Optimal Defaults with Normative Ambiguity

Jacob Goldin Daniel Reck*

February 10, 2020

Abstract

Abstract: Default effects are pervasive, but the reason they arise is often unclear. We study optimal policy when the planner does not know whether an observed default effect reflects a welfare-relevant preference or a mistake. Within a broad class of models, we find that determining optimal policy is impossible without resolving this ambiguity. Depending on the resolution, optimal policy tends in opposite directions: either minimizing the number of non-default choices or inducing active choice. We show how these considerations depend on whether active choosers make mistakes when selecting among non-default options. We illustrate our results using data on pension contribution defaults.

EconLit Codes: J2, J3, H2

[*] Goldin: Stanford Law School and NBER; email: jsgoldin@law.stanford.edu. Reck: Department of Economics, London School of Economics; email: d.h.reck@lse.ac.uk. We thank Ian Ayres, Doug Bernheim, Stefano DellaVigna, Alex Gelber, Wojciech Kopczuk, Claus Kreiner, Alex Rees-Jones, Camille Landais, Emmanuel Saez, Dmitry Taubinsky, Eric Talley, and seminar participants at Berkeley, the University of Copenhagen, the University of California San Diego, the Institute for Fiscal Studies, Vanderbilt, VATT Finland, the NTA Annual Conference, the IIPF annual congress, and the ASSA annual meeting for helpful discussion and comments. Jeanne Bomare and Jonas Mueller-Gastell provided excellent research assistance.

A fundamental challenge in behavioral economics is determining whether some observed behavior reflects a mistake by decision-makers. If the behavior violates the predictions of a neoclassical model of decision-making, one can typically modify the preferences of the agents in the model to rationalize the observed decisions. Doing so can realign the model's predictions with the data, but introduces normative ambiguity; that is, do the modified preferences actually reflect the decision-maker's welfare or do they instead reflect a divergence between welfare and choice? Conducting welfare analysis in behavioral settings requires confronting this issue.

We study this problem for the case of default effects.[1] We consider two types of mistakes that potentially cause welfare and choice to diverge. First, decision-makers may misperceive which option maximizes their welfare; for example, there is a heated debate over whether individuals save too little for retirement (Ghilarducci, 2019). Second, decision-makers may act as if there is some cost to choosing an option other than the default even when this apparent cost, if incurred, would not actually affect decision-makers' welfare. In theory, these sources of normative ambiguity might be resolved by identifying the precise structural model that generates behavior. In practice, distinguishing between alternative behavioral models is difficult in most settings in which default effects are observed. Prior research that studies optimal policy with default effects has addressed the issue either by assuming a specific behavioral model (Carroll et al., 2009) or by considering robustness to several alternative models (Bernheim, Fradkin and Popov, 2015). Neither of these approaches shed light on which features of the behavioral model drive the optimal policy results, making it difficult to extrapolate the results to settings in which other behavioral models may be at play, or settings in which additional policy tools are available.

The starting point for our approach is that for a broad class of decision-making models, the effect of defaults on behavior can be characterized in terms of two ingredients: (1) decision-makers' utility over the menu of available options, and (2) an "as-if" cost to selecting an option that is not the default (i.e., to making an active choice). This implied cost to opting out of the default is defined so as to rationalize decision-makers' observed behavior; decision-makers behave *as if* opting out of the default causes them to incur a cost of this magnitude. Unlike standard revealed preference models, we do not impose that as-if costs actually reduce welfare for those who choose to incur them. Instead, we parameterize the degree to which as-if costs are normative (that is, the degree to which they enter into decision-makers' welfare). Thus, the first type of *normative ambiguity* we study corresponds to uncertainty in the degree to which as-if costs are normative.

Alternative behavioral models imply different conclusions about the degree to which as-if costs are normative. For example, one explanation for default effects is that decision-makers rationally seek to avoid

[1] A large empirical literature documents the effect of defaults on behavior. Prominent examples include Madrian and Shea (2001); Johnson and Goldstein (2003); Haggag and Paci (2014); Handel (2013); Beshears et al. (2017).

exerting the mental effort required to choose between non-default options. In this model, all as-if costs are normative. Alternatively, decision-makers might seek to avoid exerting mental effort, but systematically over-estimate the amount of effort required to choose between non-default options. In this model, some – but not all – as-if costs are normative. A third possibility is that decision-makers inadvertently fail to consider making an active decision in the first place, in which case none of the observed as-if costs are normative. Although rejecting certain behavioral models might be possible with the right data, it is difficult to conceive of a convincing empirical test for determining the share of as-if costs that enter into decision-makers' welfare. This dilemma is worsened by the fact that there may be heterogeneity among decision-makers in the model that generates their behavior, and hence, in the degree to which their as-if costs are normative. Because resolving these issues empirically is challenging, normative ambiguity tends to arise whenever default effects are observed.

We use our framework to characterize the optimal default in terms of three components: the distribution of (1) decision-makers' preferences over the available options; (2) as-if costs; and (3) the share of as-if costs that are normative. When these components are known, the optimal default can be determined without additional knowledge of the underlying behavioral model (at least within the class of models we study). Standard revealed preferences techniques can be used to recover the first two components, but not the third. Hence, our proposed approach is to identify (1) and (2) from observed choice data and then to characterize the optimal default as a function of (3), based on the plausible range of behavioral models in the setting at hand.

We show that when as-if costs are mostly non-normative, the optimal policy induces decision-makers to make an active choice. Depending on the setting, the planner can implement this policy directly, by eliminating the presence of any default option from the decision, or indirectly, by setting as the default an option that decision-makers find sufficiently undesirable. In contrast, when as-if costs are sufficiently normative, forcing active choice is not only undesirable, doing so actually *minimizes* social welfare. Instead, we show that a better approach in this case is to set a default that leads relatively few decision-makers to opt out; doing so results in many people receiving an option that is close to the option they most prefer and few people incurring the (normative) costs of opting out. Optimal policy in this case resembles a rule of thumb that has been proposed in the literature: minimizing the number of opt-outs (Thaler and Sunstein, 2003). Our results therefore clarify the conditions under which this rule of thumb yields desirable policies.

The second source of normative ambiguity we consider is whether decision-makers choose optimally when selecting a non-default option. When active choices are sub-optimal, opting out of a default can reduce the quality of the option that one selects. In such cases, the optimal default depends on decision-makers' true (unbiased) preferences but also accounts for the possibility that decision-makers opt-out when the default is set too far from the outcome they (incorrectly) perceive to be their ideal. Such mistakes reduce the desirability of policies that promote active choice.

We illustrate our approach by applying it to data on employee contribution decisions to a 401(k) retirement plan. We characterize the optimal default as a function of the degree to which employees over-weight opt-out costs and misperceive their privately optimal saving rate. For the firm we study, we illustrate quantitatively our theoretical claim that one cannot identify the optimal policy without taking a stance on these issues. When employees' active savings decisions are optimal, the critical threshold in our data is whether the normative share of as-if costs is less than 8 percent of total as-if costs – about $160 for the median employee. When the normative component of as-if costs is below this threshold, the optimal policy induces employees to make an active contribution decision. In contrast, when the normative component of as-if costs exceeds this threshold, the optimal default is the contribution rate that minimizes employee opt-outs, which, for this firm, corresponds to the rate that maximizes the employer match. In addition, if employees under-save for reasons unrelated to the default, the optimal policy tilts away from active choice and toward setting the default to a relatively high contribution rate. Finally, the two types of mistakes we consider interact in important ways: the further opt-out decisions are from optimal, the more the optimal policy depends on the degree of under-saving by active employees.

Our results contribute to a growing literature on the welfare economics of default options. A closely-related paper to ours is Bernheim, Fradkin and Popov (2015) ("BFP"), which studies the optimal contribution rate default for an employer-provided 401(k) retirement plan. BFP consider a range of potential behavioral models for default effects and estimate the optimal policy within each model using employee contribution data. They find that the optimal default is quite stable across models for the firms in their data, suggesting that normative ambiguity does not pose a major challenge to identifying the optimal default.

Our results build on BFP in several important respects. First, we use our model to clarify the settings in which policymakers should promote active choices. In such cases, we show that determining the optimal policy is impossible without resolving normative ambiguity to some degree. This result has practical importance: in our empirical 401(k) application, for example, we find that the optimal policy plausibly promotes active choice rather than minimizing employee opt-outs (as BFP concluded). A related advantage of our approach is that it allows one to quantify in dollars the threshold of normative opt-outs costs for which promoting opt-outs becomes optimal.[2]

The second way in which we build on BFP is by generalizing the level of abstraction to highlight which

[2] BFP do consider one policy that has the effect of inducing active choice, which is to tax the default option. However, they do not consider how the desirability of this policy varies based on the underlying model. Instead, they point out that a tax on the default can yield large welfare losses if some decision-makers exogenously remain passive. That concern applies to some, but not all, forms of encouraging active choice; e.g., it would not apply to policies that simply remove the default, as in Carroll et al. (2009).

features of a behavioral model matter for shaping optimal policy.[3] We model default effects at a high level of generality so that our results are consistent with a broad class of positive models that might generate default effects, including the specific models that BFP consider.[4] This approach yields three payoffs: First, the lessons we draw for optimal policy emerge from general features of the problem rather than the specific models one happens to consider. For example, we show the generic reason why BFP find some robustness in the optimal default: normative ambiguity tends not to affect the optimal default when active choice policies are ruled out and preferences over contribution rates are sufficiently well-behaved. Second, because our framework is not tied to a specific positive model, it can easily incorporate heterogeneity in the model that explains an observed default effect or even uncertainty over the correct positive model. Third, the generality of our approach dramatically simplifies some key features of the problem, making our framework transparent and easy to apply. In fact, our approach has already been fruitfully applied to study optimal 401(k) defaults within a dynamic setting (Choukhmane, 2019).

Our third contribution is to consider settings in which decision-makers make mistakes for reasons unrelated to default effects. Such mistakes are frequently invoked to argue for using defaults to shape behavior, such as automatic enrollment into 401(k) plans to combat under-saving by employees (e.g., Thaler and Sunstein, 2008; Camerer et al., 2003). To our knowledge, however, the prior theoretical work on optimal defaults excludes "internalities" of this form. This omission is surprising, given that some of the proposed mechanisms by which default effects operate – such as present-bias – might also imply biased decision-making by those selecting non-default options, depending on the decision being observed.

Two other papers are also closely related to our own. Carroll et al. (2009) was the first to study when policies that force active choice are preferable to setting a default. Within a model in which present bias magnifies the cost of opting out of the default, they show that the desirability of active choice depends on the degree of time inconsistency that decision-makers exhibit.[5] We extend this result to settings in which decision-makers are sensitive to the default for reasons unrelated to present bias and to models in which active choosers make mistakes. More recently, Chesterley (2017) also studies the welfare effect of default options but focuses on a different set of policies than the ones we consider, such as policies that vary the cost of selecting a non-default option. In contrast, Chesterley's setup is not designed for studying normative ambiguity: he assumes the social planner knows the degree to which observed as-if costs are normative, and the only behavioral model he considers is one in which default effects are magnified because of present bias.

[3] BFP's approach to welfare analysis is first, to select a positive model, and second, to identify the optimal policy within that model (holding fixed the positive model itself). Within positive models, they do vary the welfare criteria, generating some normative ambiguity within but not between models.

[4] As described in Section 1, the as-if cost representation we study captures all but the anchoring model considered by BFP. We show in the Online Appendix that our key results extend to that model as well.

[5] The focus of Carroll et al. (2009) is an empirical comparison of active choice to opt-in 401(k) plan design; they do not apply their theoretical results to data.

Another related strand of the literature attempts to disentangle mechanisms by which default effects operate. A few recent papers examine the implications of inattention for estimates of switching costs, often in the context of choosing a health insurance plan (Abaluck and Adams, 2017; Heiss et al., 2016). Using different strategies, both of these papers estimate that as-if switching costs are in the thousands of dollars, which is consistent with the estimates from pension plans we discuss below, but that once inattention is accounted for, the estimated as-if switching costs are only in the hundreds of dollars. Blumenstock, Callen and Ghani (2017) conducts an experiment to study the mechanisms underlying default effects in a savings context; they find the cognitive costs of selecting a savings plan play an important role. These results can inform the normative judgments policy-makers must make but they do not resolve normative ambiguity. For instance, if one accepts that as-if costs are primarily driven by a reluctance to pay attention to non-default options, the planner must still determine the degree to which incurring such costs would reduce actually welfare.[6] Understanding the role of normative judgments in behavioral welfare analysis is therefore complementary to understanding the mechanisms by which default effects operate.

For clarity of exposition, we focus first on the question of whether as-if costs are normative, and then incorporate the possibility that active decision-makers might choose suboptimally. The remainder of the paper is therefore organized as follows: Section 1 sets out the model with normative ambiguity over as-if costs. Section 2 examines optimal policy-making in this model, with a particular focus on policies promoting active choice and defaults that minimize the number of opt-outs. Section 3 incorporates the possibility that active decision-makers make mistakes. Section 4 illustrates our results using data on 401(k) plan contribution defaults. Section 5 concludes.

1 Model

Consider a population of measure 1. Decision-makers choose from a fixed menu X, where $x_i \in X$ denotes the option chosen by individual i. One option, $d \in X$, is presented to decision-makers as the default. Decision-makers have well-behaved preferences over the elements of X, represented by utility function $u_i(\cdot)$.[7] We assume that $u_i(\cdot)$ is cardinal and comparable across individuals. Preferences over X do not depend on the default.

[6] For helpful discussions of issues relating to attention, information frictions, and policymaking, see Handel (2013) and Handel and Schwartzstein (2018).

[7] Note that the differentiability of $u_i(x)$ may fail in some relevant applications. For example, with 401(k) plans with an employer match, $u(x)$ might exhibit an interior kink point at the contribution rate at which the match kicks in or at which the match is maximized. We discuss this further below.

Individual behavior is characterized by the following optimization problem:

$$x_i(d) = \arg\max_{x \in X} \; u_i(x) - \gamma_i \, 1_{\{x \neq d\}} \tag{1}$$

where $\gamma_i \geq 0$ for all i.[8] We will refer to γ_i as the *as-if cost* to selecting an option that is not the default. Let $x_i^* = \arg\max_{x \in X} u_i(x)$ denote the choice that maximizes (1) when $\gamma_i = 0$. We assume that decision-makers indifferent between selecting the default and opting out will select the default. Under these assumptions, behavior is given by

$$x_i(d) = \begin{cases} x_i^* & u_i(x_i^*) - u_i(d) > \gamma_i \\ d & u_i(x_i^*) - u_i(d) \leq \gamma_i \end{cases} \tag{2}$$

We next define $a_i(d) = u_i(x_i^*) - u_i(d) - \gamma_i$, which reflects the degree to which an individual prefers opting out. We will refer to a decision-maker with $a_i(d) > 0$ as *active* at default d and a decision-maker with $a_i(d) \leq 0$ as *passive* at default d.[9] We denote the cumulative distribution of $a_i(d)$ over the population of decision-makers at a given default by $F_{a;d}$.

Our main results will apply to the class of models generating behavior that can be represented by (2). Masatlioglu and Ok (2005) derives necessary and sufficient restrictions on behavior that a model with this representation must satisfy.[10] For example, (2) requires that if a decision-maker would choose x over y when y is the default, she must also choose x over y when x is the default.

