American Economic Journal: Economic Policy 2020, 12(2): 22–45
https://doi.org/10.1257/pol.20180605

An Equilibrium Theory of Retirement Plan Design[†]

By Ryan Bubb and Patrick L. Warren*

We develop an equilibrium theory of employer-sponsored retirement plan design using a behavioral contract theory approach. The operation of the labor market results in retirement plans that generally cater to, rather than correct, workers' mistakes. Our theory provides new explanations for a range of facts about retirement plan design, including the use of employer matching contributions and the use of default contribution rates in automatic enrollment plans that lower many workers' savings. We provide novel evidence for our theory from a sample of plans. (*JEL* D86, G51, J26, J32, J41)

Employer-sponsored retirement savings plans are the predominant vehicle for private retirement savings in the United States. A growing literature shows that the design of these plans affects savings behavior in ways inconsistent with rational optimization (e.g., Madrian and Shea 2001, Thaler and Benartzi 2004). These empirical findings have informed normative claims by behavioral economists about how employers *should* design their plans. In a survey of this literature, for example, Benartzi and Thaler (2007, 99) ask, "What can employers do so that more plan participants enroll in retirement plans, contribute an amount that will build a reasonable retirement nest-egg, and allocate the funds among assets in an appropriately diversified way?" They proceed to suggest to employers a range of plan design options to improve their workers' retirement savings outcomes. Employers should paternalistically harness the stickiness of default rules, for example, by automatically enrolling workers in order to counteract present-biased workers' temptation to save too little (Thaler and Benartzi 2004, Carroll et al. 2009). These papers take a "public finance" approach to retirement plan design, modeling the employer as if it acts as a paternalistic social planner, designing its retirement plan to maximize social welfare. In response to this literature, Congress enacted the Pension Protection Act of 2006 (PPA), which removed regulatory barriers to employers automatically enrolling their workers in their retirement plan (see Beshears et al. 2010 for an account of the legislative process). Employers have adopted automatic enrollment in droves, with the

* Bubb: New York University School of Law, 40 Washington Square South, New York, NY 10012 (email: ryan.bubb@nyu.edu); Warren: John E. Walker Department of Economics, Clemson University, 222 Sirrine Hall, Clemson, SC, 29634 (email: patrick.lee.warren@gmail.com). Dan Silverman was coeditor for this article. A previous version of this paper circulated under the title "A Positive Theory of Retirement Plan Design." We are grateful to Ian Ayres, Oren Bar-Gill, John Beshears, Louis Kaplow, Lewis Kornhauser, David Laibson, Josh Schwartzstein, and workshop participants at NYU School of Law, the NYU-Penn Law and Finance Conference, the NBER Summer Institute, the American Law and Economics Association Annual Meeting, the Annual Conference of the Society of Institutional and Organizational Economics, and the Vanderbilt Financial Regulation and Consumer Choice conference for helpful comments and discussions.

[†] Go to https://doi.org/10.1257/pol.20180605 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

percentage of Vanguard-administered plans that use automatic enrollment increasing from 15 percent in 2007 to 41 percent in 2015 (Vanguard 2016).

The existing literature, however, has not considered whether employer paternalism in plan design is consistent with their incentives and, in particular, with the incentives produced by the operation of the labor market. We develop an equilibrium theory of employer-sponsored retirement plan design using a behavioral contract theory approach. Retirement plans are an important feature of compensation contracts designed by employers to attract workers. The approach we take to modeling firms is neoclassical: in our model, firms maximize profits, not social welfare. The rational benchmark entails a simple wage contract. Retirement plans serve no useful purpose for rational, time-consistent exponential discounters (tax benefits aside). However, following the behavioral literature on retirement savings, we allow workers' decision utility at the time of contracting to deviate from their experienced utility in canonical ways and characterize the equilibrium retirement plan designs that result.

In Section I, we consider present-biased workers with varying degrees of sophistication and first assume that they can costlessly enroll in their retirement plan, so that the default contribution rate of the plan is irrelevant. Perfectly sophisticated present-biased workers, who understand that they suffer from a time-inconsistency problem, value retirement plans with employer contributions as a form of commitment, and in equilibrium they receive a plan that acts as a first-best commitment device. Imperfectly sophisticated (or "naïve") present-biased workers, in contrast, overestimate their future savings. As a result, naïve present-biased workers overvalue firms' offers to match their retirement savings and hence receive retirement plans with matching in equilibrium. While matching contributions can help offset naïve workers' present bias, their level in equilibrium is not finely calibrated to workers' need for commitment. Moreover, with heterogeneous types, considered in Section II, matching results in cross-subsidization of exponential workers by naïve present-biased workers.

We then assume that enrollment is costly so that a plan's default employee contribution rate is potentially sticky. We assume that at the time of contracting, workers do not consider these costs of enrolling and hence do not anticipate that their savings will be sensitive to the default. We show that the equilibrium default *minimizes* workers' savings given the other terms of the contract. The reason is that lowering workers' savings reduces the level of matching contributions employers must make, relaxing their zero-profit constraint, and thus allows employers to offer better terms on the salient dimensions of compensation. With heterogeneous types, the equilibrium contract entails a cap on the employer's matching contributions, with the default set strictly below the cap. The equilibrium cap limits the anticipated distortion that matching would otherwise cause to exponentials' intertemporal consumption choices, while setting the default below the cap minimizes the employer's matching contributions.

To develop a useful benchmark against which to assess the empirical predictions of our equilibrium model, we formalize an alternative paternalistic employer model in Section III. In particular, we consider the set of zero-profit contracts that maximize average worker welfare under our assumptions about worker behavior.

Paternalistic employers can always implement the first best by using matching contracts with a default set at the cap on the employer's matching contributions. Under such a contract, all types stay with the default, avoiding bearing the costs of opting out, and still perfectly smooth their consumption.

In Section IV, we use our equilibrium model and the competing paternalistic employer model to analyze the effect of the PPA's liberalization of defaults on savings outcomes and welfare. Under the paternalistic employer model, allowing firms to automatically enroll their workers is not useful. The reason is that a paternalistic employer has contractual instruments other than the default to improve savings outcomes, namely employer contributions. Under the equilibrium model, in contrast, the PPA poses an important set of trade-offs. Allowing employers to automatically enroll workers can increase the savings of present-biased workers who would not opt into the plan on their own, but at a cost of lowering their total compensation through increased redistribution to less present-biased workers. Moreover, the PPA can lower the savings and welfare of more moderately present-biased workers, and it increases the redistribution to exponentials and distorts exponentials' intertemporal consumption choices.

Our theory provides novel explanations of many facts about employer retirement plan design, showing the power of applying standard models of market equilibrium to understanding these plans. Most defined contribution plans offer matching contributions, and a substantial fraction of workers in such plans fail to contribute enough to receive the full match. Moreover, most employers that have adopted automatic enrollment have chosen the minimum default initial contribution rate allowed under the regulatory safe harbor Congress created for such plans, which is widely understood to be below the optimal savings rate of most workers. Existing evidence on the effects of automatic enrollment on contribution rates is based on the study of a few specific employer plans rather than a representative sample of plan designs, but it is consistent with the basic predictions of our theory. In particular, on the one hand the adoption of automatic enrollment increases the participation rate in the plan; but on the other hand, it lowers other workers' savings rates by anchoring them at a lower default savings rate than they would have chosen had they opted in on their own (Choi et al. 2004).

The most distinctive prediction of our theory relative to existing accounts of plan design is that among automatic enrollment plans that offer matching, the default will be set below the cap on the employer match. In contrast, a paternalistic employer would generally set the default at the cap on matching contributions, which is a point of accumulation in employees' optimal savings rates (Bernheim, Fradkin, and Popov 2015). In Section V, we test these competing predictions using a large hand-coded dataset of the plan designs of a representative sample of automatic enrollment plans that use matching. We find that about three-quarters of such plans set defaults below their cap on matching, consistent with our equilibrium theory (and inconsistent with the paternalism theory). But a minority of plans use a default right at the cap, which is consistent with the paternalism theory (and inconsistent with our equilibrium theory), implying that considerations outside of our model are also important determinants of plan design. On the whole, this evidence shows that our theory provides an empirically relevant new perspective on plan design.

The approach we take to analyzing employer-sponsored retirement plans builds on an existing literature in behavioral contract theory that so far has focused on firms' product markets, such as consumer credit (DellaVigna and Malmendier 2004, Heidhues and Kőszegi 2010, Bar-Gill 2012), cell phone service (Grubb 2009), add-on goods (Gabaix and Laibson 2006), insurance (Gottlieb and Smetters 2016), and gym memberships (DellaVigna and Malmendier 2006). Our theoretical contribution to this literature is twofold. First, we include a default rule (in our application, a default contribution rate) into the contract space along with a cost of opting out. Such default rules have been examined empirically (e.g., Madrian and Shea 2001) and normatively (e.g., Carroll et al. 2009), but they have not been previously studied in a positive equilibrium model like ours.

