# Should the government be paying investment fees on $3 trillion of tax-deferred retirement assets?*

Mattia Landoni† and Stephen P. Zeldes‡

First version: March 26, 2016
This version: October 18, 2017

Keywords: mutual fund fees, taxes, retirement savings
JEL: D14, G11, G23, G28, H21, J26, J32

*The authors would like to acknowledge the helpful comments of Dan Bergstresser, Kent Daniel, Charles Jones, David Laibson, Brett Myers, Emi Nakamura, seminar participants at the Columbia Business School Finance Free Lunch, the Columbia Macro Lunch, CEIBS, Baruch College, Texas A&M Mays Business School, the Federal Reserve Bank of Chicago, EIEF, the Federal Reserve Bank of Boston, and conference participants at the Red Rock Finance Conference 2016, ESSFM Gerzensee 2017, and the World Finance Conference 2017. We thank Abdullah Al-Sabah for outstanding research assistance. Stephen Zeldes is an external advisor at FeeX.com, and is grateful to the team there for their help understanding and measuring investment management fees.

†Assistant Professor of Finance, Edwin L. Cox School of Business, Southern Methodist University

‡Benjamin M. Rosen Professor of Finance and Economics, Columbia Business School, Columbia University; and NBER

**Abstract**

Governments incentivize retirement saving by allowing individuals to contribute to tax-advantaged accounts where the returns to financial assets receive special tax treatment. In accounts with "back-loaded" taxation, the individual contributes pretax money and pays taxes when the money is withdrawn. In accounts with "front-loaded" taxation, the individual contributes aftertax money and pays no future taxes. Under some simplifying assumptions, a standard benchmark result is that both the individual and the government are indifferent between the two types of accounts. We add investment management fees to the benchmark model and show that the neutrality result breaks down. Assuming fees are fixed as a percent of assets under management (AUM), we show that individuals are still indifferent to the timing of taxation but the government is not. Under back-loaded taxation, the government implicitly owns a share of all retirement accounts and is effectively paying investment fees on this share, something it avoids under front-loaded taxation. We estimate this to cost the government $15 billion per year, representing a subsidy to the asset management industry. We then ask whether this result holds in general equilibrium, where fees as a percent of AUM are allowed to vary. The answer depends both on the nature of the cost function for asset management services, and on the nature of market competition, but we find that the result will in general continue to hold: back-loaded taxation is more expensive for the government and produces a larger asset-management industry. Finally, we use the general equilibrium model to examine welfare implications. In a rough calibration of the model, we find that this increase in the size of the asset management industry reduces consumer welfare.

# 1 Introduction

Retirement savings systems around the world incorporate tax incentives designed to increase saving and enhance retirement security. One dimension along which these incentive schemes vary is the timing of taxation. The traditional way to structure these incentives is through tax deferral—exempting contributions to retirement accounts from current income taxation and then taxing the principal and returns upon withdrawal. This "back-loading" of taxation benefits the investor, because asset returns (interest, dividends, and capital gains) can be earned on the deferred taxes, yielding a higher amount of resources during retirement than would occur in the absence of the tax deferral. Because contributions are exempt, returns are exempt, and withdrawals are taxable, this tax regime is commonly denoted as EET. An alternative system is one in which taxation is "front-loaded", i.e., contributions are made with after-tax income, but then neither the principal nor returns are taxed at any point in the future (TEE).[1,2]

EET has historically been the predominant form of taxation used for most retirement savings systems around the world. However, TEE has gradually become more widely adopted, typically as an additional option. In the U.S., the defined contribution retirement system began as EET with the introduction of employer-based accounts ("401(k)s") and individual retirement accounts ("IRAs"). "Roth" accounts (named after the legislation's original proponent, Sen. William V. Roth) were made available as an alternative to "traditional" accounts—via IRAs in 1997, and via employer-based accounts in 2001. A similar pattern occurred in the U.K. and Canada, who started with EET and later introduced TEE. Although a large majority of assets is still held in traditional EET accounts, the share of TEE assets has been growing. Recently, as part of a broader discussion of U.S. tax reform, policymakers and commentators have begun debating the merits of "Rothification", namely a shift from