Equation (2) described individual behavior. The following equation characterizes individual welfare:

$$w_i(x, d) = u_i(x) - \pi_i \, \gamma_i \, 1_{\{x \neq d\}} \tag{3}$$

where $\pi_i \in [0, 1]$ reflects the degree to which the as-if costs are normative and thus affect the decision-maker's welfare.[11] One can think of the maximand in (1) as "decision utility" and the utility function in (3) as "experienced utility" (Kahneman, Wakker and Sarin, 1997). When $\pi_i = 1$, a decision-maker's sensitivity to the default is rational. When $\pi_i = 0$, default sensitivity represents a complete mistake; the decision-maker behaves as if selecting a non-default option would reduce his welfare, but if he were to actually select a non-

[8] For simplicity, we focus on the case in which γ_i is fixed for each individual. We show in the Online Appendix that the intuition of our main results extends to the case in which γ_i depends on which option is the default.

[9] Note that a decision-maker who "actively" considers each option in the choice set before settling on the option that happens to be the default would still be referred to as "passive" in our terminology. In addition, below we will assume that such a decision-maker's welfare is the same as a decision-maker who selects the default without considering other alternatives. This assumption is innocuous for purposes of deriving the optimal default under our model, since a decision-maker who considers each option even when her most-preferred option is the default would also consider each option under alternative defaults.

[10] This representation is slightly less general than the one implied by the Masatlioglu and Ok (2005) axioms. Our main results extend to the more general representation as well (see the Online Appendix).

[11] One might extend our approach to settings in which $\pi_i > 1$, which may occur, for example, when opt-out costs are not fully salient or when individuals overestimate the gains from optimizing. This possibility tends to push the optimal default toward options that induce relatively few opt-outs.

default option, his welfare would not decrease. When $\pi_i \in (0,1)$, it would be rational for the decision-maker to exhibit some sensitivity to the default, but his behavior implies that the welfare reduction from opting out is greater than it actually is.[12] Note that (3) embeds the assumption that active choosers make optimal decisions over the (non-default) options they select, as active choosers maximize the sub-utility function $u_i(x)$. We relax this assumption in Section 3.

We denote a decision-maker's indirect utility by $v_i(d) \equiv w_i(x_i(d),\, d)$. Aggregate social welfare under default d is given by $W(d) \equiv \int_i v_i(d)\, di$. An *optimal default* $d^* \in X$ is an option that yields the highest social welfare when presented as the default, $W(d^*) \geq W(d)\ \forall d \in X$.

1.1 Relationship to Positive Models of Default Effects

In this section we briefly review alternative behavioral models that have been proposed to explain default effects and discuss the extent to which they do or do not map into our framework. The main insight is that although many behavioral models are consistent with our representation, each implies a different conclusion regarding the share of the as-if costs that are normative (π). Combinations of these models would generate an even wider range of possibilities for π.

Real Opt-Out Costs. Decision-makers select from among the available options according to their preferences over the available items (u_i), while rationally accounting for the welfare-relevant costs associated with selecting an option that is not the default. These costs might include monetary costs, such as administrative fees for selecting a non-default option, or non-monetary costs such as the hassle or mental effort required to determine one's most-preferred option from the available menu. Because decision and experienced utility are identical in this model, $\pi_i = 1$.

Status Quo Bias. Another proposed explanation for default effects is that decision-makers follow a psychological heuristic in which they behave as if it is costly to deviate from the status quo, and simultaneously perceive the default option to represent a continuation of the status quo. As suggested by the word "bias", this propensity to follow the status quo does not actually increase welfare. Hence, $\pi_i = 0$.

Endowment Effect. A related possibility is that decision-makers perceive themselves as endowed with the default option and exhibit reluctance to exchange that endowment for other options (Tversky and

[12] Close readers of BFP may wonder about the difference between the role of π in our model and the role of "frame-dependent weights" in theirs. The idea behind frame-dependent weights is that within certain positive models, the extent to which a decision-maker accounts for the normative opt-out costs will vary based on the choice environment (i.e., the "frame"). For example, in a present-bias model, observed choices would reveal larger as-if opt-out costs if the opt-out decision was made during the same time period in which the opt-out costs were potentially incurred, rather than during a prior period. Thus, frame-dependent weights account for ambiguity about the proper perspective on welfare *within* a given positive model. In contrast, we use π to reflect different perspectives on welfare either within or *across* alternative behavioral models. More importantly, in their empirical application, BFP assume a particular value for their frame-dependent weights (roughly equivalent to $\pi = 0.01$) that strikes them as ex ante reasonable. In contrast, we remain agnostic about the share of as-if costs that are normative to highlight how assumptions of this type drive welfare conclusions.

Kahneman, 1991). Whether this reluctance enters into welfare is controversial (Zeiler, 2017). When the endowment effect is fully normative, $\pi_i = 1$; when it is entirely a bias, $\pi_i = 0$. Intermediate cases would imply $\pi_i \in (0, 1)$.

Quasi-Hyperbolic Discounting. Suppose a decision-maker decides whether to opt-out of a default in period one. In all future periods, she receives flow utility from the option she selected in the prior period and decides again whether to opt out. Assuming that opt-out costs and flow utility functions are fixed over time, the individual faces the same decision problem and will make the same choice in each period; we can think of $u_i(x)$ as utility for some option x received in perpetuity.[13] As in Laibson (1997), $\delta_i \in (0, 1]$ denotes the discount factor and $\beta_i \in (0, 1)$ denotes the degree of present-bias. The contemporaneous cost of opting out is denoted by c_i.

Suppose the decision-maker correctly anticipates her future opt-out decisions. In this case, choices are described by: $x_i(d) = \arg\max_{x \in X} \ \delta_i \beta_i u_i(x) - c_i 1_{\{x \neq d\}}$. Welfare is described by: $w(x_i, d) = \delta_i \, u_i(x) - c_i 1_{\{x \neq d\}}$. It is straightforward to verify that these preferences are equivalent to (1) and (3), with $\gamma_i = \frac{c_i}{\delta_i \beta_i}$ and $\pi_i = \beta_i$.[14]

Inattention. Decision-makers may intentionally or inadvertently fail to consider the utility of the available options or the (real or perceived) costs of opting out of the default (Chetty, 2012; Goldin and Lawson, 2016). Following Masatlioglu, Nakajima and Ozbay (2012) we model inattention by supposing that decision-makers maximize utility over some subset of the available options, $\Gamma_i(X, d) \subseteq X$, where Γ_i represents an *attention filter:* $x_i(d) = \arg\max_{\Gamma_i(X,d)} \ u_i(x)$. Suppose every decision-maker either pays attention only to the default (passive choice) or to the full menu (active choice) $\forall i, \ \Gamma_i(X, \, d) \in \{\{d\}, \, X\}$.

There are two intuitive possibilities for how Γ_i is determined. One is a heuristic model of attention, in which there are simply two exogenous types of agents: attentive choosers for whom $\Gamma_i(X, d) = X$ always and inattentive choosers for whom $\Gamma_i(X, d) = d$ always (Chetty et al., 2014). This model maps into our framework with $\gamma_i \in \{0, \infty\}$ and $\pi_i = 0$. Alternatively, the set of options to which a decision-maker is attentive may depend on the utility gain from choosing actively. Decision-makers may choose to be attentive ($\Gamma_i = X$) when the gains to doing so exceed some threshold, and otherwise set $\Gamma_i = \{d\}$. This model is equivalent to the general model of default sensitivity laid out above; ambiguity over the welfare consequences of following the default corresponds to whether the costs of paying attention are themselves normative.

[13] For analysis of the costly opt-out model in more general dynamic settings, see Carroll et al. (2009) and Choukhmane (2019).

[14] When decision-makers choose not to opt out today but expect to choose according to their long-run preferences (i.e., $\beta = 1$) in the future, preferences and behavior can be represented in terms of Equations (2) and (3), with $\gamma_i = \frac{1-\beta_i\kappa_i}{\beta_i - \beta_i\kappa_i} \frac{c_i}{\delta_i}$ and $\pi_i = \frac{\beta_i - \beta_i\kappa}{1-\beta_i\kappa}$, where $\kappa \in [0, 1]$ denotes the probability the decision-maker assigns to selecting according to her long-run preferences during the subsequent period. This result follows from incorporating π into BFP's analysis of partial naivete. When $\kappa_i = 1$ (full naivete), the decision-maker will procrastinate indefinitely and never opt out. In this case an observer would conclude that as-if costs were arbitrarily large and, though such costs would never be incurred, they would be totally irrelevant for welfare, $\pi_i = 0$.

Anchoring Effects. Defaults may shape behavior through psychological anchoring effects, in which the default induces decision-makers to select an option closer to the default than they would otherwise choose (Tversky and Kahneman, 1974). Behavior under such models fall cannot be represented according to (2) because the default affects choices among those who opt out.[15] Although defaults may sometimes operate through anchoring effects, the empirical evidence reviewed in Section 1.2 suggests that there are many contexts in which the opt-out cost models appear to better fit the data.[16] Online Appendix Section C extends a number of our key findings using a model that can incorporate framing effects.

Advice. Decision-makers might select the default if they are uncertain over their preferences and they believe the planner's choice of default provides an informative signal as to which option is best for them. The optimal policy prescriptions we consider are geared towards a world in which the planner lacks ex ante information as to which option is most consistent with decision-makers' preferences, suggesting that rational (well-informed) decision-makers would not treat the default signal as having any informational content. Nonetheless, decision-makers might mistakenly construe the default as a suggestion by the planner and treat it as containing some informational content. One possibility is that decision-makers treat the suggestion as "take it or leave it" advice – i.e., they either follow the suggestion exactly or ignore it altogether, perhaps by gathering so much information on their own that the original suggestion has negligible signal value. Such a model is isomorphic to the status quo bias model when the default has no true signal value. Alternatively, decision-makers may take the suggested option into account, even if they do not accept it, and choose something closer to the default than what they otherwise would have chosen. In this case, the default affects decision-making like an anchor, where the effect of the default on a decision-maker's behavior depends on the strength of the decision-maker's prior and the weight the decision-maker attaches to the default-as-signal.

1.2 Empirical Plausibility

In practice, it is often difficult to directly test the axiomatic foundations of particular behavioral models. With respect to models of default effects, for example, difficulties may arise because individuals have heterogeneous preferences and opt-out costs, or it may be impossible to observe the same individual choosing under alternative defaults.

One prediction of our model that, with modest additional structure, does lend itself to testing is the idea that fewer individuals will select any given option when the default is close to that option than when the default is far from that option. Formally, this prediction can be stated as follows: *Suppose that the*

[15] Technically, models of anchoring violate the axiom that Masatlioglu and Ok (2005) label Status Quo Independence.

[16] Bernheim, Fradkin and Popov (2015) estimate a model incorporating anchoring and a fixed cost of opting out of the default, and note that the fit of the model improves somewhat when allowing for anchoring. However, at least some of this improvement in fit is mechanical, since the anchoring effect represents a new free parameter that can help explain why individuals choose the default.

menu X is ordered, and $u_i(\cdot)$ is single-peaked. Then for any two defaults $d^{'}$ and $d \in X$ such that $d^{'} > d$, $P(x_i(d) = x) \geq P(x_i(d^{'}) = x)$ for $x > d^{'}$, and $P(x_i(d) = x) \leq P(x_i(d^{'}) = x)$ for $x < d$.

Evidence consistent with this prediction has been documented across a range of settings, including: 401(k) contributions (e.g., Madrian and Shea 2001, Figure IIc; Choi et al., 2006, Figure 2), charitable contributions (Altmann et al., 2016); taxi ride tips (Haggag and Paci, 2014); and even thermostat temperature settings in office buildings (Brown et al., 2013). These findings support the empirical relevance of the class of behavioral models we study. Notably, the anchoring model of defaults discussed in Section 1.1 makes the opposite prediction, suggesting for example that we should observe $P(x_i(d) = x) < P(x_i(d') = x)$ for $x > d' > d$, at least at values of x that are sufficiently close to d'.

2 Characterizing the Optimal Default

In this section we characterize the optimal default in terms of the components of our model.

2.1 Generic Welfare Comparisons

We begin by examining how changes in the default affect welfare. The welfare achieved under any default can be decomposed between two groups: (1) active choosers selecting x_i^* and incurring normative costs $\pi_i \gamma_i$, and (2) passive choosers selecting d:

$$W(d) = E[u_i(x_i^*) - \pi_i \gamma_i \,|\, a_i(d) > 0] \; (1 - F_{a;d}(0)) + E[u_i(d) \,|\, a_i(d) \leq 0] \, F_{a;d}(0), \tag{4}$$

Consider a change in the default from d_0 to d_1. From (4), it is apparent that this change affects welfare directly for passive choosers, for whom it changes the option they select, and may also affect the composition of active and passive decision-makers. To study the welfare effects of this change, it will be useful to partition the population into four groups of decision-makers based on their behavior under the old default (d_0) and the new default (d_1):

Group	**Behavior when default is:**		**Characterization**
	d_0	d_1	
Always Active (AA)	$a_i(d_0) > 0$	$a_i(d_1) > 0$	$u_i(x_i^*) - \max\{u_i(d_0),\ u_i(d_1)\} > \gamma_i$
Always Passive (PP)	$a_i(d_0) \leq 0$	$a_i(d_1) \leq 0$	$u_i(x_i^*) - \min\{u_i(d_0),\ u_i(d_1)\} \leq \gamma_i$
Active-to-Passive (AP)	$a_i(d_0) > 0$	$a_i(d_1) \leq 0$	$u_i(x_i^*) - u_i(d_0) > \gamma_i \geq u_i(x_i^*) - u_i(d_1)$
Passive-to-Active (PA)	$a_i(d_0) \leq 0$	$a_i(d_1) > 0$	$u_i(x_i^*) - u_i(d_1) > \gamma_i \geq u_i(x_i^*) - u_i(d_0)$

The table describes how the composition of these four groups is determined in terms of the behavioral parameters from Equation (2). We denote the fraction of the population in each of these groups by $p(j)$ for $j \in \{AA,\ PP,\ PA,\ AP\}$. The following proposition uses this decomposition to characterize the welfare effect of a change in default.

Proposition 1. *For any two defaults d_0, $d_1 \in X$:*

$$\begin{aligned} W(d_1) - W(d_0) = E\left[u_i(x^*) - u_i(d_0) - \pi_i\gamma_i \,|\, PA\right]\, p(PA) - E\left[u_i(x^*) - u_i(d_1) - \pi_i\gamma_i \,|\, AP\right]\, p(AP) \\ +E\left[u_i(d_1) - u_i(d_0) \,|\, PP\right]\, p(PP) \end{aligned} \tag{5}$$

Several features of (5) are worth noting. First, the always-active choosers, group AA, do not enter into (5); these individuals incur the same normative cost ($\pi_i\gamma_i$) and make the same choice (x_i^*) under both defaults. Second, for those who are passive at d_0 and active at d_1 (group PA), the change induces a utility gain from choosing actively, $u_i(x_i^*) - u_i(d_0)$, but also causes them to incur normative cost $\pi_i\gamma_i$. The first term in equation (5) reflects the change in social welfare from these individuals. The second term is the analogous contribution from individuals who are active at d_0 but not at d_1 (group PA). The third term reflects individuals who are passive under both defaults (group PP); the net welfare effect for this group depends on whether they (on average) prefer the new default or the original default.

One instructive special case concerns the situation in which all individuals prefer the same option, $x_i^* = x^*$ for all i. Not surprisingly, the optimal policy is such cases is to set the default equal to decision-makers' most-preferred option, regardless of the $\pi_i' s$. Intuitively, complete preference homogeneity eliminates normative ambiguity because it avoids the need to compare the welfare of active choosers with the welfare of passive choosers; this is because no one incurs (potentially normative) opt-out costs.

Note that Proposition 1 holds regardless of the nature of the menu X – it might be discrete, continuous, or of multiple dimensionality. The next result considers situations where X is a real interval, which occurs in many applied contexts.