Second, we are the first to apply a behavioral contract theory approach to studying retirement plans, an application of central concern to economists. One justification for not taking such an approach to understanding employer-sponsored retirement plans is the view that markets do not provide important incentives for employers with respect to retirement plan design. For example, Barr, Mullainathan, and Shafir (2013, 444) argue that attempts to boost participation in retirement plans face "at worst indifferent and at best positively inclined employers and financial firms." They contrast this with other markets, such as consumer credit, in which firms have strong incentives to exploit consumer mistakes. Spiegler (2015) similarly suggests that employers act as paternalistic de facto market regulators through their retirement plan design. But as we show in this paper, a standard equilibrium model in which firms maximize profits and workers maximize their decision utility produces a rich positive theory that matches many key stylized facts about employer-sponsored retirement plan design. There are, of course, motivations for employers in designing their retirement plans that we ignore in our model—including reputational concerns, regulations such as the nondiscrimination rules, tax incentives, and employee departure incentives—but the simple theory of the firm we apply here is the natural place to begin.

Our equilibrium theory suggests that recent attempts by behavioral economists to reform employer-sponsored retirement plans by simply showing employers what plan designs would improve worker savings outcomes and removing regulatory barriers to offering them (see, e.g., Thaler and Benartzi 2004; Orszag, Iwry, and Gale 2006) may be ineffective. A general theme of our analysis is that equilibrium plan designs generally cater to rather than correct workers' mistakes. If the motivation for retirement savings policy generally and the preferential tax treatment of employer plans specifically is to correct mistakes workers make in planning and saving for retirement (Kotlikoff 1987), then our analysis shows that the delegation of plan design to employers could result in perverse outcomes for the present-biased and inertial workers that retirement savings policy aims to help.

I. Homogeneous Types

Consider a perfectly competitive labor market populated by homogeneous workers. This can also be thought of as the case in which firms observe workers' types so that each type gets its own contract. Labor contracts specify a wage $w \geq 0$ and

a retirement plan that is composed of a nonelective employer contribution to the plan $r \geq 0$, employer matching contributions of $m \geq 0$ dollars for every dollar the worker saves for retirement, and a default employee contribution rate $d \geq 0$. Total employer retirement plan contributions are thus $r + sm$, where $s \geq 0$ is the amount the worker contributes to the retirement plan.[1] Profits from an employed worker are given by $\pi = \gamma - w - sm - r$, where γ is the value the worker produces. Workers have access to a savings technology through their employer's retirement savings plan with a rate of return normalized to zero, but they cannot borrow. For simplicity, we have assumed away any motivation to save outside of the employer's retirement plan.

There are three periods in which the sequence of decisions is as follows:

- Period 0: Firms make contract offers (w, r, m, d), and workers choose among offers.
- Period 1: Workers receive wage w and decide how much of the wage to save, s, consuming the remainder, $w - s$.
- Period 2: Retired workers consume their savings and retirement plan benefits, $r + (1 + m)s$.

A worker's period 0 self (self 0) has utility $u(c_1) + u(c_2)$, where c_i is anticipated consumption in period i, $u(\cdot)$ is increasing and concave, and the discount factor is normalized to 1. Self 1, by contrast, chooses savings to maximize the utility function $u(c_1) + \beta u(c_2) - k\mathbf{1}(s \neq dw)$, where $\beta \in (0, 1]$ is the worker's time-inconsistent, present-bias factor and $k \geq 0$ is the cost of adjusting savings from the default savings rate d. We include such an opt-out cost in order to allow for defaults to be sticky, as has been documented in the empirical literature (Madrian and Shea 2001, Choukhmane 2019). Thus, facing a contract (w, r, m, d), self 1 solves

$$\max_{s \geq 0} u\big(w - s\big) + \beta u\big(r + \big(1 + m\big)s\big) - k\mathbf{1}\big(s \neq dw\big). \tag{1}$$

If $\beta < 1$, we refer to the worker as "present-biased." If $\beta = 1$, we refer to the worker as "exponential."

Self 0 chooses a contract to maximize her utility, taking into account her anticipated future savings behavior. Importantly, however, we assume that self 0 *believes* that self 1 will choose savings by applying a present-bias factor $\hat{\beta} \in [\beta, 1]$, following O'Donoghue and Rabin's (2001) approach to modeling partial naïveté. We refer to present-biased workers with $\hat{\beta} = \beta$ as sophisticated and to those with $\hat{\beta} > \beta$ as naïve.

We assume moreover that self 0 believes that opting out of the default will be costless, so that she ignores the defaults of competing contracts ex ante. Note that this assumption applies to the exponentials in our model as well as to present-biased

[1] We restrict our attention to linear contracts in this section and consider piecewise linear contracts in the next section, but the optimal contracts in a more general contract space would share the same equilibrium features. Naïve workers would anticipate lower first-period consumption and higher second-period consumption than they actually receive because their period 1 selves will be unwilling to save enough to take full advantage of employer contributions. Linear contracts are more robust to gaming (Holmstrom and Milgrom 1987) and uncertainty about workers' types (Carroll 2015), especially in a context where savings decisions and payoffs actually happen over time.

agents. There are two distinct motivations for this assumption. The first is as a form of overconfidence. Consumers are often overoptimistic about their likelihood of completing troublesome tasks, such as refinancing debt (Shui and Ausubel 2005), going to the gym (DellaVigna and Malmendier 2006), mailing back rebate forms (Silk 2004), or limiting their cell phone usage (Grubb and Osborne 2015). Believing $k = 0$ is a kind of overconfidence in which workers are overoptimistic about their likelihood of updating their 401(k) savings plan. Naïve present bias might itself produce exactly this form of overconfidence, since it can induce unanticipated procrastination.

The second motivation is based on limited attention. Attention requires effort, and thus is applied selectively and drawn differentially to certain features of the environment (Kahneman 1973). There are various theories in the economics literature for what determines which features of an environment are salient and therefore attract attention (see, e.g., Bordalo, Gennaioli, and Shleifer 2013; Gabaix 2014; Schwartzstein 2014). But the basic idea here is that small details of the benefits package—like the default contribution rate of the retirement savings plan—might not attract the attention of workers in their labor market contracting decisions. Note, however, that we assume that the employer contributions to the plan *are* salient at the time of contracting.

To investigate the plausibility of these differential salience assumptions—that the default is nonsalient and the employer contributions are salient—we look to available advice to job seekers on how to evaluate job offers. In particular, we examined the first 20 Google search results for the phrase "how to evaluate a job offer." After discarding one result from a UK website, we found that about half (9 out of 19) of the results explicitly mention employer contributions, but none mention automatic enrollment or the default contribution rate of the retirement plan. This suggests that employer contributions are commonly considered by job seekers, but the default contribution rate is not.

Note that for simplicity, we assume workers have only a single job and make only a single savings choice under that job's retirement plan. This means that there is no opportunity for workers to learn over time their true costs of opting out of the default, as is typical in models of naïveté.[2] Furthermore, this assumption means that there is no opportunity for workers to update their savings choices midcareer. Choukhmane (2019) finds that workers who were subjected to automatic enrollment at one employer and then move to a new employer that does not use automatic enrollment have lower participation rates and contribution rates at the new employer. This dynamic effect is consistent with a rational life-cycle savings model in which workers adjust their savings rates based on accumulated savings to date. But as will become clear, all of our basic results would continue to apply if we allowed for multiple jobs over a worker's career and such wealth effects.

