either value of f or π. If we restrict our attention to individuals making choices in the gain domain, therefore, all welfare comparisons will be robust. The only potential deviations from revealed preference come from individuals choosing at the reference point or in the loss domain. This finding is the basis for the statement in Section 2 that we respect revealed preference in the gain domain. The potential deviations from revealed preferences in the naturally occurring frame that we consider are material for individuals at the reference point or in the loss domain only. Obviously, the extended models in Section 4 do not necessarily have this property. Importing those extended models requires modifications to the above - for instance we would need to introduce three frames rather than two to import the model in Section 4.1 into the Bernheim-Rangel framework - but the general structure of how we could give those models a behavioral revealed preference interpretation remains the same.