Proposition 2. *Let X be any interval in $\mathbb{R}$, and suppose $u_i(x)$ is everywhere differentiable for all i. If d^* represents an interior solution to the optimal default problem, the following first-order condition is satisfied:*

$$\begin{aligned} 0 = W^{'}(d^*) \quad &= \quad E[(1-\pi_i)\gamma_i \,|\, a_i(d^*) = 0,\ u_i'(d^*) < 0]\, f_{a|u'<0}(0)\, F_{u'}(0) \\ &- \quad E[(1-\pi_i)\gamma_i \,|\, a_i(d^*) = 0,\ u_i'(d^*) > 0]\, f_{a|u'>0}(0)\ (1 - F_{u'}(0)) \\ &+ \quad E\left[u'(d^*) \,|\, a_i(d^*) < 0\right]\, F_{a;d^*}(0) \end{aligned} \tag{6}$$

where $f_{a|u'>0}$ is the probability density function of $a_i(d^*)$ conditional on $u_i'(d^*) > 0$; $F_{u'}$ is the cumulative density function of $u_i'(d^*)$; and $F_{a;d^*}$ is the cumulative density function of $a_i(d^*)$.

As in Proposition 1, the three terms represent the welfare effects of the default change on decision-makers in the AP, PA, and PP groups. The first term represents the PA group; a decision-maker for whom $a_i(d) = 0$ and $u_i^{'}(d) < 0$ will be passive at the original default and active following a marginal increase in the default (which they prefer slightly less than the original default). Similarly, the second term represents decision-makers in the AP group, who are slightly better off after the marginal increase in the default, and therefore more willing to acquiesce to it. Decision-makers in the third, inframarginal group, with $a_i(d) < 0$, remain passive even after a small change in the desirability of the default.

How does the normative share of as-if costs affect the optimal default? Proposition 2 highlights that π matters for weighting the relative welfare effects of a change in the default for decision-makers in the PA and AP groups against the welfare effects for decision-makers in the PP group. When $\pi_i = 1$, the welfare effects depend only on decision-makers in the PP group, who experience a marginal change in welfare from moving to a slightly better or slightly worse default. The reason why is that decision-makers in the PA and AP groups behave as though they are indifferent between following the default and making an active choice ($a_i(d) = 0$). When $\pi_i = 1$ for decision-makers in these groups, that behavior fully reflects their welfare, and the envelope theorem implies that their welfare is not affected by a policy change that makes them active or passive. In contrast, when $\pi_i < 1$, the welfare of the PA and AP groups will be weighted more heavily in determining the optimal default, because although these groups are small because they are on the margin of choosing actively, their welfare changes discretely when they begin or cease to choose actively. The next two sections further explore how the optimal policy depends on π_i.

2.2 Forcing Active Choice

This Section applies our framework to evaluate the desirability of policies that induce decision-makers to make active choices. In practice, such policies might take the form of (1) a "penalty default" (Ayres and Gertner, 1989) set to an option so undesirable that virtually all decision-makers opt out, or (2) restricting the opportunity set so that decision-makers are forced to make an active choice (e.g., Carroll et al., 2009). As an example of the former approach, one could imagine setting intestacy law – law governing inheritances in the absence of a will – so that individuals who die without leaving a will would have all of their assets taxed at a 100% rate. An example of the latter approach would be requiring new employees to make an active decision about how much to contribute to their 401(k) plans as a condition of employment.[17] For convenience, we will model both types of policies as penalty defaults, since their effects are the same under our assumptions.

[17] Other ways to promote active choice are to reduce the costs of opting out of a default, considered by Chesterley (2017), or taxing decision-makers who select the default option, considered by BFP.

Formally, we define a *penalty default* as some option $d_p \in X$ for which $a_i(d) > 0$ for all i. It is straightforward to show that whenever $u_i(d_p)$ is sufficiently low for all individuals, d_p will be a penalty default. Compare a change in the default to a penalty default d_p from an arbitrary alternative d. Using Proposition 1, we have

$$W(d_p) - W(d) = E[u_i(x^*) - u_i(d) - \pi_i\gamma_i|PA]\,p(PA) \tag{7}$$

Because individuals are never passive at d_p, only the first term of (5) matters for welfare. The following proposition stems from (7) and illustrates the importance of resolving normative ambiguity when policies that promote active choice are available:

Proposition 3. *Suppose that X is any menu and there exists a penalty default $d_p \in X$. There exist thresholds $\underline{\pi} \in [0, 1)$ and $\overline{\pi} \in (0, 1]$ such that*

(3.1) *$\pi_i \leq \underline{\pi}$ for all i implies d_p maximizes social welfare over all $d \in X$.*

(3.2) *$\pi_i \geq \overline{\pi}$ for all i implies d_p* minimizes *social welfare over all $d \in X$.*

Proposition 3 shows that when forcing active choice is a feasible policy, it is never possible to identify the optimal default without taking a stance on whether or to what degree opt-out costs are normative.[18] Moreover, the stakes are high: forcing active choices can be either the best or the worst possible outcome for social welfare, depending on what π turns out to be.

To interpret (3.1), start from the benchmark case where $\pi_i = 0$ for everyone. In that case, forcing active choice results in everyone receiving the option they prefer and no one incurring any normative opt-out costs. The result in (3.1) generalizes this idea to the case where π is small but not necessarily zero. In contrast, (3.2) implies that forcing active choice is extremely undesirable under high values of π. Note that when π_i is sufficiently close to 1 for all i, the right-hand side of (7) must be negative, because individuals who are passive at default d have $u_i(x^*) - u_i(d) < \gamma_i$. Such individuals reveal a preference for choosing passively. Hence, when π_i is sufficiently close to 1 for all i, forcing active choice is dominated by *every other potential default.*

2.3 Minimizing Opt-Outs

A frequently discussed rule of thumb for setting defaults, proposed by Thaler and Sunstein (2003), is to select as the default whichever option minimizes the number of decision-makers who opt-out. In our notation, the opt-out minimizing default, d^m, is defined as the value of d that maximizes: $W^m(d) \equiv F_{a;d}(0)$, where,

[18] The case in which x_i^* is homogeneous is a knife's edge exception to this statement. In that case, setting $d = x^*$ achieves the highest possible social welfare for any value of π – including tying d_p when $\pi_i = 0\ \forall i$.

as above, $F_{a;d}(\cdot)$ is the cumulative density function of $a_i(d)$, i.e. the fraction of the population choosing passively under default d.

Evaluating this expression at two possible defaults, d_0 and d_1, it is straightforward to derive that under W^m, social welfare is improved by changing the default from d_0 to d_1 if and only if $p(PA) < p(AP)$. That is, the default change must cause more decision-makers to become passive than it causes to become active. To illustrate how this condition relates to welfare in our model, note that we may decompose (5) as:

$$\begin{aligned} W(d_1) - W(d_0) = \quad & \underbrace{(p(AP) - p(PA))\,\overline{\pi\gamma}}_{1} + \underbrace{p(AP)\,E\left[\pi_i\gamma_i - \overline{\pi\gamma}\,|\,AP\right] - p(PA)\,E\left[\pi_i\gamma_i - \overline{\pi\gamma}\,|\,PA\right]}_{2} \\ & + \underbrace{E[u_i(x^*) - u_i(d_0)|PA]\,p(PA) - E[u_i(x^*) - u_i(d_1)|AP]\,p(AP)}_{3} \\ & + \underbrace{E[u_i(d_1) - u_i(d_0)|PP]\,p(PP)}_{4} \end{aligned} \tag{8}$$

where $\overline{\pi\gamma} = E[\pi_i\gamma_i]$. As an initial matter, note that term 1 compares $p(AP)$ and $p(PA)$ exactly as in W^m. Individuals who are active at d_1 but not d_0 (group PA) will incur opt-out costs under d_1 valued at $\pi\gamma$, which has a negative effect on their welfare. The opposite is true for the AP group, who incur costs under d_0 but not d_1. Term 1 therefore favors whichever default minimizes opt-outs. When all of the other terms in 8 are negligible or have the same sign as the first term, the opt-out minimizing default coincides with the optimal default.

The other terms in 8 represent factors that may cause the optimal default to diverge from the default that minimizes opt-outs. Term 2 reflects the fact that even when the size of the AP and PA groups are similar, the magnitude of the normative opt-out costs of each may differ. Similarly, term 3 reflects that, aside from whatever cost they incur from being active, the utility gain from being active may differ across the AP and PA groups. Finally, the fourth term captures how the change in the default affects welfare for the decision-makers who remain passive. Notably, the preferences of this group are completely neglected by the minimizing opt-outs rule, even though the choices of this group are directly affected by a change in the default. When the preferences of group PP differ systematically from those of the PA and AP groups, the default selected by W^m may be sub-optimal because it fails to reflect the preferences of the decision-makers who remain passive under both defaults. Hence, when the PP group is large and tends to prefer defaults that induce many decision-makers to opt-out, the minimizing opt-outs rule of thumb may perform poorly.

The following proposition provides sufficient conditions under which minimizing opt-outs yields the optimal default:

Proposition 4. *Suppose that* $X = [x_{min}, x_{max}] \subseteq \mathbb{R}$ *and that:*

(A4.1) *As-if costs γ_i are distributed independently of x_i^*.*

(A4.2) *Preferences are given by $u_i(x) = u(x - x_i^*)$ for some map $u : \mathbb{R} \to \mathbb{R}$, with $u'(0) = 0$, $u'' < 0$ and $u(c) = u(-c)$ for any c.*

(A4.3) *x_i^* follows a single-peaked and symmetric distribution about some mode x^m.*

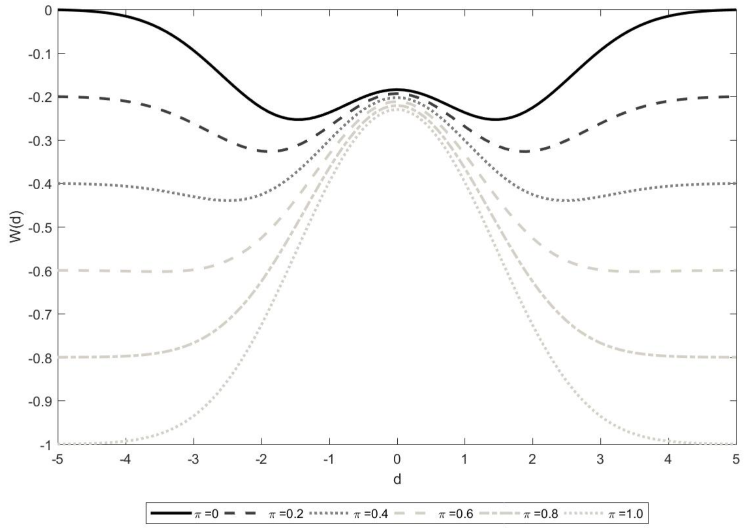
Under these conditions, there exists a threshold $\overline{\pi} \in (0, 1]$ such that $\pi_i \geq \overline{\pi}$ for all i implies that the optimal default is the default that minimizes opt-outs.

Proposition 4 provides conditions under which minimizing opt-outs yields the optimal policy. Loosely speaking, these conditions occur when as-if costs are sufficiently normative, the distributions of the underlying behavioral parameters are independent, and decision-makers' preferences are symmetric and single-peaked. We can understand the sufficient conditions in terms of their implications for the various terms in Equation (8). In particular, (A4.1) rules out a relationship between as-if costs γ_i and preferences that could cause the sign of term 2 in Equation (8) to have the opposite sign of term 1. Next, (A4.2) makes the comparison of the utility differences in the last two terms of Equation (8) straightforward, as all heterogeneity in $u_i(\cdot)$ derives from heterogeneity in the distribution of optimal choices x_i^*. Third, (A4.3) rules out features of the distribution of x_i^* that could pull the optimal default away from the opt-out-minimizing default via the third and fourth terms in Equation (8). Together, (A4.2) and (A4.3) guarantee that the effect of a change in the default in opt-outs among the PA and AP groups in (8) is a strong signal about the change in welfare of the PP group. The symmetry assumptions in (A4.2) and (A4.3) ensure that, when $d = x^m$, (1) the effect of a marginal change in the default on the AP and PA groups cancel each other out, and (2) for every member of the PP group made better off by a marginal change in the default, there is another member of that group made worse off by the same amount. These cancellations ensure that the first-order condition for x^m to be a local optimum is satisfied. Given these assumptions, all that remains is to examine when $d = d^m$ is indeed the global optimum, which is guaranteed for sufficiently high π.[19]

Together, Propositions 3 and 4 imply that, under the regularity conditions in (A4.1) - (A4.3), the optimal policy rule is relatively simple. When π is sufficiently large, the social planner should minimize opt-outs. When π is sufficiently small, the social planner should force active choice. For intermediate values of π, other policies may be optimal. When the conditions in (A4.1)-(A4.3) are not satisfied, minimizing opt-outs may not be optimal for large π, but one can use the expression in Equation 8 to correct for asymmetries in $u(\cdot)$ or the distribution of x^*, or a correlation between as-if costs γ_i and optimal choices x_i^*.

[19] BFP show that a distinct factor that tends to favor minimizing opt-outs when as if costs are normative is the presence of agglomerations in the distribution of x^*, such as might occur in a 401(k) savings context with an employer match (see BFP Theorem 2).

Figure 1: Social Welfare with Quadratic Preferences and Gaussian x_i^*



Note: This figure plots social welfare, $W(d)$, for a simulated model in which $u(x-x_i^*) = -\alpha(x-x_i)^2$ with $\alpha = 0.25$, $x_i^* \sim N(0,1)$, and $\gamma_i = 1$ for all individuals. We assume π_i is homogeneous across decision-makers, and we plot $W(d)$ for several values of π. Setting $d = E[x_i^*] = 0$ will minimize opt-outs in this model, and setting an extreme default will force active choices. The simulation illustrates that the optimal policy follows a threshold rule over π in this model, so that minimizing opt-outs is optimal for high values of π and forcing active choice is optimal for low values of π.

Figure (1) plots social welfare for a stylized model that satisfies (A4.1)-(A4.3). To fill out the model, we assume that π_i is uniform across decision-makers, x_i^* follows a Gaussian distribution in the population, and $u(x - x_i^*)$ is quadratic – i.e., $u(x - x_i^*) = -\alpha(x - x_i^*)^2$ for suitably chosen $\alpha > 0$. To interpret the figure, recall that forcing active choice is equivalent in the model to selecting a default sufficiently extreme that all decision-makers choose to opt out, and, because we plot $W(d)$ assuming that $x^m = 0$, setting a default of zero will minimize opt-outs. The figure shows that as π varies, the default that minimizes opt-outs remains a local optimum; the feature that varies with π is the relative attractiveness of forcing active choice. As suggested by the figure, the optimal policy in this stylized setting takes the form of a threshold rule around some threshold $\overline{\pi} \approx 0.2$. When $\pi > \overline{\pi}$, setting the default to minimize opt-outs is optimal. Instead, when $\pi < \overline{\pi}$, the optimal policy is to force active choice. Our results in this section also shed additional light on previous results from the literature. Specifically, Carroll et al. (2009) consider the optimal policy within a model of default effects similar to the one we describe in Section 1.1, where present bias magnifies opt-out costs relative to true opt-out costs and the individual is a sophisticated quasi-hyperbolic discounter according to some (homogeneous) factor β. The authors add some additional structure to the model, namely a uniform distribution of costs over some finite interval (c.f. our Assumption A3.1), quadratic loss preferences (c.f. our assumption A3.2), and a uniform density of optimal choices x_i^* over some finite interval (c.f. our assumption A3.3). Within this model, they show that active choices are optimal when (1) β is sufficiently low and (2) optimal choices (x_i^*) are sufficiently heterogeneous. Conversely, when (1) β is sufficiently high and (2) preferences are less heterogeneous, the optimal choice will tend to be a "center default" that minimizes opt-outs (c.f. x^m in Proposition 4). Recall from Section 1.1. that in the sophisticated present-bias model of default effects , $\pi = \beta$. Our Proposition 3 therefore illuminates the generic reason why active choices are optimal when β is low: these are the cases in which as-if costs are deemed normatively irrelevant. Similarly, our Proposition 4 illuminates the generic reason why minimizing opt-outs is optimal when $\beta \approx 1$.