Firms are willing to offer any contract that would result in nonnegative profits, given workers' actual savings behavior, but perfect competition implies that firms

[2] As evidence that such naïveté persists despite labor market experience, Choi et al. (2002) report results from a survey of employees at a large US food corporation. Two-thirds of respondents reported that their current retirement savings rate was "too low." Of these, 35 percent reported that they intended to increase their 401(k) contribution rate, with most planning to do so in the next 2 months. But of those who planned to increase their contribution rate in the next few months, only 14 percent actually increased their contribution rate in the next 4 months.

must break even in equilibrium. Equilibrium labor contracts are the zero-profit contracts that maximize self 0's utility given her beliefs about self 1's savings behavior. They are thus the solution to

$$\max_{w,r,m,d} u\big(w - s\big(w,r,m|\hat{\beta}\big)\big) + u\big(r + (1+m)s\big(w,r,m|\hat{\beta}\big)\big), \tag{2}$$

subject to

$$w + r + ms\big(w,r,m,d|\beta\big) = \gamma, \tag{3}$$

$$s\big(w,r,m|\hat{\beta}\big) = \arg\max_{s\geq 0} u\big(w-s\big) + \hat{\beta}u\big(r + (1+m)s\big), \tag{4}$$

and

$$s\big(w,r,m,d|\beta\big) \in \arg\max_{s\geq 0} u\big(w-s\big) + \beta u\big(r + (1+m)s\big) - k\mathbf{1}\big(s \neq dw\big). \tag{5}$$

Self 0 wants to maximize the sum of her utility from consumption in the two periods, as reflected in the objective function in (2). The zero-profit constraint (3) requires that total compensation paid across the two periods must equal the worker's product γ. By concavity of the utility function, the first-best outcome equates consumption in each of the two periods at $\gamma/2$. Self 0 chooses a contract based on her belief that self 1 will put a present-bias factor of $\hat{\beta}$ on second-period utility when choosing how much to save under the contract; her anticipated savings level is determined by (4). Her self 1 will actually make savings decisions according to (5), using a present-bias factor of $\beta \leq \hat{\beta}$ and a cost of opting out of the default of k.

A. *Costless Opt Out*

We begin with the special case of costless opt out, $k = 0$, in which defaults are not sticky and play no role. We thus ignore defaults in this subsection and characterize contracts as triplets (w,r,m).

Consider first a sophisticated present-biased worker. A sophisticated worker's self 0 beliefs about her self 1's savings are correct, since $\hat{\beta} = \beta$. The problem for a sophisticated worker's self 0 is to choose a contract that induces her present-biased self 1 to save optimally. It is easy to see that a sophisticated worker will be willing to choose $r = w = \gamma/2$ to solve her time-inconsistency problem through r and achieve the first best. This contract will give self 1 exactly what self 0 wants her to consume. Self 1 will want to consume even more than $\gamma/2$ in the first period, but the remaining $\gamma/2$ of her compensation is only paid in the second period through r.

A sophisticated worker can also achieve the first best through m. The first-order condition for self 1's choice of savings in (5) is

$$-u'\big(w - s\big(w,r,m|\beta\big)\big) + \beta(1+m)u'\big(r + (1+m)s\big(w,r,m|\beta\big)\big) = 0. \tag{6}$$

Thus, choosing m such that $1+m = 1/\beta$ will perfectly counterbalance self 1's present bias, inducing self 1 to make savings decisions according to self 0's

preferences, i.e., to equate her consumption in the two periods. Denote this m as $m^{FB} \equiv (1-\beta)/\beta$.

In the case of an exponential worker, $m^{FB} = 0$ because matching would inefficiently subsidize second-period consumption, leading to a costly distortion in exponentials' intertemporal consumption choices. Note that this result depends on our assumption that workers cannot borrow, since borrowing would allow them to avoid any intertemporal distortion. Exponentials are better off receiving their compensation through the lump-sum payments of w and r. They are indifferent among zero-profit contracts with $r \leq \gamma/2$, since they can simply choose savings to achieve the first-best levels of consumption under any such contract. Because both sophisticated workers and exponential workers have correct beliefs about their future behavior, they receive first-best contracts in equilibrium.

In contrast, a naïve worker's self 0 underestimates her degree of present bias and hence her need for commitment. But a naïve worker also has a different motivation for using m: she overestimates how much she will save under a given m and therefore the amount of matching contributions she will receive. Matching is therefore a relatively cheap way to deliver period 0 utility. Even a completely naïve worker with $\hat{\beta} = 1$, who has no awareness of her time inconsistency and therefore no demand for commitment devices per se, will nonetheless demand some amount of matching contributions due to this mistake. The following proposition formally characterizes the equilibrium under costless opt out.

PROPOSITION 1: *In equilibrium with* $k = 0$,

(*i*) *Sophisticated workers receive either a matching contract with* $m = m^{FB}$ *or a nonelective contribution contract with* $r = \gamma/2$ *and achieve the first best.*

(*ii*) *Exponential workers receive contracts with* $r \leq \gamma/2$, $w = \gamma - r$, *and* $m = 0$ *and achieve the first best.*

(*iii*) *Naïve workers receive contracts with* $m > 0$. *When* $u(c_i)$ *takes CRRA form with coefficient of relative risk aversion equal to* θ, *then:*

(*a*) *if* $\theta < 1$, *then* $m > m^{FB}$;

(*b*) *if* $\theta = 1$, *then* $m = m^{FB}$;

(*c*) *if* $\theta > 1$, *then* $m < m^{FB}$.

Moreover, CRRA utility with $\theta = 1$ *is also necessary for* $m = m^{FB}$ *for all* $\hat{\beta} \in (\beta, 1]$.

All proofs are in the online Appendix.

The results for sophisticated and exponential workers are as described above. The results for naïve workers deserve further discussion. The zero-profit constraint (3)

implicitly defines w as a function of r and m. Denote that function $w(r,m) = \gamma - r - ms(w,r,m,\beta)$. The first-order condition for m can then be written as

$$\left[u'\left(c_2\left(\hat{\beta}\right)\right) - u'\left(c_1\left(\hat{\beta}\right)\right)\right] \frac{\partial}{\partial m}\left[s(\beta)m + s\left(\hat{\beta}\right)\right] + u'\left(c_2\left(\hat{\beta}\right)\right) \frac{\partial}{\partial m}\left[\left(s\left(\hat{\beta}\right)m - s(\beta)m\right)\right] = 0, \tag{7}$$

where we have suppressed the dependence of the savings functions on the contract terms and where $c_1(\hat{\beta}) = w(r,m) - s(\hat{\beta})$ and $c_2(\hat{\beta}) = (1+m)s(\hat{\beta}) + r$ are the worker's anticipated consumption levels.

The first line of (7) represents the commitment motivation for matching. If the worker anticipates his period 1 self saving less than optimally ($c_2(\hat{\beta}) < c_1(\hat{\beta})$), then he desires a larger m to move more consumption to the second period. The first term in brackets represents the strength of that commitment motivation and goes to zero if $1 + m = 1/\hat{\beta}$. Because, for naïve workers $\hat{\beta} > \beta$, this implies that naïve workers' commitment motivation alone is insufficient to lead them to demand a first-best commitment device in which $1 + m^{FB} = 1/\beta$.

The second line of (7) represents the overestimation motivation for matching. In particular, $(\partial/\partial m)\left[(s(\hat{\beta})m - s(\beta)m)\right]$ represents how much the worker anticipates his total compensation will grow as the match increases. To see why, note that $s(\hat{\beta})m$ is the amount of anticipated matching payments. In fact, a match m will generate only $s(\beta)m$ in actual matching payments. Zero profits thus requires that the wage be reduced by $(\partial/\partial m)\, s(\beta)m$ as the match is increased. Note that for $\hat{\beta} = \beta$, this expression is equal to zero—sophisticates have correct expectations and hence no overestimation motivation. In contrast, for naïve workers this term is always positive, since they overestimate their savings under the match. The overestimation motivation can therefore help make up for naïve workers' lack of commitment motivation.

In general, however, the naïve worker's preferred match is not finely calibrated to his need for additional commitment. In the behavioral contract theory literature, naïve present-biased agents generally do not demand first-best commitment contracts (see, e.g., DellaVigna and Malmendier 2004, Heidhues and Kőszegi 2010), and this is also generically true in our setting. Whether equilibrium m overshoots or undershoots m^{FB} for naïve workers depends on their elasticity of intertemporal consumption (EIS), which determines the worker's willingness to tolerate unequal consumption across periods. With CRRA utility, the EIS is equal to $1/\theta$, where θ is the coefficient of relative risk aversion, and for workers with EIS > 1, the equilibrium entails a match above m^{FB} in which the worker anticipates receiving a relatively high total amount of compensation at a cost of backloading his anticipated consumption into the second period. Workers with EIS < 1, in contrast, will undershoot m^{FB}. In the knife-edge case of CRRA utility with EIS $= 1$ (i.e., log utility), workers receive the first-best commitment contract for all levels of naïveté. Moreover, log utility is *necessary* for all naïve types to receive m^{FB}. That is, for every other increasing concave function $u(\cdot)$, there exists a type of naïve worker such that in equilibrium $m \neq m^{FB}$.

Perceptive readers will note that Proposition 1 says nothing about the equilibrium r for naïve workers. In fact, any nonelective contribution r that is less than the naïve worker's equilibrium second-period consumption is consistent with equilibrium. Equilibrium contracts with all such r's deliver the same anticipated and realized consumption streams as workers perfectly offset increased r with lower (actual and anticipated) savings. As negative savings are not permitted, this neutrality breaks down if r is too large, and workers strictly prefer the nonbinding rs to any binding one.