2.4 Measuring the Normative Share of Opt-Out Costs

This section has described how optimal policy turns on the share of opt-out costs that are welfare-relevant. Because this parameter typically cannot be directly inferred from data, our proposed approach is to derive optimal policy as a function of the normative share of opt-out costs; doing so allows policymakers and advocates to assess the assumptions on which alternative policies lie. Still, depending on the setting, various data can be informative on the share of opt-out costs that are normative. For example, one might identify how opt-out costs affect a decision-maker's welfare by extrapolating from the decision-maker's choices in other settings, such as how much the decision-maker is willing to pay to avoid undertaking tasks that are similar to opting out of a default like filling out other paperwork or making active decisions with similar complexity. Alternatively, one could try to learn how opting out affects welfare more directly, such as by estimating the effect of opting out of a default on decision-makers' subjective well-being. Of course, both of these approaches have weaknesses: other choices made by the decision-maker might differ in important ways or might themselves be biased. How opt-outs affect subjective well-being is difficult to measure and may miss other welfare-relevant features of the choice. But in some settings, these approaches may provide clues as to the range of plausible values for π.

3 Mistaken Active Choices

Thus far we have restricted our focus to optimization frictions that arise because a default is present. That is, although our model has allowed decision-makers to err in deciding whether to opt-out of a default, we have assumed that those decision-makers who do opt out go on to choose optimally from the available options. However, choices may also be distorted by biases unrelated to default effects. In the retirement savings decision, for example, present-biased employees may under-save even when making an active choice.

3.1 Modeling Internalities

To incorporate this possibility into our model, we continue to assume that behavior is described by Equation (1) but now assume that welfare is given by

$$w_i(x) = u_i(x) + m_i(x) + \pi_i \gamma_i 1_{\{x \neq d\}}, \tag{9}$$

where $m_i(x)$ is the internality imposed on the individual by his or her choice of x – i.e., the component of the welfare effect of x that is not taken into account by the decision-maker. Here, $a_i(d) > 0 \implies x_i(d) = x_i^a$. The active choice, x_i^a, maximizes $u_i(x)$ but not $u_i(x) + m_i(x)$ due to the internality.

For simplicity, we focus on the case in which X is a real interval and assume that both $u_i(x)$ and $m_i(x)$ are differentiable. In the retirement savings example, for an active chooser who under-saves, we would have $u_i'(x_i^a) = 0$ and $m_i'(x_i^a) > 0$. Indirect utility and social welfare are defined as above. We will describe how the optimal policy changes with the addition of internalities, first in general and then under some restrictions that provide additional intuition. Let $W_0^{'}(d)$ be the effect of a marginal change in the default on welfare in our original model, as derived in Proposition 2. With internalities, the analogue to this expression is given by:

$$\begin{aligned} W^{'}(d) = \quad & W_0^{'}(d) + E\left[m_i(x_i^a) - m_i(d) \,|\, PA\right] \ P(PA) \\ & -E\left[m_i(x_i^a) - m_i(d) \,|\, AP\right] \ P(AP) + E\left[m_i'(d) \,|\, PP\right] \ P(PP). \end{aligned} \tag{10}$$

There are two changes in this expression relative to the one in Proposition 2. First, the PA and AP groups experience a discrete change in the internality from becoming active or becoming passive as the default changes. Second, the always-passive (PP) group in the last term of Equations (10) experience an additional marginal welfare effect, m_i', from the change in the default. As before, the welfare of always-active choosers does not enter into the evaluation of the welfare effect of a change in the default.

In order to compare optimal policy with and without internalities, it is instructive to place additional simplifying restrictions on $m_i(x)$. The following proposition illustrates how the presence of internalities affects the determination of the optimal default derived above:

Proposition 5. *In the model with internalities, suppose that*

(A5.1) *For all i, $u_i(x) = -\frac{\alpha}{2}(x - x_i^a)^2$ with $\alpha > 0$.*

(A5.2) *Normative preferences are given similarly by $u_i(x) + m_i(x) = -\frac{\alpha}{2}(x - x_i^*)^2$.*

(A5.3) *The error in active choice $x_i^a - x_i^*$ is independent of x_i^a and γ_i.*

Then the social welfare effect of a marginal change in the default is $W'(d) = W_0^{'}(d) + \mu\, X^{'}(d)$, where $W_0(d)$ denotes social welfare without internalities (Equation 6), $\mu = E[m_i'(x_i(d))]$, and $X(d) = E[x_i(d)]$.

Proposition 5 highlights that optimal policy considerations here balance the concerns of the previous model, summarized by $W_0^{'}(d)$, with a new goal, which is to correct the internality generated by the decisions of the active choosers. For example, if $\mu > 0$ represents the average degree of under-saving among a population of decision-makers, the optimal savings contribution default would induce more saving than when the social planner assumed no internalities were present. The larger the mean marginal internality, μ, the further the deviation from the no-internality optimum. In addition, it is straightforward to show that $\frac{\partial E[x_i(d)]}{\partial d} = E\left[x_i^a - d|PA\right]\ P(PA) + E\left[d - x_i^a|AP\right]\ P(AP) + P(PP)$. In words, the effect of a change in the default on total activity (X) has two components. First, both of the marginally active groups choose discretely lower x_i – recall that $x_i^a < d$ for the PA group, and $x_i^a > d$ for the AP group. Second, the always-passive group experiences a marginal increase in x_i.

To understand how optimal policy differs in this model relative to the model without internalities, suppose we initially have a default $\hat{d}$ that is optimal when no internalities are present, so $W_0^{'}(\hat{d}) = 0$. If, as in the under-saving example, $\mu > 0$, then a deviation from this default in whichever direction *increases* $X(d)$ would constitute an improvement in social welfare. Importantly, whether total activity increases with an increase in the default or with a decrease in the default is an empirical question. One might presume that the presence of positive internalities from saving would lead the planner to prefer a higher default savings rate, but the results here shows that this intuition is only correct when there are relatively few marginally active choosers. In the case where total saving is relatively unaffected by the default, the presence of internalities from under-saving is actually irrelevant for the optimal default.

The assumptions under which Proposition 5 holds are instructive but not guaranteed. Assumptions (A5.1) and (A5.2) make the problem more tractable by ensuring that the internality is approximately linear. Assumption (5.3) ensures that the marginal internality $\mu_i \equiv m_i'(x)$ is independent of other structural parameters governing individual behavior (though not necessarily of the *other* parameter summarizing mistakes, π_i). Relaxing (A5.3) but maintaining (A5.1) and (A5.2), it is straightforward to derive that:

$$\begin{aligned} W^{'}(d) = \quad & W_0^{'}(d) + \mu X'(d) + E[\mu_i - \mu|PP]P(PP) \\ & + E[(\mu_i - \mu)(x_i^a - d)|PA]P(PA) - E[(\mu_i - \mu)(x_i^a - d)|AP]P(AP) \end{aligned} \tag{11}$$

Equation (11) shows how to modify the expression in Proposition 5 to account for the fact that the mean marginal internality may be different across the three groups of decision-makers we are interested in. For example, individuals with high values of γ_i, who are particularly sensitive to defaults, might be particularly bad under-savers. In this case, the last term of Equation (11) would become more important, as the PP group have higher values of γ_i than the other groups. Such a modification would make higher defaults more attractive on the margin, relative to the expression in Proposition 5. Alternatively, we might suppose that individuals with low values of x_i^a are especially bad under-savers. As individuals with low x_i^a tend to be in the PA group, this modification would make the third term in Equation (11) larger, making an increase in the default less attractive on the margin.

In addition to affecting the optimal default, the presence of internalities affects the desirability of forcing active choice. As in Section 2.2, we can compare a penalty default d^p and a generic default d. Adapting Equation (7) to the case of internalities yields:

$$W(d^p) - W(d_0) = E\left[u_i(x^a) - u_i(d) - \pi_i\gamma_i + m_i(x_i^a) - m_i(d) \,|\, PA\right] p(PA) \tag{12}$$

Unlike in Proposition 3.1, it is no longer the case that a sufficiently small value of π guarantees that d^p is the optimal default. To see why, suppose that $\pi_i = 0$ for all individuals. As before, we know that $u_i(x_i^a) > u_i(d)$. However, it may be the case that $m_i(x_i^a) < m_i(d)$. When the difference $m_i(d) - m_i(x_i^a)$ is sufficiently large for enough individuals, d will be a better default than d^p. In the under-saving example, it is possible – though not assured – that a default under which some individuals choose passively will increase total saving relative to an active choice regime to such an extent that the active choice regime is not optimal, even when as-if costs are completely irrelevant for welfare.

To summarize, incorporating mistakes by active choosers into the model means that the planner may be able to raise social welfare by choosing a default that makes those mistakes less likely to occur, and reduces the benefits of penalty defaults that cause decision-makers to choose actively. Once internalities are considered, the case for active choice policies hinges on decision-makers making enough of one type of mistake (i.e., over-weighting the costs of opting out of a default) but not making too much of a different type of mistake (i.e., making a biased choice when selecting among non-default options). Uncertainty by the planner over the distribution of $m_i(x)$ thus creates a new difficulty for determining the optimal default that parallels the normative ambiguity caused by uncertainty over π.

3.2 Measuring Internalities

We have shown how the degree to which decision-makers make privately suboptimal choices affects optimal default policy, but empirically estimating such internalities is widely acknowledged to be one of the central challenges in behavioral public economics. When the optimal default depends on the degree to which internalities are present, there are several approaches one might take for estimating this parameter from data. First, an observer might investigate a potential internality by focusing on choices made in other contexts that do not exhibit the same biases; for example, decision-makers might be observed in a setting in which the feature of the choice environment suspected of causing a bias is known to be absent. Similarly, one might estimate the structural parameters that govern preferences over some menu by observing other choices made in a related domain. Alternatively (and more basically), one might simply ask people about their preferences or well-being; although such an approach can itself be biased, in some cases it may shed light on welfare when decision-makers' observed choices do not. Finally, a researcher might identify decision-makers' unbiased preferences by extrapolating from other decision-makers whose decisions are thought to be unbiased, such as subject-matter experts or those not subject to framing effects (Bronnenberg et al., 2013; Allcott, Lockwood and Taubinsky, 2019). For all of these approaches, once decision-makers' true (i.e., welfare-relevant) preferences are known, they can be compared to the preferences revealed by their choice behavior to infer the presence and magnitude of any internality. As others have emphasized, the feasibility and convincingness of these alternative approaches varies widely by application.

4 Empirical Illustration

This section illustrates our results using data on 401(k) plan contribution decisions. We choose to focus on this setting for two reasons. The first is that it is a setting in which defaults have been shown to affect behavior and in which the choice of default is of significant practical importance. The second is that it has been the focus of a recent and influential literature on optimal default policy; holding the setting constant in our analysis relative to this prior literature helps clarify the value added by our approach.

To preview our results, we draw two substantive conclusions from this analysis. First, we generalize the result from BFP that the uncertainty over optimal defaults is small when the range of policies considered does not include policies that promote active choice. Specifically, we estimate the mapping between values of π and the optimal default. This allows us to conclude that the optimal policy BFP identifies applies not only for the illustrative models they consider, but rather for all behavioral models within a more general class (i.e., any model that is consistent with the opt-out cost representation). Our second contribution is to show that, in contrast to the results in BFP, normative ambiguity *does* generate meaningful uncertainty as to optimal policy once the policy space is expanded to include policies that promote active choice. When the as-if costs associated with default effects are mostly irrelevant from a welfare perspective, the optimal policy is to adopt a penalty default (e.g., setting a very high default contribution rate) or to require active choice as a condition of employment. In contrast, these policies are dominated when even a modest fraction of the observed as-if opt-out costs are normatively relevant. Finally, we discuss additional analyses that help to resolve the normative ambiguity over π, and we discuss how the planner's uncertainty over π matters for welfare. The illustration also highlights two of the chief benefits of our approach, namely the simplicity with which it can be applied and the transparency between our assumptions and the welfare conclusions that emerge.

Empirical Setup. The data we use consists of 401(k) contribution rates for newly eligible employees of several firms first analyzed by Choi et al. (2004, 2006) and Beshears et al. (2008).[20] We describe the relevant features of these data here and refer readers to the earlier studies for additional detail. The first step in our approach is to estimate the distribution of parameters in the opt-out cost representation of behavior. To do so, we follow BFP and rely on a structural as-if cost model fitted on employee contribution rate data for each employer. This model assumes preferences over individual contribution rates are $u_i(x) = \rho_i \ln(x + M(x) + \alpha) + \ln(z)$, where ρ_i and α are preference parameters governing the overall preference for contributing and the price sensitivity of contributions respectively, $M(x)$ is the employer match as a function of x, and $z = 1 - (1-t)x$ is residual income pinned down by the budget constraint. We assume a marginal tax rate t of 20 percent. The firm matches 50 percent of employee contributions up to 6 percent of earnings, so we have $M(x) = 0.5x$ for $x \leq 0.06$ and $M(x) = 0.5 * 0.06$ for $x > 0.06$.

Estimating this model structurally requires assumptions about the distribution of ρ_i and γ_i. First, we assume these two variables are independent.[21] Second, we assume that γ follows an exponential distribution, modified to have a point mass λ_1 at zero, with cdf $\Phi(\gamma) = \lambda_1 + (1-\lambda_1)(1-e^{-\lambda_2 \gamma})$ for $\gamma \geq 0$ and $\Phi(\gamma) = 0$ for $\gamma < 0$. Thus, a fraction λ_1 of individuals are assumed to have zero as-if costs and choose actively under all defaults. Third, we assume ρ_i follows a censored normal distribution: $\rho_i = \max\{0, \tilde{\rho}\}$ where $\tilde{\rho}$ is distributed normally with mean μ_ρ and variance σ^2.

The cap on the employer match creates a large kink in the budget constraint at 6 percent of earnings, which induces bunching in the optimal contribution rate at 6 percent. The degree of bunching implicitly identifies the price sensitivity preference parameter, α. The model also predicts bunching at the corner solutions of 0 and the maximum contribution rate of 15 percent. Finally, we evaluate $W(d)$ using equivalent variation relative to a benchmark in which all individuals receive their most-preferred option x_i^* without incurring any costs of opting out of a default, so that indirect utility is $v_i(d) = u_i(x_i(d)) - \pi_i \gamma 1\{x_i(d) \neq d\} - u_i(x_i^*)$. As the units of x_i are in percentages of annual salaries contributed to a 401(k) plan, the units of welfare thus correspond to the fraction of annual salary that would make individuals receiving x_i^* (without any costs)

[20] For simplicity, we present results for a single firm only (the one labeled "Company 3" in BFP). Results for the other firms are very similar and are presented in the Online Appendix.

[21] This assumption allows us to extrapolate from the decisions of active choosers at some defaults, whose decisions are informative about ρ, to the population of decision-makers. As the active decision-makers will tend to have low γ_i, covariance between ρ_i and γ_i poses a challenge to identifying the preferences of passive choosers (Goldin and Reck, 2018).

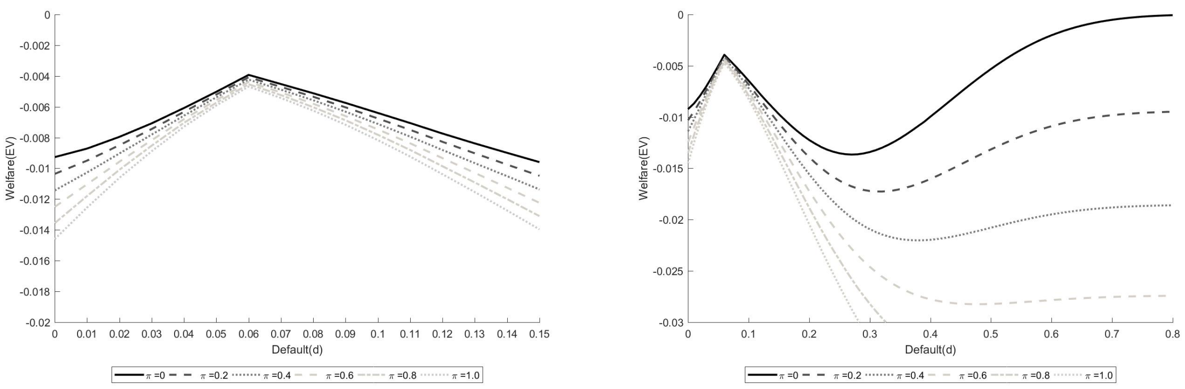
willing to switch to a default d. Because welfare is lower under a given default d than under the benchmark, equivalent variation is typically negative. Throughout the analysis, we assume a uniform value of π_i for the population, denoted π for simplicity. The calculation of welfare for any distribution of π_i, including those in which it covaries with other heterogeneous parameters, is straightforward.

Results. We first use the structural model to estimate the distribution of as if opt-out costs (γ_i). We estimate that 60 percent of employees have strictly positive opt-out costs. Among this group, the mean opt-out cost is \$3,386 for an employee with the median salary in the data (\$40,000). We estimate a mean opt-out cost for *all* employees of about 5.07 percent of their salary, or \$2,028 at the median salary. Ten percent of all employees have as-if costs greater than 15 percent of their salary, i.e. \$6,000 for an employee with median salary.

We next use this estimated distribution of as if opt-out costs to solve for the optimal default as a function of π. Figure 2a depicts equivalent variation for alternative default contribution rates between 0 and 15 percent of earnings, for values of π ranging from zero to one. We find that regardless of π, the optimal policy is to set a default of 6 percent of earnings, which is where, for all three firms, employer matching contributions are maximized. This analysis generalizes the main finding in BFP to any positive model of default effects consistent with the opt-out cost representation, and to any view of welfare within such a model. The large kink in the welfare functions at a 6 percent default contribution in Figure 2a arises because of the large kink in the budget constraint at 6 percent, due to the employer match. Without the match, there is no such kink in welfare and the optimal default tends toward the middle of the distribution of individual preferences for saving, governed by the parameter ρ_i. Relatedly, because the employer match exerts such a strong influence on welfare, these estimates are robust to very different distributions of ρ_i. As a result, concerns about how to identify heterogeneous underlying preferences over saving are immaterial here. This is a fortunate coincidence in the 401(k) context and not a general feature of the optimal default problem. We next extend the analysis to consider policies that promote active choice. To do so, we expand the range of contribution rate defaults we consider. As suggested by Proposition 3, extending the analysis in this way causes the optimal policy to depend on π. As shown in Figure 2b, extremely high defaults dominate when π is low, but the 6 percent default dominates when π is moderate or large.[22] There is a strong qualitative similarity between the stylized model in Figure 1 and the estimated model in Figure 2b; the main difference between these is the kink in the latter at the 6 percent default, which is caused by the kink in the budget

[22] A similar phenomenon is evident in Figures A.7 to A.12 in the Online Appendix of BFP, who do not focus on it because they state that the extreme contribution rates exceed statutory limits on 401(k) contributions. However, setting an extremely high default contribution rate is not the only way to force employees to make active choices – the employer could simply require it as a condition of employment, as in Carroll et al. (2009). In addition, legal constraints in this policy area may be modified, as illustrated by the Pension Protection Act of 2006, which allowed employers to automatically enroll employees in 401(k) plans.

Figure 2: Equivalent Variation by Default Contribution Rate


(a) Defaults between 0 and 15 percent of salary

(b) Expanded set of contribution rates

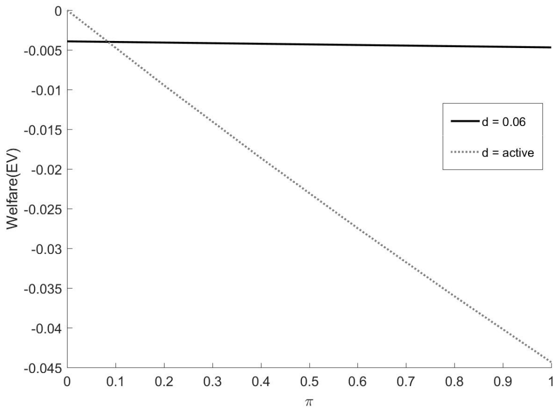
Note: This figure depicts welfare in equivalent variation units as a function of the default contribution rate for varying values of π. We observe that 6 percent is the robust optimal default when we restrict the analysis to contribution rates between 0 and 15 percent (left panel), but extreme contribution rates that promote active choice are desirable when π is small (right panel).

constraint from the employer match.

Figure 3 compares welfare under the 6 percent default to an active choice regime, for values of π ranging from zero to one. When $\pi = 0$, the active choice regime leads to the same outcome as our benchmark in which all individuals costlessly receive x_i^*, so its equivalent variation is zero. Consistent with Propositions 3 and 4, we find that at higher values of π, the 6 percent, employer-contribution-maximizing default dominates the active choice default. The 6 percent default is also the default that minimizes opt-outs. The optimal policy thus takes the form of a threshold rule: active choices dominate below the threshold, the 6 percent default dominates above the threshold. The threshold below which active choice dominates is about $\pi = 0.08$ for this firm.

Discussion. Because the threshold value of π for which active choice is optimal is $\pi = 0.08$, finding the optimal policy requires determining whether opting out of a default reduces employee welfare by 8% or more of the employee's as if opt-out cost. Above, we estimate the mean opt-out cost to be $2,028 for an employee with median salary. Eight percent of this amount is $162, which corresponds to about 0.4 percent of the median employee salary. When opting out of the default reduces welfare by at least $162 on average, the optimal policy is therefore to set the default to the contribution rate that minimizes opt-outs, which here corresponds to the match-maximizing rate of 6%. In contrast, if the normative component of opt-out costs is below $162 on average, the optimal policy is to promote active choice, either by setting the default to an extremely high contribution rate, or, perhaps more realistically, requiring employees to make an active

Figure 3: Welfare Under Active Choice versus Minimizing Opt-Outs in 401(k) Plans



Note: This figure compares welfare under the 6 percent contribution default with welfare under an active choice regime. At low values of π, the active choice regime leads to higher welfare.

choice as a condition of employment.[23]

Identifying the normative component of opt-out costs poses a methodological challenge since it cannot be directly inferred from observed choice data. As discussed in Section 2.4, however, researchers may look to other types of information for guidance. In this context, for example, an observer might estimate the normative component of opt-out costs from choices made in other settings, which do not exhibit the same biases from default effects. For example, one might look to the price decision-makers are willing to pay to avoid filling out other forms or making other decisions of similar complexity, using data on the price of paid tax preparers or financial planners. Alternatively, an observer might estimate the number of non-work hours required to actively choose one's contribution rate and fill out the form, and price that time according to the employee's implied wage rate.

In the 401(k) setting we focus on, this type of analysis suggests the normative component of as-if costs is smaller than 8 percent, which would imply that active choice is optimal. Suppose that employees value their time at \$19 per hour, which is the equivalent hourly wage rate of an employee in our sample with median salary and a 40-hour work week. The threshold value of \$162 would imply that, on average, the process of opting out and making an active choice would need to take more than 8 hours of the employee's time in order for the 6 percent default to dominate an active choice policy. We suspect that this amount of time is on the high side of what most casual observers would consider plausible. The actual process of filling out forms and selecting an option typically takes less than an hour, though of course researching the available options

[23] Our focus is on the optimal default option but we acknowledge that other policy parameters, such as the match rate, also significantly affect welfare. Relatedly, an employer's goal may not be solely to maximize employees private welfare (Bubb and Warren, 2018). A higher default could increase employer's matching contributions, and the associated cost might lead employers to set a lower default or reduce matching contributions, relative to the optimal policy considered here. Such issues would not be present for national automatic pension enrollment policies like those recently adopted in the United Kingdom, where the employee's pension contribution has no direct effect on the employer's contribution.

to determine which one will be best takes more time. Relatedly, given the cognitive difficulty of making pension choices, individuals may be willing to pay more to avoid making such choices than the $19/hour benchmark implied by their wages. Still, if we suppose the process takes 2 hours, employees would need to value the time it takes to opt out at about $80/hour, which seems high for an individual making $40,000 per year. These conclusions match the intuition of others as well – BFP calibrate some of their models using parameters that effectively impose a value of $\pi = 0.01$.

However, our results also imply a sense in which promoting active choice is riskier than minimizing opt-outs. Figure 3 shows that minimizing opt-outs leads to equivalent variation ranging from 0.4 to 0.5 percent of annual earnings (about $160 to $200 at the median earnings) relative to the first-best benchmark, for values of π from zero to one. Active choice leads to equivalent variation ranging from 0 to 4.4 percent of annual earnings ($0 to $1,760 at the median earnings). We can incorporate this idea into our analysis by formalizing the planner's uncertainty over π. Given a uniform distribution over some range of π deemed plausible by the planner, we can compare the expected equivalent variation between active choices and minimizing opt-outs by integrating the difference between $W(d)$ for these two policies in Figure 3. With a uniform distribution over $[0, 1]$, the expected equivalent variation from active choice is -2.3 percent of annual earnings and the expected equivalent variation of minimizing opt-outs is -0.4 percent of earnings. The expected equivalent variation of minimizing opt-outs is much higher partly as a consequence of Propositions 3 and 4: when π is small, minimizing opt-outs is sub-optimal, but remains a local optimum; in contrast, when π is large, active choice minimizes social welfare. More fundamentally, when minimizing opt-outs, π only matters for the welfare of active choosers, who tend to have low as-if costs already. When setting an active choice policy, π matters for the welfare of all decision-makers, even those with very high as-if costs.

Incorporating Internalities. We next consider how internalities shape the optimal 401(k) contribution default. Whether and how much individuals' are biased in their assessments of how much to save for retirement is a subject of debate in the literature. On the one hand, many individuals report wishing they saved more for retirement (Thaler and Sunstein, 2008, p.107); on the other hand, individuals' actual savings choices may be more reliable than the preferences they state when responding to surveys (Carroll et al., 2009). Similarly, alternative methodological approaches yield quite different conclusions about the share of individuals who are optimally saving for retirement, with estimates ranging from less than 50% (Munnell et al., 2014) to greater than 80% (Scholz, Seshadri and Khitatrakun, 2006). For a summary of this literature, see Poterba (2015).[24]

[24] Much of this literature imposes strong assumptions on the content of decision-makers' preferences. An alternative path for shedding light on the magnitude of the internality would be to compare the distribution of actual contribution decisions to the decisions employees make in settings where their choices are more likely to be optimal, such as those made following an

To incorporate under-saving into our analysis, one approach would be to calibrate the magnitude of such internalities for our population by applying the under-saving estimates reported in the literature. A limitation of our data for this purpose is that we lack information on wealth and on contribution rates by age. In addition, the literature does not distinguish between under-saving due to passive choices at low defaults and under-saving for any other reason, while our results suggest that distinguishing between these is key for understanding the optimal defaults. For these reasons, and because the literature offers no clear guidance on the magnitude of under-saving, we focus instead on illustrating how various assumptions about the degree of internality shape the optimal 401(k) contribution rate default.

To derive the optimal default as a function of the internality, we follow the approach described in Section 3. We assume that each individual is subject to an internality over their chosen savings rate, including employer matching contributions, of $m_i(x) = \mu \ (x_i + M(x_i))$, for some uniform constant μ.[25] We can interpret μ in terms of a lump-sum-equivalent increase in salary in percentage terms; for example, when $\mu = 0.1$, the increase in the individuals' welfare of switching from contributing 0 percent of one's salary to the pension to contributing 10 percent is equivalent to a lump-sum transfer of 1 percent of the worker's salary.

Figure 4a presents the optimal policy as a function of the internality (μ) and degree to which opt-out costs are normative (π). Our earlier results are nested in the case where $\mu = 0$: the optimal policy is active choice when $\pi < 0.08$ and a 6 percent employee contribution otherwise. As μ increases, active choice becomes sub-optimal for even low values of π. The primary advantage to active choice is that it allows each individual to select an individually optimal contribution rate; this advantage is eliminated when internalities are sufficiently large. As μ continues to rise, the optimal default increases up to the maximum contribution of 15 percent. This occurs because when μ is very large, the effect of an increase in the default on the saving of (always) passive savers dominates everything else from a welfare perspective.[26] This increase happens for any value of π, but it happens fastest when $\pi = 0$. When π is larger, opt-outs are costly, so the fact that increasing the default above 6 percent also leads to an increase in opt-outs makes higher defaults less desirable than when $\pi = 0$.

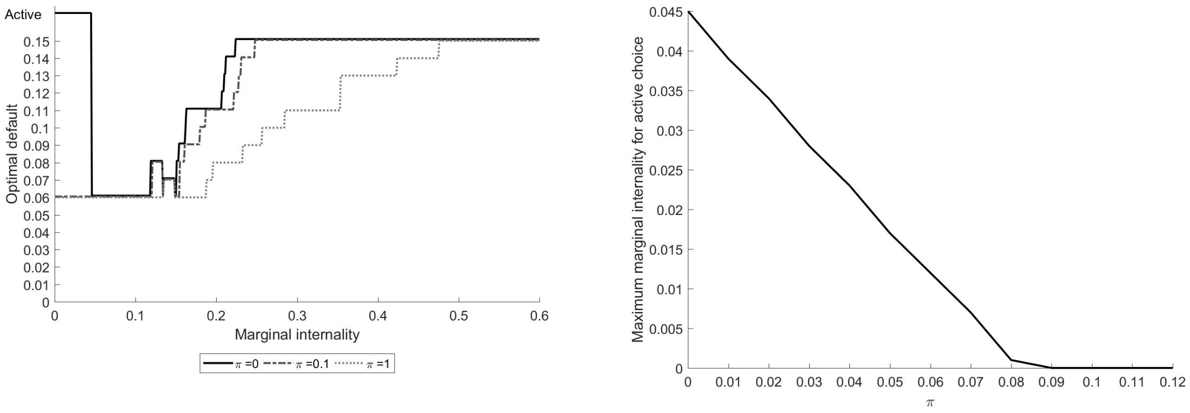
Figure 4b illustrates this phenomenon, plotting the largest possible value of μ for which the active choice policy is optimal over different values of π. When $\mu > 0.045$, the active choice policy cannot be optimal. Recall that a value of $\mu = 0.045$ would imply that a 10 percent increase in contributions (potentially

expert-guided retirement planning session.

[25] From the proof of Proposition 5, we can interpret $\mu_i \approx -u''(x_i^* - x_i^a)$ under (A5.1) and (A5.2). Hence, the assumption that μ is constant basically requires that the amount of under-saving under active choice, $x_i^* - x_i^a$, is relatively homogeneous among employees.