Our analysis of this simplest case of our baseline model illustrates a recurring theme throughout our analysis: equilibrium retirement plans maximize workers' decision utility at the time of contracting, not their experienced utility, which is the appropriate criterion for welfare analysis. For workers subject to biases that imply a disjuncture between their decision utility and experienced utility, equilibrium retirement plans do not maximize their welfare. In the case of naïve present-biased workers with costless opt-out, this disjuncture stems from mistakes workers make in predicting their future savings behavior. The equilibrium plan caters to rather than corrects these mistakes.

B. *Costly Opt out*

Suppose now that $k > 0$ so that defaults are (potentially) sticky. Note that the default d does not appear in the objective function in (2), since we have assumed that the worker believes her self 1 will always make an active choice based on the preference parameter $\hat{\beta}$. Defaults matter in this model only in determining self 1's savings choice as reflected in (5), which in turn affects the zero-profit constraint (3). The following proposition characterizes the equilibrium.

PROPOSITION 2: *In equilibrium with* $k > 0$, *all worker types choose matching contracts with* $m > 0$ *with the default contribution rate* d *that minimizes worker savings, given the other terms of the contract. If* $\beta > u'(\gamma)/u'(0)$, *these contracts have* $d > 0$.

In our model, firms do not use automatic enrollment to paternalistically increase savings as urged in much of the literature in behavioral economics. Rather, in equilibrium, the default is designed to *minimize* worker savings conditional on the other terms of the contract. To see why, suppose there were an equilibrium contract with a different default. Then holding fixed the w, r, and m of that contract, using the default that minimizes savings given those other terms would lower realized matching payments, relaxing the zero-profit condition. But that implies that there exists an alternative contract in the constraint set that delivers higher levels of salient forms of compensation than the supposed equilibrium contract. Key to this result, of course, is our assumption that defaults are not salient at the time of contracting, so that their only substantive effect is through relaxing firms' zero-profit constraint. Similarly, there is no incentive for firms to try to make defaults less sticky by forcing workers to make an active choice as suggested by Carroll et al. (2009). The reason is that doing so would always increase savings under the contract relative to the optimal default.

This dynamic is why the equilibrium contracts always entail a positive match, even for exponential discounters with $\hat{\beta} = \beta = 1$. Because all workers are in effect naïve about their sensitivity to defaults, even otherwise exponential workers overestimate how much they will save in equilibrium, since the defaults are designed to minimize savings. And that overestimation mistake makes matching attractive to workers in much the same way as with naïve present-biased workers under costless opt-out. This mistake results in first-order gains in utility as you increase m from $m = 0$. For exponentials, the effect on intertemporal consumption choices, which are distorted by the match, is second-order. Moreover, for agents with $\hat{\beta} < 1$, increasing m from $m = 0$ *improves* their anticipated intertemporal consumption choices so that they also have the commitment motivation for using m discussed in the costless opt-out case above.

With sufficiently high marginal utility at $c = 0$, the default is strictly positive. To understand why, note that a contract with the worker staying at a zero default will always be dominated by an identical contract with a higher matching rate up to the match that makes the worker indifferent between sticking with a zero default and opting out. Consider then whether it would be attractive to increase the matching rate above this threshold level and raise the default in order to keep the worker from opting out. Doing so requires a reduction in the wage to maintain zero profits, since it increases employer matching payments. But with significantly diminishing marginal utility of consumption, the default needs to be increased and the wage decreased only very slightly. Indeed, with CRRA utility, which has $u'(0) = \infty$, the default need only be increased by a second-order amount.

II. Heterogeneous Types

Consider now the case of heterogeneous worker types in which firms do not observe workers' types. In this section, for tractability, we assume that $u(c_i) = \ln(c_i)$, which enables us to derive simple closed-form expressions for savings functions, which are useful in our proofs. We focus on the tractable but still analytically rich case with two types. A fraction κ^e of workers are exponential discounters with $\beta^e = \hat{\beta}^e = 1$, and a fraction $1 - \kappa^e$ are naïvely present-biased with $\beta^n = \beta < 1$ and $\hat{\beta}^n = 1$. Our assumption that both types have $\hat{\beta} = 1$ means that in terms of period 0 preferences, there is only a single type, which allows us to use the same basic equilibrium concept we used in Section I. This assumption in effect entails assuming that exponential and naïve workers pool but can be relaxed with little substantive change to our results.[3]

In this section, we also make two changes to the contract space. First, for brevity we omit the nonelective contribution, r, since it plays no substantive role

[3] We considered the more general heterogeneous type case in which naïve workers have $\hat{\beta} < 1$ in an earlier draft of this paper and showed that naïve and exponential workers always pool in equilibrium. The reason is that the equilibrium contract has a matching cap set at the naïve workers' anticipated savings level. Naïve workers therefore anticipate receiving *mcw* in matching contributions, which is the maximum possible matching benefit, which makes them (wrongly) prefer the equilibrium pooling contract to any zero-profit separating contract. This produces similar results in terms of contracts and utility to the full naïve case but requires much more complicated analysis and lengthy proofs.

for exponential and naïve workers, as our analysis of homogeneous types above showed.[4] Second, we allow for nonlinear matching contributions by including a cap $c \in [0,1]$ on savings, measured as a fraction of the wage, that are matched.[5] Otherwise, the timing and choices remain as above. The equilibrium is thus the solution to the following constrained optimization problem:

$$\max_{w,m,c,d} u\big(w - s\big(w,m,c|1\big)\big) + u\Big(s\big(w,m,c|1\big) + m\min\big\{s\big(w,m,c|1\big), cw\big\}\Big), \tag{8}$$

subject to

$$w + m\Big[\big(1-\kappa^e\big)\min\big\{s\big(w,m,c,d|\beta\big), cw\big\} + \kappa^e \min\big\{s\big(w,m,c,d|1\big), cw\big\}\Big] = \gamma, \tag{9}$$

$$s\big(w,m,c|b\big) = \arg\max_{s\geq 0} u\big(w-s\big) + bu\big(s + m\min\big\{s, cw\big\}\big), \tag{10}$$

and

$$s\big(w,m,c,d|b\big) \in \arg\max_{s\geq 0} u\big(w-s\big) + bu\big(s + m\min\big\{s, cw\big\}\big) - k\mathbf{1}\big(s \neq dw\big). \tag{11}$$

The key change from the homogeneous type case is in the zero-profit condition (9), which now includes matching on the basis of the weighted average of savings of the two types. For brevity, we focus on the costly opt-out case. The following proposition characterizes the equilibrium.

PROPOSITION 3: *With heterogeneous types and $k > 0$, in equilibrium workers receive contracts such that*

(*i*) *Savings are matched at a rate $m > 0$ up to a cap $c > 0$;*

(*ii*) *The default contribution rate d is the one that minimizes average worker savings in the contract, given the other terms of the contract, and $d < c$; and*

(*iii*) *If the opt-out cost k is sufficiently small, then*

(*a*) *both types anticipate saving to the cap.*

[4] As in that section, workers are indifferent between all levels of nonelective contributions for which the no-borrowing constraints do not bind, and at those levels equilibrium actual and anticipated consumption are independent of nonelective contributions.

[5] We omitted the cap from our homogenous type model in Section I because with homogenous types, the equilibrium cap will always be set at or above both the actual and anticipated savings levels of the worker and hence plays no meaningful role.

(*b*) *In fact, one type saves to the default and one type opts out. For sufficiently small κ^e, exponentials opt out and raise savings to the matching cap. For sufficiently large κ^e, naïve workers opt out and lower savings to below the default.*

(*c*) *Total compensation received by the naïve workers is less than that received by the exponentials.*

The equilibrium is broadly similar to the case with homogeneous types. As before, the equilibrium involves a matching contract and a default that minimizes savings, given the other terms of the contract. But allowing for heterogeneous types produces four key new results.

First, the equilibrium contract entails a default set strictly below the contract's cap on matching. The reason is that doing so results in strictly lower matching payments under the contract than if the default were set at or above the cap, which in turn enables the firm to offer more generous terms on the salient (nondefault) dimensions of compensation and still break even.

Second, if *k* is sufficiently small, the cap on matching is set at workers' anticipated savings level. The reason workers prefer such a cap to an uncapped (or higher-cap) contract is that it reduces the anticipated distortion to workers' intertemporal consumption choices produced by matching. To see the intuition, note that all workers believe in period 0 that they are time-consistent exponential discounters and hence ceteris paribus would prefer to receive compensation as a lump sum wage rather than through a distortionary instrument like a match. Matching is nonetheless attractive here because workers overestimate how much they will save, and hence matching results in workers anticipating receiving compensation greater than their marginal product, γ. The cap on matching lets workers enjoy that anticipated boost to compensation while limiting the anticipated distortion to intertemporal consumption choice. In fact, as we discuss in Section V below, employers' matching contributions almost always include a cap on the match; allowing for heterogeneous types enables our model to explain that fact.