[26] This finding is not mechanical: if the model suggested that a default below the maximum contribution maximized total saving, then that could be the optimal default for arbitrarily large μ. Such an internal default does not exist here because the as-if costs are quite large, so that a large number of employees remain passive for even relatively high defaults.

Figure 4: The Influence of Internalities from Under-Saving on the Optimal Default



(a) Optimal Default by the Marginal Internality

(b) How Large an Internality Makes Active Choice Sub-Optimal?

Note: This figure illustrates how internalities affect the optimal default policy. Figure (4a) depicts the optimal default for varying values of the internality μ and three values of π. We observe that active choices are dominated by the 6 percent default when μ is sufficiently large. As μ continues to increase, the optimal default increases up to the maximum allowable contribution. Figure (4b) plots the highest value of μ under which the active choice policy is optimal, for varying values of π. As π becomes large, the active choice policy becomes less desirable regardless of μ because of opt-out costs. Hence, for larger π, a smaller internality is necessary to rule out the active choice policy. For reference, marginal internalities of 0.06, 0.12, and 0.18 correspond to the after-match savings rate being 10, 20, and 30 percent below the optimum, respectively (see Appendix Figure 5).

including the employer match) increases the total internality by the equivalent of a lump-sum transfer of 0.45 percent of earnings. We show in Appendix Figure 5 that in our model, this value of μ corresponds to an average degree of under-saving of approximately 7 percent (or 0.6 percent of workers' salary). Similarly, once $\pi > 0.08$, our previous results imply that active choice is always sub-optimal, regardless of μ. In contrast, for sufficiently low values of both π and μ, active choice remains optimal. We take three main conclusions from this analysis. First, the introduction of internalities introduces another caveat to the use of active choice policies: if individuals are systematically biased toward under-saving, a default that induces higher saving and relatively few opt-outs can be preferable. Second, when internalities are sufficiently large, they can dominate all other considerations with respect to the optimal default. Finally, if the main lesson of our previous analysis was that the optimal default pension contribution depends on difficult normative judgments, the possibility of internalities only makes the difficulties worse. Policymakers must decide not only whether to respect decision-makers' revealed opt-out costs, but also whether to respect the preferences decision-makers reveal when they make active choices.

5 Conclusion

Uncertainty over the decision-making model that generates an observed behavior is a pervasive source of difficulty in behavioral economics. Under a range of positive models of default effects, decision-makers' behavior can be described using "as-if" preferences over opt-out costs revealed by their observed choices. Revealed preference analysis can recover information about these as-if preferences, but cannot answer whether these as-if preferences accurately reflect individuals' welfare. Our analysis of the optimal default problem clarifies the conditions in which optimal policy determinations do and do not depend on the degree to which these as-if opt-out costs are normatively relevant.

Two kinds of empirical evidence can help with the normative judgments necessary to pin down optimal defaults. The first is to use various interventions attempting to reduce or enhance default effects to shed light on the positive mechanisms driving default effects, as in Blumenstock, Callen and Ghani (2017). Even with such evidence, however, one must make a judgment on whether a mechanism acts by imposing normative costs or by driving a wedge between choices and welfare. The second potential strategy is to gather external evidence to make an informed judgment. This approach requires assumptions about what choices in other settings, such as those that reveal the monetary value of workers' time, or those in which workers describe a target retirement consumption level, tell us about normative parameters. We gave examples of this type of reasoning in our examination of pension defaults. While both of these strategies can help the planner make an informed choice, neither can resolve the problem without external judgments about how preferences are revealed by choices.

Finally, although the form of normative ambiguity we focus on is new, our proposed approach follows an established tradition in public finance of parameterizing certain normative judgments. In particular, optimal policy analyses typically incorporate judgments about the value of equity by employing social welfare weights (Mirrlees, 1971). Indeed, optimal policy analyses can be usefully divided based on whether their prescriptions do or do not rely on such judgments (e.g., Pareto versus non-Pareto improvements). Our results suggest that a similar division, between policy conclusions that require resolving normative ambiguity and those that do not, will be fruitful for "behavioral" policy analysis.

References

Abaluck, Jason, and Abi Adams. 2017. "What do consumers consider before they choose? Identification from Asymmetric Demand Responses." Working Paper.

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

Altmann, Steffen, Armin Falk, Paul Heidhues, and Rajshri Jayaraman. 2016. “Defaults and donations: Evidence from a field experiment.” Working Paper.

Ayres, Ian, and Robert Gertner. 1989. “Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules.” *The Yale Law Journal*, 99(1): 87–130.

Bernheim, B Douglas, Andrey Fradkin, and Igor Popov. 2015. “The Welfare Economics of Default Options in 401(k) Plans.” *American Economic Review*, 105(9): 2798–2837.

Beshears, John, James J Choi, David Laibson, and Brigitte Madrian. 2008. “The Importance of Default Options for Retirement Savings Outcomes: Evidence from the United States.” In *Lessons from Pension Re-form in the Americas.* , ed. Stephen J Kay and Tapen Sinha, 59–87. Oxford University Press.

Beshears, John, Schlomo Benartzi, Richard Mason, and Katherine L Milkman. 2017. “How Do Consumers Respond When Default Options Push the Envelope?”

Blumenstock, Joshua, Michael Callen, and Tarek Ghani. 2017. “Why do Defaults Affect Behavior? Experimental Evidence from Afghanistan.” Working Paper.

Bronnenberg, Bart, Jean-Pierre Dube, Matthew Gentzkow, and Jesse Shapiro. 2013. “Do Pharmacists Buy Bayer? Sophisticated Shoppers and the Brand Premium.”

Brown, Zachary, Nick Johnstone, Ivan Haščič, Laura Vong, and Francis Barascud. 2013. “Testing the effect of defaults on the thermostat settings of OECD employees.” *Energy Economics*, 39: 128–134.

Bubb, Ryan, and Patrick L Warren. 2018. “An Equilibrium Theory of Retirement Plan Design.” Working Paper.

Camerer, Colin, Samuel Issacharoff, George Loewenstein, Ted O’donoghue, and Matthew Rabin. 2003. “Regulation for Conservatives: Behavioral Economics and the Case for" Asymmetric Paternalism".” *University of Pennsylvania Law Review*, 1211–1254.

Carroll, Gabriel D, James J Choi, David Laibson, Brigitte C Madrian, and Andrew Metrick. 2009. “Optimal Defaults and Active Decisions.” *The Quarterly Journal of Economics*, 124(4): 1639–1674.

Chesterley, Nicholas. 2017. “Defaults, Decision Costs and Welfare in Behavioural Policy Design.” *Economica*, 84(333): 16–33.

Chetty, Raj. 2012. “Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply.” *Econometrica*, 80: 969–1018.

Chetty, Raj, John N Friedman, Søren Leth-Petersen, Torben Nielsen, and Tore Olsen. 2014. “Active vs. Passive Decisions and Crowd-Out in Retirement Savings Accounts: Evidence from Denmark.” *The Quarterly Journal of Economics*, 129(3): 1141–1219.

Choi, James J, David Laibson, Brigitte C Madrian, and Andrew Metrick. 2004. “For Better or for Worse: Default Effects and 401 (k) Savings Behavior.” In *Perspectives on the Economics of Aging.* , ed. David A Wise, 81–126. University of Chicago Press.

Choi, James J, David Laibson, Brigitte C Madrian, and Andrew Metrick. 2006. “Saving for Retirement on the Path of Least Resistance.” In *Behavioral Public Finance: Toward a New Agenda.* , ed. Edward J McCaffery and Joel Slemrod. Russell Sage Foundation.

Choukhmane, Taha. 2019. “Default Options and Retirement Saving Dynamics.”

Ghilarducci, Teresa. 2019. “Americans Do Not Have Enough Retirement Savings, Really.” *Forbes.*

Goldin, Jacob, and Daniel Reck. 2018. “Revealed Preference Analysis with Framing Effects.”

Goldin, Jacob, and Nicholas Lawson. 2016. “Defaults, Mandates, and Taxes: Policy Design with Active and Passive Decision-Makers.” *American Law and Economic Review.*

Haggag, Kareem, and Giovanni Paci. 2014. "Default Tips." *American Economic Journal: Applied Economics*, 6(3): 1–19.

Handel, Benjamin, and Joshua Schwartzstein. 2018. "Frictions or Mental Gaps: What's Behind the Information We (Don't) Use and When Do We Care?" *Journal of Economic Perspectives*, 32(1): 155–178.

Handel, Benjamin R. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *The American Economic Review*, 103(7): 2643–2682.

Heiss, Florian, Daniel McFadden, Joahim Winter, Amelie Wupperman, and Bo Zhou. 2016. "Inattention and Switching Costs as Sources of Inertia in Medicare Part D."

Johnson, Eric J, and Daniel Goldstein. 2003. "Do Defaults Save Lives?" *Science*, 302(5649): 1338–1339.

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

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *The Quarterly Journal of Economics*, 443–477.

Madrian, Brigitte, and Dennis Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *Quarterly Journal of Economics*, 116(4): 1149–1187.

Masatlioglu, Yusufcan, and Efe A Ok. 2005. "Rational choice with status quo bias." *Journal of Economic Theory*, 121(1): 1–29.

Masatlioglu, Yusufcan, Daisuke Nakajima, and Erkut Y Ozbay. 2012. "Revealed Attention." *The American Economic Review*, 102(5): 2183–2205.

Mirrlees, James A. 1971. "An Exploration in the Theory of Optimum Income Taxation." *The Review of Economic Studies*, 38(2): 175–208.

Munnell, Alicia H, Wenliang Hou, Anthony Webb, et al. 2014. *NRRI update shows half still falling short.* Center for Retirement Research at Boston College Chestnut Hill, MA.

Poterba, James M. 2015. "Saver heterogeneity and the challenge of assessing retirement saving adequacy." National Tax Association.

Scholz, John Karl, Ananth Seshadri, and Surachai Khitatrakun. 2006. "Are Americans saving optimally for retirement?" *Journal of political economy*, 114(4): 607–643.

Thaler, Richard H, and Cass R Sunstein. 2003. "Libertarian Paternalism." *American Economic Review*, 175–179.

Thaler, Richard H, and Cass R Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Yale University Press.

Tversky, A., and D. Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science.*

Tversky, Amos, and Daniel Kahneman. 1991. "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *Quarterly Journal of Economics.*

Zeiler, Kathryn. 2017. "Mistaken About Mistakes." *European Journal of Law and Economics.*

Appendix (For Online Publication Only)

A Appendix: Proofs

Lemma 1:

$$W(d) = E[u_i(x_i^*) - \pi_i\gamma_i | a_i(d) > 0]\ (1 - F_{a;d}(0)) + E[u_i(d) | a_i(d) \leq 0]\, F_{a;d}(0)$$

where $F_{a;d}(\cdot)$ denotes the cumulative density function of $a_i(d)$.

Proof: From the definition of the social welfare function we know that $W(d) = E[v_i(d)]$. By the law of iterated expectations,

$$W(d) = E[v_i(d) | a_i(d) > 0] P(a_i(d) > 0) + E[v_i(d) | a_i(d) \leq 0] P(a_i(d) \leq 0)$$

We know from the consumer's problem and the definition of $a_i(d)$ that 1) $a_i(d) \leq 0 \implies x_i(d) = d$ and 2) $a_i(d) > 0 \implies x_i(d) = x_i^* = \arg\max u_i(x)$. Substituting these into $v_i(d) = w_i(x_i(d), d) = u_i(x_i(d)) - \pi_i\gamma_i 1\{x_i(d) \neq d\}$ gives the result. ■

Proposition 1: *For any two defaults $d_0, d_1 \in X$:*

$$W(d_1) - W(d_0) = E[u_i(x^*) - u_i(d_0) - \pi_i\gamma_i | PA]\, p(PA) - E[u_i(x^*) - u_i(d_1) - \pi_i\gamma_i | AP]\, p(AP) + E[u_i(d_1) - u_i(d_0) | PP]\, p(PP).$$

Proof: We know that $W(d_1) - W(d_0) = E[v_i(d_1) - v_i(d_0)]$. We partition individuals into the four groups (PA, AP, PP and AA) and apply the law of iterated expectations to express the change in welfare as a probability-weighted sum over these four groups. As before, $a_i(d) \leq 0 \implies x_i(d) = d$ and 2) $a_i(d) > 0 \implies x_i(d) = x_i^* = \arg\max u_i(x)$. In the PA group, $a_i(d_1) > 0$ so $v_i(d_1) = u_i(x_i^*) - \pi_i\gamma_i$,and $a_i(d_0) \leq 0$, so $v_i(d_0) = u_i(d_0)$. Thus $E[v_i(d_1) - v_i(d_0) | PA] = E[u_i(x^*) - u_i(d_0) - \pi_i\gamma_i | PA]$. Proceeding similarly for the other four groups and substituting in the resulting expressions yields the desired result. ■

Proposition 2: *Let X be any interval in $\mathbb{R}$. If d^* represents an interior solution to the optimal default problem, the following first-order condition is satisfied:*

$$\begin{aligned} 0 = W^{'}(d^*) \quad &= \quad E[(1-\pi_i)\gamma_i|a_i(d^*)=0,\ u_i'(d^*)<0]\, f_{a|u'<0}(0)\, F_{u'}(0) \\ &- \quad E[(1-\pi_i)\gamma_i|a_i(d^*)=0,\ u_i'(d^*)>0]\, f_{a|u'<0}(0)\ (1-F_{u'}(0)) \\ &+ \quad E\left[u'(d^*)\,|\, a_i(d^*)<0\right]\ F_{a;d^*}(0) \end{aligned}$$

where $f_{a|u'>0}$ is the probability density function of $a_i(d^*)$ conditional on $u_i'(d^*)>0$; $F_{u'}$ is the cumulative density function of $u_i'(d^*)$; and, as above, $F_{a;d^*}$ is the cumulative density function of $a_i(d^*)$.

Proof: One can obtain this result by direct calculation of the derivative of the welfare function, as divided into active and passive choosers in Lemma 1 (i.e. expressing the expectations as integrals and applying Leibniz rule). One can also obtain the result by plugging in $d_1 = d_0 + \Delta d$ in Proposition 1, taking the limit as Δd approaches zero, plugging in the definitions of the primitives, and noting that the PA and AP groups now both have $a_i(d)=0$, which implies that $u_i(x^*) - u_i(d) = \gamma_i$ by construction. ■

Proposition 3 Suppose that there exists a penalty default $d_p \in X$.

(3.1) *There exists a threshold $\underline{\pi} \in [0,1)$ such that $\pi_i \leq \underline{\pi}$ for all i implies d_p maximizes social welfare.*

Proof: We will prove the existence of a threshold $\underline{\pi} \in [0,1)$ such that when $\pi_i \leq \underline{\pi}$, $W(d^p) \geq W(d)$ for any d.

Let $X^A \subset X$ be the subset of X such that for any $d \in X^A$, $P(a_i(d) \leq 0) > 0$.

Let $d \in X$ be an arbitrary default. We know $W(d^p) \geq W(d)$ is trivially true when d is also a penalty default, i.e. $d \notin X^A$ as then $W(d) = W(d_p)$ for any π. Next suppose $d \in X^A$, so $p(PA) > 0$. Let $\tilde{\pi}(d) = \sup_{i \in PA(d)} \pi_i$ be the largest possible value of π_i for the PA group for default d. We know from Equation (7) that

$$W(d_p) - W(d) \geq p(PA)\{E[u_i(x^*) - u_i(d)|PA] - \tilde{\pi}(d)E[\gamma_i|PA]\} \tag{13}$$

The RHS of this expression is a continuous and strictly monotonically decreasing function of $\tilde{\pi}(d)$ (so long as $E[\gamma_i|PA] > 0$, which must be true because PA individuals choose passively). When $\tilde{\pi}(d) = 0$, the RHS of this expression is weakly positive because $u_i(x^*) \geq u_i(d)$ for all i.[27] When $\tilde{\pi}(d) = 1$, the RHS is strictly negative because $u_i(x^*) - u_i(d) < \gamma_i$ for all individuals that are passive at d, which is the PA group in this

[27] The expression is *strictly* positive if we presume x_i^* is a *unique* maximum for every individual. Under this assumption, we know that the penalty default d_p is the *uniquely* optimal default.

situation. The Intermediate Value Theorem then implies there is a value of $\tilde{\pi}(d)$, such that we know that the expression on the RHS of (13) is 0. Denoting this threshold by $\underline{\pi}(d)$, we have that $W(d_p) - W(d) \geq 0$ when $\pi_i \leq \underline{\pi}(d)$ for all i. The result then follows from letting $\underline{\pi} = \inf_{d \in X^A} \underline{\pi}(d)$, so that $\pi_i < \underline{\pi}$ implies $W(d_p) - W(d) \geq 0$ for any d. ■

(3.2) *There exists a threshold* $\overline{\pi} \in (0, 1]$ *such that* $\pi_i \geq \overline{\pi}$ *for all* i *implies* d_p minimizes *social welfare.*

Proof: The proof is analogous to the proof of (3.1). For any default d, let $\hat{\pi}(d) = \inf_{i \in PA(d)} \pi_i$. Using equation (7) and a similar Intermediate Value Theorem argument to the above we derive that there is a threshold $\overline{\pi}(d)$, such that $\pi_i \geq \overline{\pi}(d)$ implies $W(d_p) - W(d) \leq 0$. The result then follows from letting $\overline{\pi} = \sup_{d \in X^A} \overline{\pi}(d)$. ■.

Proposition 4 *Suppose that* $X = [x_{min}, x_{max}] \subseteq \mathbb{R}$ *and that:*

(A4.1) *As-if costs* γ_i *are distributed independently of* x_i^*.

(A4.2) *Preferences are given by* $u_i(x) = u(x - x_i^*)$ *for some map* $u : \mathbb{R} \to \mathbb{R}$, *with* $u'(0) = 0$, $u'' < 0$ *and* $u(c) = u(-c)$ *for any* c.

(A4.3) x_i^* *follows a single-peaked and symmetric distribution about some mode* x^m.

Under these conditions, there exists a threshold $\overline{\pi} \in (0, 1]$ *such that* $\pi_i \geq \overline{\pi}$ *for all* i *implies that the optimal default is the default that minimizes opt-outs.*

Proof: We provide the proof of the theorem for the case when $\pi_i = 1$ for all i. It is straightforward to show that if the theorem holds when $\pi_i = 1$ for all i, it must hold for sufficiently high π_i.

Starting from the case where $\pi_i = 1$ for all i, we first prove that $W'(x^m) = 0$, $W'(d) > 0$ for $d < x^m$, and $W'(d) < 0$ for $d > x^m$, which implies that W has a unique global maximum at x^m. We then prove that opt-outs are minimized under x^m. We start by letting $d \in X$ be some default.

Step 1: Characterizing the first and second derivative of $W(d)$.

Let $W_\gamma(d) = E[v_i(d) | \gamma_i = \gamma]$. By (A4.1) we know that $W(d) = \int_\gamma W_\gamma(d) f(\gamma) d\gamma$. To prove our result, it therefore suffices to prove that for any fixed γ, $W_\gamma^{'}(d) = 0$ if $d = x^m$, and $W_\gamma''(d) < 0$ always.

We first introduce some notation involving the function $u()$. Without loss of generality $u(0) = 0$. Taking γ as given, by (A4.2) there is some unique value ξ such that $u(\xi) = u(-\xi) = \gamma$. Note that when $x^* = d - \xi$, utility at the default is given by $u(d - x^*) = u(d - (d - \xi)) = u(\xi) = \gamma$, and similarly when $x^* = d + \xi$, $u(d - (d + \xi)) = \gamma$. By (4.2), an individual is active when $x_i^* \leq d - \xi$ or $x_i^* \geq d + \xi$.

We next characterize $W'_\gamma(d)$. For illustrative purposes, suppose $\pi_i = \pi$ is homogeneous for all i, which is true when $\pi_i = 1$ for all i. Welfare for people with given γat d is given by

$$W_\gamma(d) = \int\limits_{x^*=-\infty}^{d-\xi} -\pi\gamma f(x^*)dx^* + \int\limits_{x^*=d-\xi}^{d+\xi} u(d-x^*)f(x^*)dx^* + \int\limits_{x^*=d+\xi}^{\infty} -\pi\gamma f(x^*)dx^*,$$

Where $f(x^*)$ is the pdf of x_i^*. Note that $f(x^*)$ does not depend on γ by (A4.1). Differentiating the above with respect to d and applying $u(d - x_i^*) = \gamma$ at $x_i^* = d - \xi$ or $d + \xi$, we obtain

$$W'_\gamma(d) = \gamma(1-\pi)[f(d-\xi) - f(d+\xi)] + \int\limits_{x^*=d-\xi}^{d+\xi} u'(d-x^*)f(x^*)dx^* \tag{14}$$

This is an analogue of Proposition 2 for some fixed γ, with the added structure of (A4.2). When $\pi = 1$ (A4.4), the first term of this expression, which corresponds to the PA and AP groups, vanishes, leaving only the PP group, which we now split into those with $x^* < d$ and those with $x^* > d$:

$$W'_\gamma(d) = \int\limits_{x^*=d-\xi}^{d} u'(d-x^*)f(x^*)dx^* + \int\limits_{x^*=d}^{d+\xi} u'(d-x^*)f(x^*)dx^*. \tag{15}$$

Step 2: For any constant ζ, $f(d+\zeta) \geq f(d-\zeta) \iff x^m \geq d$.

Suppose $x^m \geq d$ and take a constant ζ. If $x^m > d+\zeta > d-\zeta$, the result immediately follows from the assumption in (A4.3) that $f()$ is single-peaked. If $d+\zeta \geq x^m \geq d > d-\zeta$ take a constant c such that $d+\zeta-x_m = x_m - c$. By symmetry about x^m, $f(c) = f(d+\zeta)$. We know that $c < x_m$, because $x_m - (d+\zeta) \leq 0$. We also know that $c \geq d-\zeta$, because we presumed $x_m \geq d$. We then have $x^m \geq c \geq d-\zeta$. The single-peaked assumption then implies $f(d+\zeta) = f(c) \geq f(d-\zeta)$.

Supposing $x^m < d$ and proceeding analogously proves the converse.

Step 3: $x^m \geq d \iff W'_\gamma(d) \geq 0$.

Starting from equation (15), note that by (A4.2) the first term is positive ($u' > 0$ when $x^* < d$) and the second term is negative ($u' < 0$ when $x^* < d$). We can compare the signs of the two terms in the previous expression by re-writing this equation, using the symmetry of the utility function, as:

$$W'_\gamma(d) = \int\limits_{x^*=d-\xi}^{d} u'(d-x^*)[f(x^*) - f(\tilde{x})]dx^*$$

where $\tilde{x} = 2d - x^*$, so that $d - x^* = -(d-\tilde{x})$. We know from symmetry that when $d = x^m$, $f(x^*) = f(\tilde{x})$, so $W'(x^m) = 0$.

As $u'(d-x^*)>0$ in the range of integration we use above. When $x^m > d$, the result in Step 3 implies that $f(x^*) \geq f(\tilde{x})$ for $x^* \in [d-\zeta, d]$, so we know that $W'_\gamma(d) \geq 0$. When $x^m < d$, the result in step 2 implies that $f(x^*) \leq f(\tilde{x})$ for $x^* \in [d-\zeta, d]$, and we know that $W'_\gamma(d) \leq 0$.

Step 3 proves that there is a unique global maximum of W at x^m.

Step 4: Setting $d = x^m$ minimizes opt-outs.

Let the frequency of opt-outs be given by $A(d) = P(a_i(d) > 0)$. Using $\xi = u^{-1}(\gamma)$ from before and letting F be the cdf of x_i^*, we know that

$$A(d) = F(d-\xi) + 1 - F(d+\xi)$$

Taking a derivative with respect to d, we have that

$$A'(d) = f(d-\xi) - f(d+\xi).$$

Setting $d = x^m$, it is straightforward to verify using (A4.3) that $A'(d) = 0$ if $d = x^m$, $A'(d) < 0$ if $d < x^m$, and $A'(d) > 0$ if $d > x^m$, which is sufficient to prove that x^m minimizes $A(d)$. ■

Proposition 5 *In the model with internalities, suppose that*

(A5.1) *For all i, $u_i(x) = -\frac{\alpha}{2}(x - x_i^a)^2$ with $\alpha > 0$.*

(A5.2) *Normative preferences are given similarly by $u_i(x) + m_i(x) = -\frac{\alpha}{2}(x - x_i^*)^2$.*

(A5.3) *The error in active choice $x_i^a - x_i^*$ is independent of x_i^a and γ_i.*

Then the marginal social welfare effect of a change in the default is given by $W_0'(d) + \mu X'(d)$, where $W_0(d)$ denotes social welfare without internalities (see Equation (6)), $\mu = E[\mu_i]$, and $X(d) = E[x_i(d)]$.

Proof **Step 1:** (A5.1) and (A5.2) imply that the internality $m(x)$ is linear.

Note that $u_i'' = -\alpha$ under (A5.1). By (A5.1) we can write

$$u_i(x) = \frac{u''(0)}{2}(x - x_i^a)^2.$$

By (A5.2) we can write

$$u_i(x) + m_i(x) = \frac{u''(0)}{2}(x - x_i^*)^2.$$

Subtracting the previous expression from this one and simplifying we obtain

$$m_i(x) = -u''(x_i^* - x_i^a)x + \frac{u''}{2}(x_i^{*2} - x_i^{a2}). \tag{16}$$

The second term is a constant with respect to x, and may therefore be safely ignored.

Step 2: Proving the result.

The result essentially follows from Equations (10) and the following equation from the text

$$\frac{\partial E[x_i(d)]}{\partial d} = E\left[x_i^a - d|PA\right]\ P(PA) + E\left[d - x_i^a|AP\right]\ P(AP) + P(PP) \tag{17}$$

Specifically, apply the linear internality to this equation to obtain:

$$\begin{array}{rcl} W^{'}(d) & = & W_0^{'}(d) + E\left[\mu_i(x_i^a - d)\,|\,PA\right]\ P(PA) \\ & - & E\left[\mu_i(x_i^a - d)\,|\,AP\right]\ P(AP) \\ & + & E\left[\mu_i\,|\,PP\right]\ P(PP). \end{array}$$

Next, note that $\mu_i = m_i'(x) = -u''(x_i^* - x_i^a)$ by (16). Applying (A5.3) then implies that we can pull out the $E[\mu_i]$ terms.

$$W^{'}(d) = W_0^{'}(d) + \mu\{E[x^a - d|PA]P(PA) - E[x^a - d)|AP]P(AP) + P(PP)\}$$

Noting that the term inside curly brackets is the expression for $X'(d)$ in Equation (17), we obtain the desired result. ■

B Relationship to the Axiomatization of Masatlioglu and Ok (2005)

Masatlioglu and Ok (2005) provides an axiomatic characterization of a model very similar to the fixed as-if cost model we use. Their paper seeks to rationalize status quo bias; recall that we showed in Section 1.1 that giving extra utility to the status quo is the same as having a fixed cost of not choosing the status quo (see Section 1.1). The representation of choices used by Masatlioglu and Ok (see their equations (3) and (4)) is isomorphic to our own (see our equation (2), and Section 1.1), with one exception: the fixed as-if cost could depend on the default in their model. Whether and to what extent γ depends on d is difficult to test empirically, but we know of no evidence suggesting that it does. Nevertheless, here we discuss further the implications of our restriction that γ does not depend on d by relaxing it and examining welfare.

Consider a model that is identical to our baseline model except that the fixed cost is a function of d for each individual, denoted $\gamma_i(d)$. It is straightforward to show that the derivative from Proposition 2 becomes

$$\begin{aligned} 0 \;=\; W'(d) \;=\;& \quad E\left[\pi_i \gamma_i'(d)\right| AA] \, P(AA) \\ &+ \; E\left[(1-\pi_i)\gamma_i(d)\right| PA] \, P(PA) \\ &- \; E\left[(1-\pi_i)\gamma_i(d)\right| AP] \, P(AP) \\ &+ \quad E[u'(d)|PP]P(PP). \end{aligned} \tag{18}$$

This expression is identical to the expression in Proposition 2 except for the first term. In our basic model, individuals that are always active for a change in the default do not experience any change in their welfare. When the fixed costs depend on d and $\pi_i > 0$, changing the default can affect the welfare of these decision-makers because . The analogue of equation (5) is also straightforward to derive for this model.

First, we note that the argument in Proposition 3 (see the proof above) for active choices being optimal for sufficiently low π is unaffected by this addition. When as-if costs are not normative, forcing active choices still leads all individuals to receive x_i^* without incurring any costs. Whether forcing active choices *minimizes* welfare for sufficiently high π is unclear. The difficulty is that the penalty default d^p could in principle have a lower fixed cost ($\gamma(d^m)$) than other defaults, which can make the penalty default relatively more attractive than some other defaults.

We know by the same logic as Proposition 4 (proof above) that the last three terms of (18) will all be zero under (A3.1)-(A3.3) when we minimize opt-outs, and that ignoring the changes in $\gamma(d)$ for active choosers we would get to a global optimum by minimizing opt-outs when π_i is sufficiently high for all individuals. The additional term in Equation (18) therefore implies that minimizing opt-outs will not be optimal in general when the change in $\gamma(d)$ for a marginal change in the default is zero. Intuitively, if increasing the default from the opt-out minimizing default would reduce the cost incurred by active decision-makers, we know the aggregate effect on all other decision-makers is zero (by Proposition 3), so such an increase in the default would be an improvement on minimizing opt-outs. For a more extreme example, suppose there is a default d^* such that $\gamma_i(d^*) = 0$ for all i. Such a default is obviously the optimal default regardless of the π_i's.[28]

To summarize, our result that active choices are desirable when default effects are purely driven by behavioral frictions survives the extension implied by the model of Masatlioglu and Ok (2005). Minimizing opt-outs will still be a good rule of thumb when default effects are real costs and the dependence between the costs and the default is not too strong, but if the costs vary strongly with the default it may be possible to improve on the opt-out minimization rule of thumb.

[28] When $\pi = 0$, both the active choice policy and the default d^* are optimal defaults.

C Variable Opt-Out Costs

Thus far we have assumed that as-if opt-out costs are constant (for a given individual) and do not depend on which non-default option the decision-maker selects. An alternative behavioral model is that defaults "pull" decision-makers towards options near the default in addition to making them more likely to select the default itself. For example, defaults may serve as an anchor (Example 1.2.7).

Ultimately, the question of whether defaults effects can be better described by including variable as-if costs in the model is an empirical question. Empirical evidence, reviewed in Section 1.2, regularly finds that increases in the default can affect choices far away from the default, suggesting that fixed costs are likely present. A variable costs model alone, such as a model of anchoring and adjustment where a higher default tends to lead to higher $x_i(d)$, would not predict, for example, that the fraction of individuals who contribute nothing to their pension would increase when the default rate of contribution is increased. Whether adding variable costs gives the model *additional* explanatory power relative to the fixed-cost-only model is more difficult to test. One possibility is to look closely at choices around the default. The fixed costs model with no variable cost predicts a "hole" in the observed distribution of choices around the default, whereas adding variable costs model predicts a "hill" around the default when fixed costs are sufficiently low. Still, given that both fixed and variable costs are plausibly heterogeneous, separately identifying these two components of decision-makers' revealed preferences without strong assumptions about distributions of the two costs is difficult. Here, we show how the inclusion of a variable costs affects the conclusions of our main analysis, especially the desirability of active choices versus minimizing opt-outs.