Third, with heterogeneous worker types, the model produces heterogeneity in opt-out decisions, with one type sticking with the default in equilibrium and the other type opting out. In the more natural case with many naïve workers, the savings-minimizing default will be one set below the savings rate naïve workers would actively choose (based on the preference parameter β), and naïve workers would stick with the default while exponentials would opt out to raise their savings to the matching cap. This matches the empirical evidence on defaults (see, e.g., Choi et al. 2004).

Fourth, with heterogeneous worker types, naïve workers cross-subsidize exponential workers. The reason is that naïve workers save less than exponential workers and consequently receive lower matching payments. In equilibrium, naïve workers are paid less than the value of their marginal product of labor, and exponentials are paid more than their marginal product. In this model, retirement plans thus lower the compensation of the naïve present-biased workers who generally undersave.

III. A Paternalistic Model of Retirement Plan Design

In our model, even if some employers would like to act paternalistically, due to intrinsic preferences or otherwise, competition in the labor market leaves no room for such paternalistic motivations to be expressed. Meaningful employer paternalism in retirement plan design requires a concurrence of two factors that is unlikely to be widespread: employers must *both* be paternalistically motivated and have significant market power. Without paternalistic motivations, even a monopsonist would offer contracts of the form described above but with a lower base wage. Without market power, any truly paternalistic zero-profit contract would be rejected by workers. In this section, we nonetheless formalize a model of paternalistic retirement plan design to serve as a useful benchmark against which to assess the empirical predictions of our equilibrium model. We will maintain the behavioral assumptions in the previous section and characterize the set of zero-profit contracts that maximizes workers' average experienced utility. In particular, we define the "paternalistic employer contract set" as the set of contracts that solves the problem

$$\begin{aligned}(12)\quad \max_{w,r,m,c,d}\ & \kappa^e\Big[\ln\big(w-s(1)\big)\\ &\qquad +\ln\big(r+s(1)+m\min\{s(1),cw\}\big)-k\mathbf{1}\big(s(1)\neq dw\big)\Big]\\ &+\big(1-\kappa^e\big)\Big[\ln\big(w-s(\beta)\big)\\ &\qquad +\ln\big(r+s(\beta)+m\min\{s(\beta),cw\}\big)-k\mathbf{1}\big(s(\beta)\neq dw\big)\Big],\end{aligned}$$

subject to

$$(13)\qquad w+r+m\Big[\big(1-\kappa^e\big)\min\{s(\beta),cw\}+\kappa^e\min\{s(1),cw\}\Big]=\gamma$$

and

$$(14)\quad s(b)\in\arg\max_{s\geq 0}\ln\big(w-s\big)+b\ln\big(r+s+m\min\{s,cw\}\big)-k\mathbf{1}\big(s\neq dw\big),$$

where we have suppressed the dependence of the savings functions on the contract terms to economize on notation. Note that unlike in the equilibrium model in which d did not appear directly in the objective function, here we explicitly maximize over d. Hence, the nonsalience of the default to workers no longer matters in this alternative paternalistic model.

In this simple framework, a paternalistic contract can always achieve the utilitarian first best. The most straightforward approach is to use the match to offset the naïve worker's time-inconsistency problem, cap the amount of savings that is matched to keep the exponential from benefiting from cross-subsidization, and default workers to save at the cap. One complication that arises is a multiplicity of contracts that achieves first best. As a selection device, we introduce an arbitrarily small fraction of "active savers" of each type with $k=0$. The introduction of an arbitrarily small share of active savers would have no impact on our equilibrium

model in Section II, since they have insignificant effects on average savings, but it allows us to narrow down the paternalistic set here.

PROPOSITION 4: *All paternalistic employer contracts implement the first-best allocation:*

(*i*) *The paternalistic employer contract set includes one with* $d = c > 0$ *and* $m > 0$ *and one with* $w = r = \gamma/2$.

(*ii*) *With the addition of an arbitrarily small share of workers of each type with* $k = 0$, *all paternalistic employer contracts must have* $d = c > 0$ *and* $m > 0$ *or have* $w = r = \gamma/2$.

There are only two basic types of contracts that achieve the first best when active savers are also in the market: contracts that use r to finance retirement consumption and matching contracts. It is easy to see that $w = r = \gamma/2$ will induce all workers to choose $s = 0$ and achieve the first best. The more interesting case is with matching. All paternalistic employer contracts that use matching must use a cap on matched savings so that both the naïve and exponential active savers save the same amount (at the cap) and must set the default equal to the cap so that workers with $k > 0$ do not have to bear opt-out costs to save the optimal amount.

Note that we have assumed that there is no heterogeneity in the normatively optimal savings rate across workers. Bernheim, Fradkin, and Popov (2015) consider a setting with heterogeneity in optimal savings rates and point out that a matching cap generates a discontinuity in returns to saving that results in a point of accumulation at the cap in workers' optimal savings rates under the contract. Accordingly, they argue that a paternalistic employer in their setting would also generally set the default at the matching cap. The prediction that $d = c$ in a paternalistic contract is thus robust beyond the assumptions considered in our model.

IV. The Pension Protection Act of 2006

The Pension Protection Act of 2006 (PPA) removed regulatory barriers to the adoption of automatic enrollment in employer-sponsored retirement plans. In particular, the PPA shielded employers from fiduciary liability for plans that automatically enroll employees, preempted state wage laws that had prevented employers in some states from using automatic enrollment, and provided a new safe harbor from the nondiscrimination rules for automatic enrollment plans. In this section, we investigate the effects of the PPA on savings outcomes and welfare under the equilibrium model from Section II and the paternalistic model from Section III. In particular, to capture the effect of the PPA on the employers who prior to the PPA were inhibited from using automatic enrollment due to regulatory concerns—or more generally any liberalization of restrictions on automatic enrollment—we contrast the models' predictions under two policies: (i) the case in which employers are prohibited from using automatic enrollment so that d is restricted to 0 exogenously; and (ii) the case considered in the models above in which d is unrestricted (PPA).

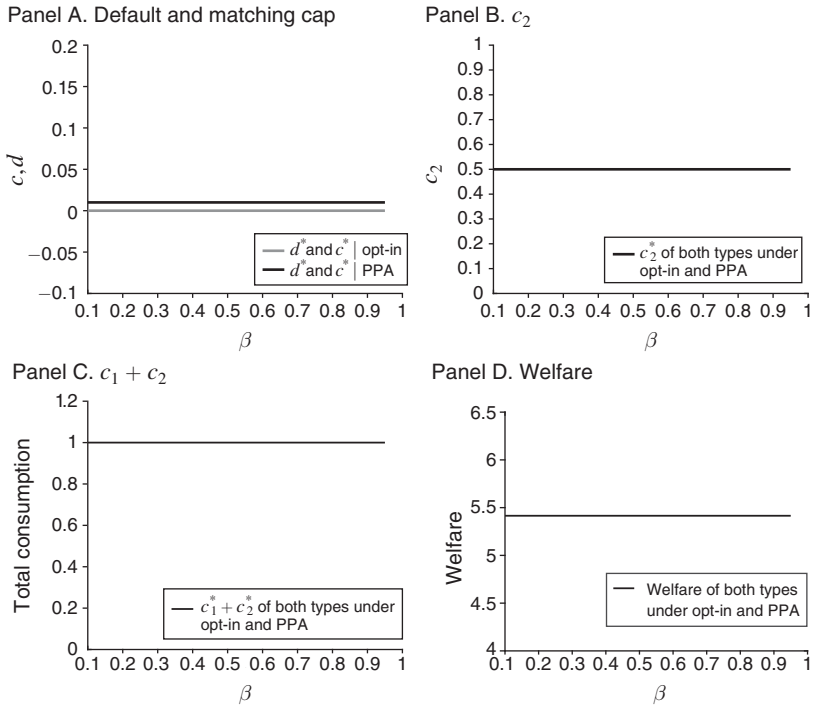

FIGURE 1. PATERNALISTIC PLAN DESIGN MODEL COMPARATIVE STATICS

Notes: Equilibrium contracting terms and outcomes as a function of β for CRRA utility with coefficient of relative risk aversion equal to 0.7. The lighter lines denote outcomes when d is exogenously restricted to 0. The heavier lines denote outcomes when d is unrestricted. The other parameters are $\kappa^e = 0.1$, $k = 0.25$, and $\gamma = 1$.

In the previous two sections we assumed, for tractability, that the utility from consumption took log form. While the resulting model produced what we think are many useful insights, the log functional form assumption has two unattractive properties for the purpose of analyzing the PPA. First, zero consumption in either period results in negative infinity utility, which implies that workers will always opt out of a zero default. This feature is shared by any CRRA utility function with a coefficient of relative risk aversion above one. Second, log utility produces constant consumption shares in each period under the equilibrium model: the average equilibrium consumption across types in each period is always equal to $\gamma/2$ regardless of the default policy. Together, these implications of log utility essentially exclude the possibility that the PPA could improve savings outcomes under the equilibrium model. Accordingly, for our analysis of the PPA in this section, we assume that the utility from consumption takes a more general CRRA form with a coefficient of relative risk aversion less than one. As a result, we must rely on numerical solutions to characterize outcomes.

We begin with the paternalistic model benchmark, for which Figure 1 shows comparative statics on contract parameters and savings outcomes. Given the prevalence of matching among automatic enrollment plans (discussed in Section V), in the figure, we depict the matching contracts under the PPA rather than the *r*-based contracts characterized in Section III. Both under opt-in and under

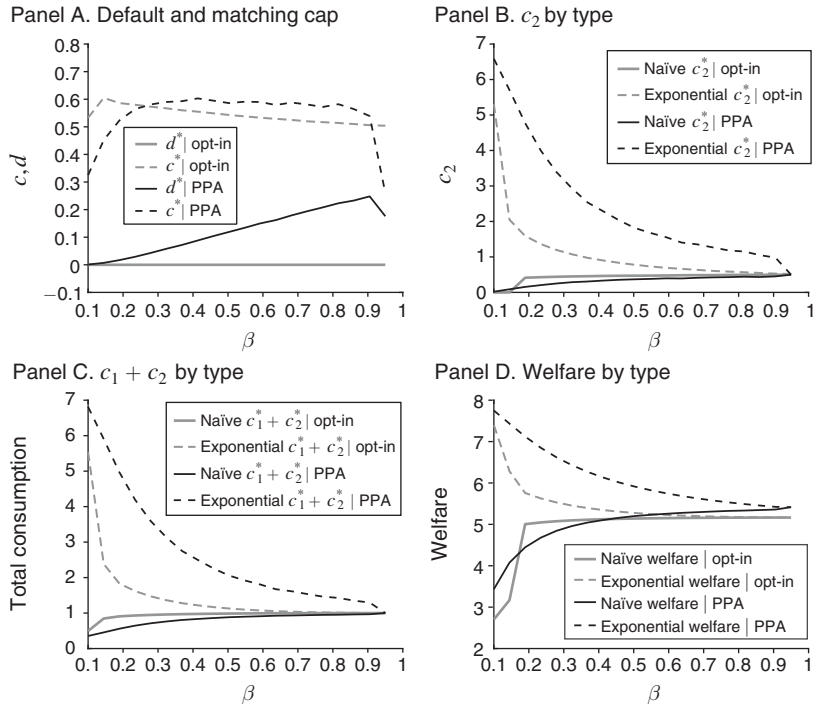

FIGURE 2. EQUILIBRIUM PLAN DESIGN MODEL COMPARATIVE STATICS

Notes: Equilibrium contracting terms and outcomes as a function of β for CRRA utility with coefficient of relative risk aversion equal to 0.7. The lighter lines denote outcomes when d is exogenously restricted to 0. The heavier lines denote outcomes when d is unrestricted. The other parameters are $\kappa^e = 0.1$, $k = 0.25$, and $\gamma = 1$.

the PPA, paternalistic contracts achieve the first-best consumption allocation of $c_1 = c_2 = \gamma/2$ for both types. In the case of the PPA, employers set a default and a match structure such that saving at that default results in first-best consumption and results in both types choosing to stay at that default. Under opt-in, in contrast, employers simply use $r = \gamma/2$ in order to avoid the opt-out cost k. The upshot is that under the paternalistic model, the PPA is not useful in our setting. The basic reason is that there are contractual instruments other than d to improve the savings outcomes of naïve workers, namely employer contributions.

The outcomes under the equilibrium model, shown in Figure 2, are more complicated. When β is very low, naïve workers save more for retirement under the PPA than under opt-in, as shown in panel B. The reason is that for these parameter values, under opt-in they stay at the zero default, whereas under the PPA, naïve workers are automatically enrolled at a small positive d. Note, however, that the PPA results in greater cross-subsidization of exponentials by the naïve workers than under opt-in, as shown in the total consumption levels in panel C. The reason is that under the PPA, employers can offer a higher match than under opt-in without inducing the naïve workers to leave the default. The net effect is still to increase naïve workers' welfare despite the reduction in their total consumption, as shown in panel D.

For higher values of β (greater than about 0.18), however, the PPA *lowers* naïve workers' retirement consumption relative to the opt-in outcome, since naïve workers

would have opted in on their own but are instead defaulted into a lower savings rate and stick with that default. On the other hand, the PPA does allow naïve workers to avoid incurring the opt-out cost k. Which of the offsetting effects is bigger in utility terms depends on the value of β; as β goes up, the utility cost of the distortion to naïve workers' consumption goes down. Thus for moderate values of β, the PPA lowers naïve workers' welfare, but for very high values of β, the PPA increases naïve workers' welfare.[6]

Exponentials earn greater total compensation and second-period consumption under the PPA for all levels of β. The reason is that allowing $d > 0$ enables employers to offer a more generous match without inducing naïve workers to leave the default and thereby results in greater redistribution from naïve workers to exponentials. As a result, the average matching payment as a fraction of the wage is higher under the PPA than under opt-in, despite the fact that the automatic enrollment default is set to minimize savings and matching payments conditional on the equilibrium values of the other terms of the contract.

To summarize, under the equilibrium model, the adoption of automatic enrollment under the PPA can increase retirement consumption of naïve workers with very low β and of exponentials. For naïve workers with very low β, the increase in retirement consumption comes at a cost of lower total consumption but can on net increase their welfare. For exponentials, the result is a steeply sloped consumption path heavily weighted toward the second period, which is inefficient. For more moderately biased naïve workers with higher levels of β, the PPA can actually lower their savings and welfare. Because we used $\kappa^e = 0.1$ in our estimation, average welfare across the two types tracks naïve workers' welfare plotted in panel D of Figure 2 quite closely. The PPA thus raises social welfare for both very high and very low β but lowers social welfare for intermediate β. Under the equilibrium model, then, the PPA poses a set of tradeoffs as a policy matter that are absent under the paternalism model. Our equilibrium model identifies and explains these tradeoffs theoretically.

The key theoretical downside of automatic enrollment we identify, moreover, has been documented empirically: existing studies of the adoption of automatic enrollment show that it lowers the savings rate of many households who would have opted in on their own by anchoring workers at a low default savings rate (Choi et al. 2004). This occurs in our model for naïve workers with relatively high β. The empirical literature also identifies an important benefit of automatic enrollment: it increases participation rates. This occurs in our model for naïve workers with low β. Empirically, these two effects in combination lower the variance in savings rates but may or may not increase average savings. In contrast, in our simple two-type model, adoption of automatic enrollment *increases* the variance in savings rates. But with a richer type space incorporating either (or both) a continuous distribution of β or variation in k, automatic enrollment would lower the variance of savings rates in our model as well.

[6] Whether the opt-out cost k should matter from a policy or welfare perspective, however, is not entirely clear. It should count if we think of it as a real resource cost, but to generate sticky defaults requires a level of k that is very large given the monetary stakes of opting out of the default. An arguably better interpretation of k is as a reduced-form way to capture procrastination and self-control problems that should not count in policymakers' welfare criterion. In that case, the welfare of naïve workers under opt-in would remain above their welfare under the PPA even at high values of β.

V. Evidence

We begin with some key stylized facts about plan design which line up well with the predictions of the equilibrium model. First, the vast majority of defined contribution plans—about 80 percent—offer employer matching contributions with a cap on matched savings (PSCA 2011). Second, failure to receive the full match offered by the employer is indeed widespread, as implied by the equilibrium theory. Choi, Laibson, and Madrian (2011) find that about 50 percent of employees do not save enough to receive their full employer match, foregoing on average 1.3 percent of their salary. Third, most employers that have adopted automatic enrollment plans have chosen relatively low default contribution rates. Indeed, about three-quarters of automatic enrollment plans default workers into a 3 percent initial contribution rate or less (PSCA 2011). Summarizing the empirical literature on automatic enrollment, Choi et al. (2006, 316) write:

> [M]ost employers that have adopted automatic enrollment have chosen very low default contribution rates and very conservative default funds. … *These default choices are inconsistent with the retirement savings goals of most employees* (emphasis added).

Our model helps explain why employers have not chosen the plan design—including the default contribution rate—that would maximize employee welfare.