We focus on the case where X is a real interval. Suppose that instead of (1), individual behavior is given by

$$x_i(d) = \arg\max_{x \in X} \; u_i(x_i) - c_i(x_i - d) - \gamma 1\{x_i \neq d\}. \tag{19}$$

For simplicity, we will assume that u_i is single-peaked, with $u_i'(x_i^*) = 0$ and $u_i'' < 0$ everywhere. For this extension, we assume that the as-if cost associated with choosing a non-default option increases the further the chosen option is from d, so that $c_i'(x_i - d) \geq 0$ when $x_i - d > 0$, and $c_i'(x_i - d) \leq 0$ when $x_i - d < 0$. The as-if cost function is twice differentiable, with $c'' \geq 0$. We normalize $c(0)$ to zero. In this model individuals choose the default when passive, or $\tilde{x}(d) = \arg\max u_i(x_i) - c_i(x_i - d)$ when active. The individual is active if $\tilde{a}_i(d) \equiv [u_i(\tilde{x}_i(d)) - c_i(\tilde{x}_i(d) - d)] - u_i(d) - \gamma_i > 0$.

Similar to before, welfare is given by

$$w_i(x) = u_i(x) - \rho_i c_i(x - d) - \pi_i \gamma_i 1\{x_i \neq d\}, \tag{20}$$

where ρ_i denotes the normative relevance of variable costs $c_i(\cdot)$ and π_i the normative relevance of fixed costs as before. Indirect utility and social welfare are also defined similarly to before.

Given any change in the default, we can divide individuals into four groups as before, except now these groups are based on $\tilde{a}_i(d)$. Taking a derivative of the welfare function with respect to d, we have that the necessary condition from Proposition 2 becomes, with the addition of variable costs,

$$\begin{array}{rcrl} 0 & = & W'(d) = & E\left[\rho_i c_i' + (1-\rho_i) c_i' \frac{c_i''}{c_i''-u_i''} \middle| AA\right] P(AA) \\ & & + & E\left[(1-\rho)c + (1-\pi_i)\gamma_i \right| PA] P(PA) \\ & & - & E\left[(1-\rho)c + (1-\pi_i)\gamma_i \right| AP] P(AP) \\ & & + & E[u'(d)|PP]P(PP). \end{array} \tag{21}$$

where all components involving $c_i(\cdot)$ are evaluated at $x = \tilde{x}(d)$.

Adding variable costs changes this expression in two ways. First, the always-active choosers (AA) are affected by a change in the default. The sign of the welfare effect on an always-active chooser is positive if and only if $x_i^* < d$. For an individual with $x_i^* < d$, we will have that $x_i^* < x_i(d) < d$, and an increase in the default makes it costlier to choose an option close to x_i^*. The $\rho_i c_i'$ term of the welfare effect for members of the AA group in Equation 20 corresponds to the direct welfare effect of increasing this cost. Such an individual also increases x_i in response to this change in costs: it is straightforward to show that $\tilde{x}_i'(d) = \frac{c_i''}{c_i''-u_i''} \in [0,1)$. The second term of the welfare effect for the AA group corresponds to the welfare impact of this change in behavior.[29] As before, when as-if costs are fully normatively relevant for all individuals, $\rho_i = 1$, and the envelope theorem eliminates the indirect welfare effect from the behavioral response. However, when $\rho_i < 1$, the individual over-reacts to the increase in costs, reducing their welfare. The opposite intuition applies when $x_i^* > d$; such individuals in the AA group are made better off by an increase in the default. The second addition to the welfare calculation is the extra variable cost incurred by marginally active decision-makers in the PA and AP groups. As it changes welfare discretely when the individual switches between choosing actively and choosing passively, this component affects welfare in exactly the same fashion as the fixed cost.

Our key result that forcing active choice is optimal when default effects are driven purely by behavioral frictions will still be true in this model, but properly examining an active choice policy requires subtle reasoning here. In this model, setting an extreme default so that everyone opts out will not necessarily be equivalent to forcing active choices directly. One might naturally suppose that when forcing active choices, the planner sets no anchor, which eliminates the variable costs, whereas when a penalty default acts as an anchor, the variable costs will matter for behavior and welfare. Suppose there is a policy that forces decision-makers to make active choices and eliminates variable costs (i.e. it does not set an anchor). It

[29] Note that the behavioral response is $\tilde{x}'(d) = 0$ when costs are linear, i.e. $c'' = 0$.

is straightforward to show that such a policy will be globally optimal when π_i is sufficiently small for all individuals (regardless of ρ_i), exactly as in Proposition 3. However, whether such a policy becomes extremely undesirable when π_i and ρ_i are sufficiently high is not clear in this model, because the policy that forces active choices also eliminates the variable costs and this can improve welfare. Conversely, a penalty default will surely minimize welfare when π_i and ρ_i are sufficiently high, but due to the large distortions on active choices it may have through the variable costs, it may not be optimal when π_i and ρ_i are large.

By a very similar procedure to the one we use in Proposition 4, one can show that minimizing opt-outs is optimal when π_i and ρ_i are sufficiently large, under some regularity conditions. Specifically, we could maintain Assumptions (A4.1)-(A4.3), and add the assumption that the variable cost function is the same for all individuals, $c_i(x_i - d) = c(x_i - d)$. Under these assumptions minimizing opt-outs will still be globally optimal when default effects are driven by real components of individual welfare.

D Additional Details from Empirical Application

This Appendix provides additional results for our empirical application. First, we show in Figure 5 how the marginal internalities from Figure 4 map to the mean optimal savings rate, $E[x_i^*]$. At $\mu = 0$, the mean savings rate corresponds to the observed savings rate when we simulate the model under the active choice policy, which is a 7 percent contribution not including the employer matching contributions, or just over 9 percent when we add in the match. As μ reaches larger values, the optimal savings rate increases and approaches the maximum 15% contribution asymptotically. Interestingly, as μ increases, the optimal default in Figure 4a approaches the maximum contribution more quickly than the mean optimal savings rate does. This finding might seem counter-intuitive at first, but it occurs because there is a mass of individuals contributing 15%, and the mass grows as μ increases. Like the 6% default before, this mass point is an attractive default because it gives a large number of people their exactly ideal option and it leads few people to opt out (see BFP Theorem 2).

Next, we show some results from two other employers, labeled “Company 1” and “Company 2” in BFP, along with additional details on the parameters used to calculate welfare in the model. In the model used here, the difference in distributions of contributions at different companies is used to identify differences in μ_ρ, which governs overall preferences over savings rates. Different companies also have different matching contribution rates. All other parameters are the same across companies. A complete table of parameter values is contained in Table 1.

Figure 5: Marginal Internalities and Mean Optimal Contribution Rates

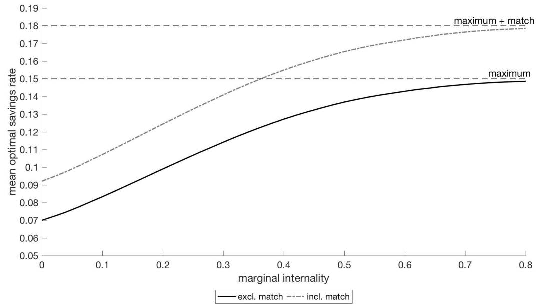


Table 1: Model Parameters and Plan Characteristics

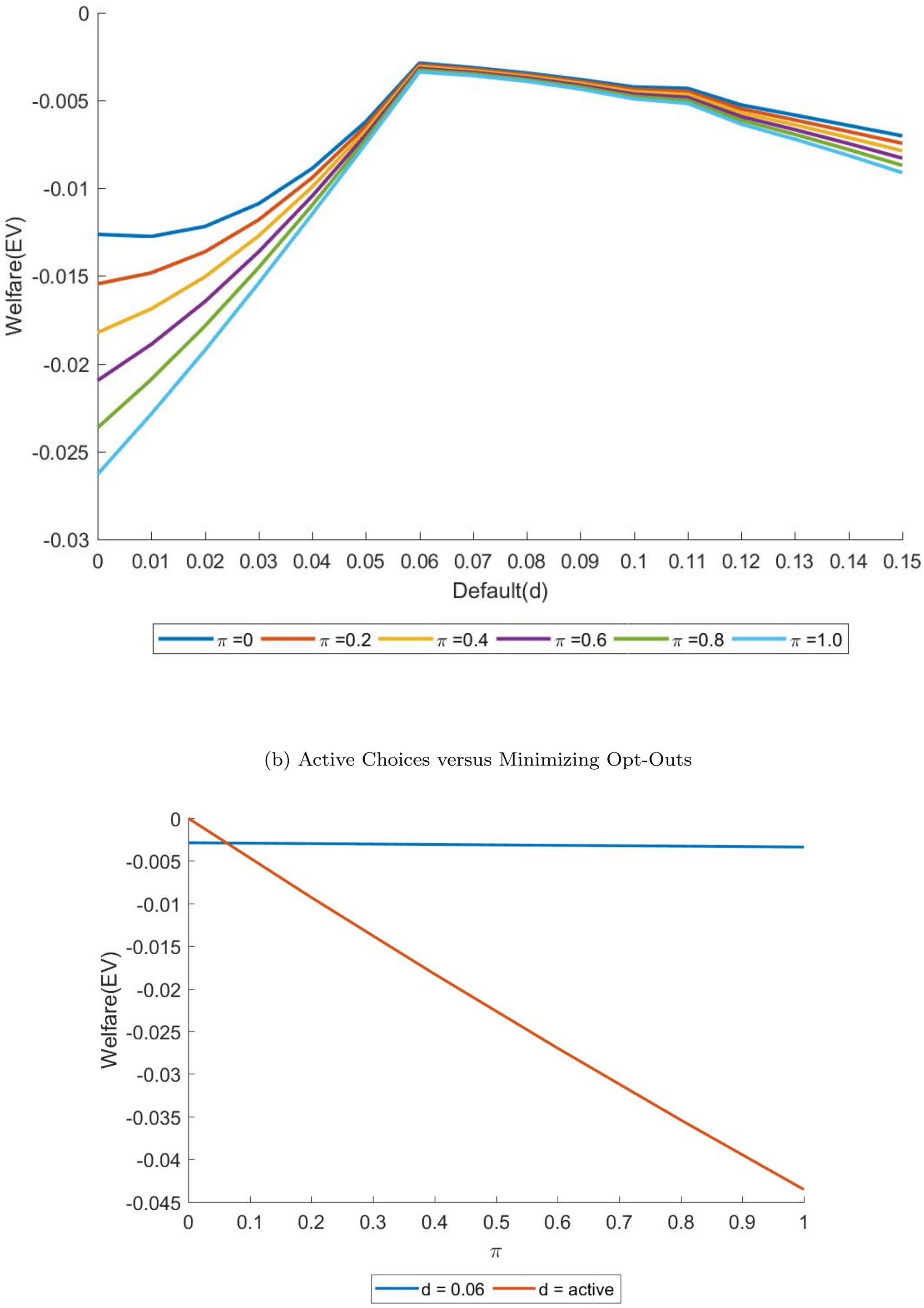
Parameter	Value
Mean savings utility weight, μ_ρ, company 1	0.2150
Mean savings utility weight, μ_ρ, company 2	0.1313
Mean savings utility weight, μ_ρ, company 3	0.1570
Standard deviation of savings utility weight, σ	0.0910
Savings shift parameter, α	0.1340
Fraction with zero as-if costs, λ_1	0.4011
As-if costs distribution parameter, λ_2	11.81
Maximum matched contribution (all companies)	0.06
Employer match rate, company 1	1.0
Employer match rate, company 2	0.5
Employer match rate, company 3	0.5

Note: this table reports the parameter values we use in our empirical illustration. The parameter values come from Table 2 of Bernheim, Fradkin and Popov (2015), for the “basic model.”

Figure 6 repeats Figures 2 and 4 in the body of the text for Company 1. Figure 7 does the same for Company 2. We can see that apart from relatively minor differences, we obtain the same results for all three companies. The most noticeable difference is actually that the higher, 100 percent match rate at company 1 makes defaults lower than 6 percent much less desirable, which is intuitive.

Figure 6: Results for Company 1

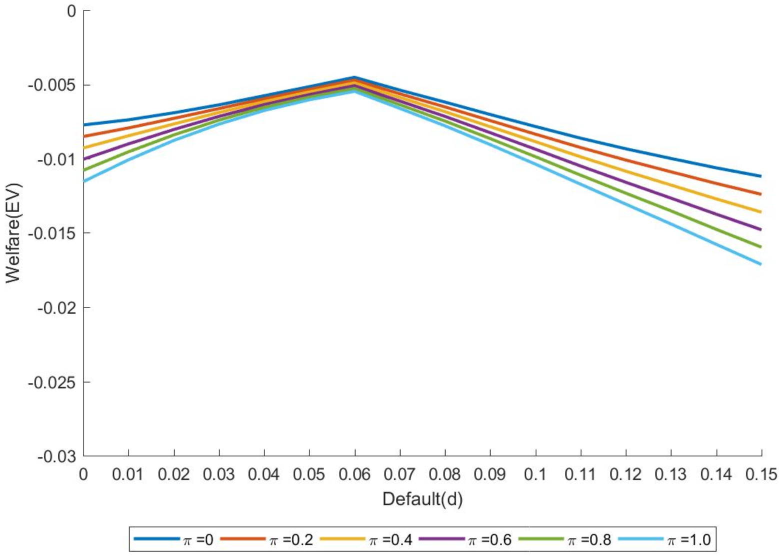
(a) Equivalent variation over defaults, by π



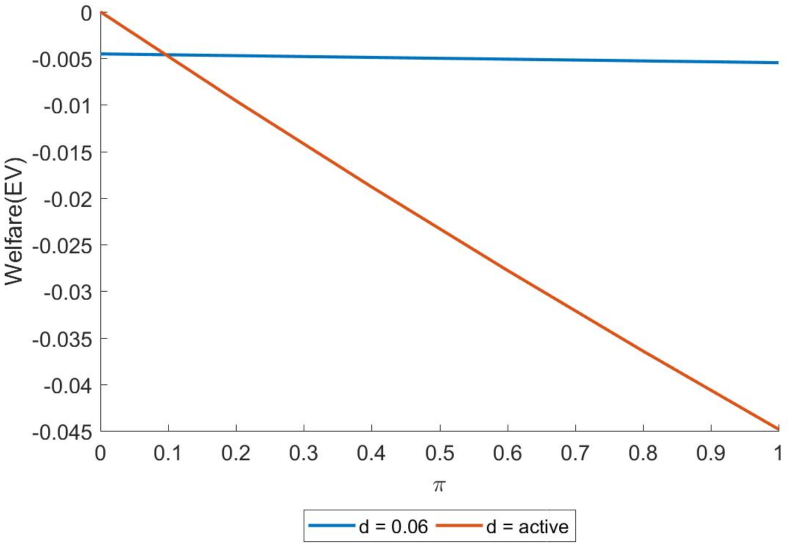
(b) Active Choices versus Minimizing Opt-Outs

Figure 7: Results for Company 2

(a) Equivalent variation over defaults, by π



(b) Active Choices versus Minimizing Opt-Outs


Welfare(EV)
π
d = 0.06
d = active