One complication in interpreting evidence on plan design in terms of the two models is that employer-sponsored plans are subject to a complicated regulatory regime. Among the most consequential of these regulations are the nondiscrimination rules, which in effect require that highly compensated employees' share of the benefits of an employer's plan not be too much greater than that of lower-paid workers. Generally, however, these rules result in incentives for employers to adopt plan features that increase worker savings—that is, to adopt plans that appear to be paternalistic—and thus cannot explain the full cluster of stylized facts described above. Indeed, our equilibrium theory provides a *rationale* for such regulatory requirements. (If employers acted as paternalistic social planners, then there would be little need for these regulations.)

In this section, we provide additional evidence on what we consider to be the most distinctive prediction of our equilibrium theory relative to the paternalistic benchmark: under our equilibrium theory, automatic enrollment plans that offer matching will use a default strictly below the cap on the employer match, whereas under the paternalistic model the default will be set equal to the cap. We test these competing predictions using a novel hand-coded dataset of automatic enrollment plan design.

A. *Data*

Our data come from public filings of Form 5500, which is required of all pension plans covered by ERISA that cover 100 or more employees.[7] Beginning in 2009, the

[7] More specifically, all pension benefit plans with 100 or more "participants" must file Form 5500, where "participants" is defined to include all individuals who are eligible to make contributions to the plan, whether or not

administrative Form 5500 dataset includes a flag identifying whether the plan uses automatic enrollment. The administrative data do not include information about the specific default contribution rate or the structure of employer contributions offered in the plan. That information is generally disclosed, however, in narrative form on Schedule H of employers' original Form 5500s, which are available for download from the Department of Labor's website.

Beginning with all defined contribution single-employer plans that were the employer's *only* pension plan in the plan year, we selected the subpopulation of plans for which the automatic enrollment flag indicated they had adopted automatic enrollment in 2010 or 2011, totaling 3,318 plans. We then selected a random sample of 1,984 plans from this subpopulation and hand-coded information about the default contribution rate and the employer's matching contributions from Schedule H. We coded a plan as having automatic enrollment if any Schedule H filed from 2009 to 2014 for the plan indicated that the plan uses automatic enrollment. In total, 1,276 of the hand-coded plans indicated that they used automatic enrollment in this period. The remaining 708 plans in the sample either erroneously included the automatic enrollment flag on their Form 5500 or simply omitted information about automatic enrollment from their Schedule H. Of the 1,276 hand-coded automatic enrollment plans, 1,213 described the plan's default contribution rate on their Schedule H. Of those, 896 plans provided matching contributions, and of those, 785 described the cap on the plan's matching contributions on their Schedule H and thus form our analysis sample.

B. *Analysis*

Figure 3, panel A, provides a histogram of the initial default contribution rate in our sample. Consistent with other data sources (see, e.g., PSCA 2011), the bulk of plans use a default contribution rate of 3 percent or less. Figure 3, panel B, provides a histogram of the cap on the employer's matching contribution as a percentage of pay. On average, the matching cap is clearly greater than the default contribution rate, with most of the mass above 3 percent.

The key competing predictions between the two models are on the relationship between the default and the cap on employer matching. Figure 3, panel C, provides a histogram of the ratio of the default contribution rate to the matching cap. The distribution is bimodal, with one mode corresponding to the prediction of each of the two models. The majority of plans in the sample—73 percent—have a ratio below 1, with the mode centered around 0.5. The plans in this mode conform to the predictions of the equilibrium model, defaulting workers at a contribution rate strictly less than the minimum they must contribute to receive the full employer match. The second mode, with 24 percent of the sample, is at a ratio of exactly 1, which is the point predicted by the paternalistic model. The balance of plans have a ratio greater than 1.

they actively participate. Plans with fewer than 100 participants can file a Form 5500-SF and are not required to file Schedule H, which is the basis of our hand coding of plan designs.

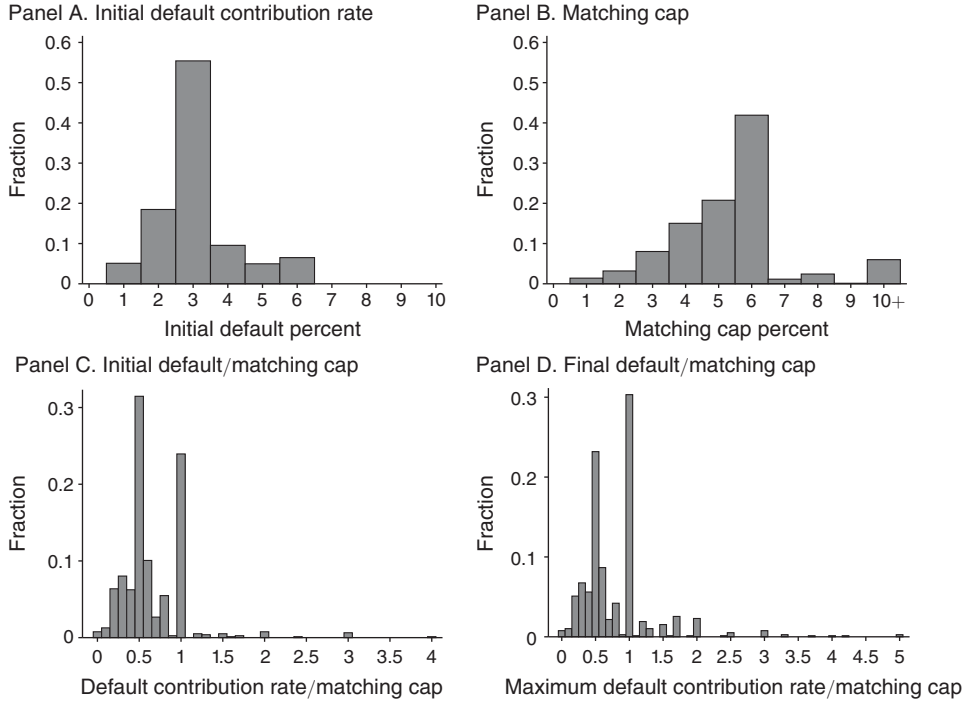

FIGURE 3. DISTRIBUTIONS OF AUTOMATIC-ENROLLMENT PLAN CHARACTERISTICS

Notes: Plan characteristics for a random sample of US firms with at least 100 workers who indicated that they have adopted automatic enrollment to the Department of Labor on Form 5500 in 2010 or 2011 and who disclosed the default rate and cap on the employer-matched savings (if any) in the accompanying financial auditing statement. Observations = 785.

About 20 percent of the sample use automatic escalation, in which participants' contribution rates are incremented by some amount (typically 1 percentage point) each year unless participants opt out, up to some maximum default contribution rate. The majority of these use a specific automatic escalation design—initial default of at least 3 percent, increasing each year by 1 percentage point until the contribution rate reaches 6 percent—that qualifies the plan for a special safe harbor to the nondiscrimination rules created by the Pension Protection Act of 2006.[8] In Figure 3, panel D, we provide the histogram of the ratio of the *maximum* default contribution rate in the plan (i.e., the default at which the automatic increase, if any, plateaus) to the matching cap. The basic results are the same: there are 2 modes, with the bulk of plans having a ratio less than 1 and a substantial minority of plans right at 1.

We interpret these results as showing that there is evidence for both of the two competing theories of employer plan design. On the one hand, the majority of employer plans use low defaults below the plan's cap on employer matching contributions. It is noteworthy that even among the automatic escalation plans, most employers are using the *minimum* default necessary to qualify for a regulatory safe

[8] See Internal Revenue Code §401(k)(13)(C)(iii)(I)–(IV) (2006).

harbor despite being free under the regulations to choose a higher default. These defaults are lower than what a paternalistic employer would choose.

An alternative interpretation is that the emergence of a 3 percent default as an early focal point in automatic enrollment plan design might be the key reason most plans use a default below the cap on matching. But almost half of the automatic enrollment plans in the data use a default other than 3 percent. Moreover, of those, 55 percent use a default strictly less than the cap on matched savings. The phenomenon we document thus cannot be dismissed as an artifact of pooling on 3 percent defaults. Our equilibrium theory also explains the persistence of low 3 percent default contribution rates.

In sum, this evidence suggests that the new theory developed in this paper provides an important and empirically relevant new perspective in understanding plan design. On the other hand, a minority of employer plans use default contribution rates at or above the plan's cap on employer matching contributions, which is consistent with the paternalistic model (and inconsistent with the equilibrium theory in the absence of nondiscrimination rules), suggesting that considerations outside of the equilibrium model are also important determinants of plan design.

VI. Conclusion

Federal retirement savings policy has long been premised on the notion that left to their own devices, households will make mistakes in saving for retirement (Kotlikoff 1987). This paternalistic concern motivates both mandatory savings schemes such as Social Security as well as incentive-based policy tools such as tax subsidies for retirement savings that together shape retirement savings in the United States. The special tax subsidies provided for employer-sponsored retirement savings plans amount to an attempt to harness employers to address this policy problem. In effect, each employer designs a microcosm of the broader federal policy regime through the mix of mandatory savings rules, savings incentives, default rules, and investment options they offer workers in their retirement savings plans.

Previous work in economics has considered the problems raised by mandating or subsidizing certain forms of employer benefits such as pensions and health insurance to achieve public policy goals. Summers (1989) argues, for example, that in the presence of wage rigidities, such policies can distort employment levels, in some cases disproportionately harming the very workers the policy seeks to help. Similarly, the predominance of employer-provided health insurance, due in large part to its tax treatment, can cause an inefficient reduction in labor mobility (job lock) (Gruber 2000).

We identify a new type of dysfunction caused by attempts to use employers to implement social policy. We show that if workers are subject to behavioral biases that affect retirement savings decisions, then employers have incentives to cater to rather than correct those biases. Such biases generally imply a wedge between workers' decision utility at the time of contracting and their experienced utility that is the appropriate criterion for welfare analysis. The equilibrium in the labor market will produce plan designs that maximize the "fictional surplus" measured by workers' ex ante decision utility rather than the true surplus measured by workers'

experienced utility. Our analysis thus calls into question the longstanding delegation to employers of the design of the primary tax-advantaged vehicle for retirement savings. If behavioral economists are right that workers make systematic mistakes in saving for retirement, then the labor market gives employers incentives that undermine the field's "public finance" approach to employer plan design.

While we focus in this paper on the rules that structure contributions to the plan, the same approach can be taken to other aspects of plan design. For example, in an earlier draft of the paper, we consider the set of investment options available within a retirement plan. We show that when employers contract with a mix of exponential workers and naïve diversifiers, in equilibrium each employer's plan offers a set of investment options that includes a low-fee option (for exponentials) and higher-fee options (for naïves) rather than a single price. Consistent with this prediction, Ayres and Curtis (2015) find that more than half of plans in their sample include so-called "dominated funds," defined as options within the plan menu that have an optimal portfolio weight of less than 1 percent and that are more than 50 basis points more expensive than funds in the same style either (i) offered within the plan or, if none, (ii) available in the marketplace. Our approach could also be applied to other forms of employment benefits for which behavioral biases likely play an important role, such as health insurance (Baicker, Mullainathan, and Schwartzstein 2015), but we leave such an analysis for future work.

REFERENCES

Ayres, Ian, and Quinn Curtis. 2015. "Beyond Diversification: The Pervasive Problem of Excessive Fees and 'Dominated Funds' in 401(k) Plans." *Yale Law Journal* 124 (5): 1476–1552.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein. 2015. "Behavioral Hazard in Health Insurance." *Quarterly Journal of Economics* 130 (4): 1623–67.

Bar-Gill, Oren. 2012. *Seduction by Contract: Law, Economics, and Psychology in Consumer Markets*. Oxford, UK: Oxford University Press.

Barr, Michael S., Sendhil Mullainathan, and Eldar Shafir. 2013. "Behaviorally Informed Regulation." In *The Behavioral Foundations of Public Policy*, edited by Eldar Shafir, 440–64. Princeton: Princeton University Press.

Benartzi, Shlomo, and Richard Thaler. 2007. "Heuristics and Biases in Retirement Savings Behavior." *Journal of Economic Perspectives* 21 (3): 81–104.

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 Choi, David Laibson, Brigitte C. Madrian, and Brian Weller. 2010. "Public Policy and Saving for Retirement: The Autosave Features of the Pension Protection Act of 2006." In *Better Living through Economics*, edited by John J. Siegfried, 274–90. Cambridge, MA: Harvard University Press.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2013. "Salience and Consumer Choice." *Journal of Political Economy* 121 (5): 803–43.

Carroll, Gabriel. 2015. "Robustness and Linear Contracts." *American Economic Review* 105 (2): 536–63.

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions." *Quarterly Journal of Economics* 124 (4): 1639–74.

Choi, James J., David Laibson, and Brigitte C. Madrian. 2011. "$100 Bills on the Sidewalk: Suboptimal Investment in 401(k) Plans." *Review of Economics and Statistics* 93 (3): 748–63.

Choi, James J., David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2002. "Defined Contribution Pensions: Plan Rules, Participant Choices, and the Path of Least Resistance." *Tax Policy and the Economy* 16: 67–113.

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*, edited by David A. Wise, 81–125. Cambridge, MA: NBER.

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*, edited by Edward J. McCaffery and Joel Slemrod, 304–51. New York: Russell Sage Foundation.

Choukhmane, Taha. 2019. "Default Options and Retirement Saving Dynamics." https://tahachoukhmane.com/wp-content/uploads/2019/01/Choukhmane_JMP.pdf.

DellaVigna, Stefano, and Ulrike Malmendier. 2004. "Contract Design and Self-Control: Theory and Evidence." *Quarterly Journal of Economics* 119 (2): 353–402.

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying Not to Go to the Gym." *American Economic Review* 96 (3): 694–719.

Edlin, Aaron S., and Chris Shannon. 1998. "Strict Monotonicity in Comparative Statics." *Journal of Economic Theory* 81 (1): 201–19.

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics* 129 (4): 1661–1710.

Gabaix, Xavier, and David Laibson. 2006. "Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets." *Quarterly Journal of Economics* 121 (2): 505–40.

Gottlieb, Daniel, and Kent Smetters. 2016. "Lapse-Based Insurance." https://faculty.wharton.upenn.edu/wp-content/uploads/2016/11/Insurance41.pdf.

Grubb, Michael D. 2009. "Selling to Overconfident Consumers." *American Economic Review* 99 (5): 1770–1807.

Grubb, Michael D., and Matthew Osborne. 2015. "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock." *American Economic Review* 105 (1): 234–71.

Gruber, Jonathan. 2000. "Health Insurance and the Labor Market." In *Handbook of Health Economics*, Vol. 1A, edited by Anthony J. Culyer and Joseph P. Newhouse, 645–706. Amsterdam: North-Holland.

Heidhues, Paul, and Botond Kőszegi. 2010. "Exploiting Naïvete about Self-Control in the Credit Market." *American Economic Review* 100 (5): 2279–2303.

Holmstrom, Bengt, and Paul Milgrom. 1987. "Aggregation and Linearity in the Provision of Intertemporal Incentives." *Econometrica* 55 (2): 303–28.

Kahneman, Daniel. 1973. *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kotlikoff, Laurence J. 1987. "Justifying Public Provision of Social Security." *Journal of Policy Analysis and Management* 6 (4): 674–96.

Madrian, Brigitte C., and Dennis F. Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *Quarterly Journal of Economics* 116 (4): 1149–87.

O'Donoghue, Ted, and Matthew Rabin. 2001. "Choice and Procrastination." *Quarterly Journal of Economics* 116 (1): 121–60.

Orszag, Peter R., J. Mark Iwry, and William G. Gale. 2006. *Aging Gracefully: Ideas to Improve Retirement Security in America*. New York: Century Foundation.

Plan Sponsor Council of America (PSCA). 2011. *54th Annual Survey*. Chicago: Plan Sponsor Council of America.

Schwartzstein, Joshua. 2014. "Selective Attention and Learning." *Journal of the European Economic Association* 12 (6): 1423–52.

Shui, Haiyan, and Lawrence M. Ausubel. 2005. "Time Inconsistency in the Credit Card Market." https://pdfs.semanticscholar.org/a6e3/d841e960666adc8e32a8bc2a3bc4d1446db6.pdf.

Silk, Timothy Guy. 2004. "Examining Purchase and Non-redemption of Mail-In Rebates: The Impact of Offer Variables on Consumers' Subjective and Objective Probability of Redeeming." http://etd.fcla.edu/UF/UFE0004380/silk_t.pdf.

Spiegler, Ran. 2015. "On the Equilibrium Effects of Nudging." *Journal of Legal Studies* 44 (2): 389–416.

Summers, Lawrence H. 1989. "Some Simple Economics of Mandated Benefits." *American Economic Review* 79 (2): 177–83.

Thaler, Richard H., and Shlomo Benartzi. 2004. "Save More Tomorrow™: Using Behavioral Economics to Increase Employee Saving." *Journal of Political Economy* 112 (S1): S164–87.

Vanguard. 2016. "How America Saves 2016." https://web.archive.org/web/20160710101904/https://pressroom.vanguard.com/nonindexed/HAS2016_Final.pdf.