[1]This notation is standard in publications by the World Bank (Holzmann and Hinz, 2005) and the OECD (Antolín et al., 2004; OECD, 2015a). Plain taxable savings accounts are denoted as TTE because contributions are taxable, returns are taxable, and withdrawals are exempt. Note that we follow the World Bank and Beshears et al. (2017) in using the terms "front-loaded" and "back-loaded" to refer to the timing of the *taxation*. Income contributed to TEE accounts is taxed upfront, hence the term "front-loaded taxation". Income contributed to EET accounts is deductible upfront and taxed upon withdrawal, hence the term "back-loaded taxation". A source of potential confusion is earlier use of the terms "front-loaded" and "back-loaded" to refer to the timing of the *tax break*. Since the tax break for TEE accounts does not occur upfront, some of those involved in the discussion of the 1997 law that introduced Roth accounts referred to Roth accounts as "back-loaded IRAs" (Committee on Finance of the U.S. Senate, 1997). Several authors including Thaler (1994) and Burman et al. (2001) follow this latter convention as well.

[2]In addition to taxing money flows when contributed to the account or when withdrawn, there is a wide range of other possibilities. For instance, income contributed to the Australian superannuation scheme is usually taxed at two of the three stages (contribution and returns) but at favored rates (OECD, 2015a,b). We focus on EET and TEE because they are the two simplest and most common schemes.

EET to TEE.[3]

Under a few simplifying assumptions, including the constancy of the tax rate across working and retirement years, some basic math shows a strong benchmark neutrality result: EET and TEE accounts yield identical consumption for individuals, both in the working years and in retirement, and individuals are indifferent between the two types of account.[4] Under the additional standard assumption that the discount rate for the government is the same as the expected return on the underlying assets, the net present value of government cash flows is also identical under the two accounts. While EET and TEE are equivalent on a present-value basis, they differ in terms of the timing of cash flows to the government. The front-loaded taxation in TEE accounts generates more government revenue in the individual's working years and less revenue in the retirement years relative to EET accounts.

To facilitate comparison of the two types of accounts, we decompose the EET account into two virtual accounts: i) a TEE account and ii) a separate implicit government account. The government account contains the assets necessary to pay future taxes when the investor takes distributions from the account. The investor is indifferent between owning an actual TEE account and owning a virtual TEE account as part of an EET account. From the government's perspective, whether it collects its revenue now or later is irrelevant to the present-value calculation.

While the benchmark neutrality result is quite general, it does not survive the addition of one crucial bit of realism: record keepers, asset managers, and financial advisors charge fees for running retirement plans, managing assets, and advising clients. These fees are typically charged as a percentage of assets under management (AUM). Assuming these fees are not fully offset by higher performance, they introduce a wedge between the net returns and the government discount rate, breaking the present-value government neutrality result described above. Under back-loaded taxation, the government is effectively paying investment fees on its substantial implicit portfolio, something it avoids under front-loaded taxation.

It is possible, of course, that the additional fees that the government pays on its virtual accounts are compensation for better performance, or other services provided to the government by asset managers. We think this is unlikely for two reasons. First, it is not clear that individuals who pay higher fees on retirement accounts receive any additional benefits as a result. But even if some or all of them do, it is unlikely that the government captures these benefits on its virtual account, because it is implicitly holding a fraction of all retirement portfolios, and many of the potential benefits of higher fees will cancel out

[3]See, for instance, Cumings (2017). For a similar debate in U.K. in 2015, see Osborne (2015).

[4]This benchmark result abstracts from differences that exist across EET and TEE in features such as contribution limits, withdrawal penalties and required minimum distributions. We briefly discuss these features in Section 2.

in the aggregate. For example, some of the higher costs might be associated with creating funds or asset allocations that are customized to a particular group of individuals, such as target date funds that adjust asset allocations as individuals age and get closer to the target retirement date. While this might create value for individuals, holding target date funds of *all* target dates will not create value to the government. Similarly, the government will not benefit from paying higher fees to invest in *all* the funds that focus on style (conservative/aggressive, value/growth), bond maturity (long/short), sector (small cap/large cap, junk/investment grade) or industry. Finally, with regard to equity funds, the government is unlikely to benefit, in aggregate, from active asset management. In the words of Fama and French (2010), "The aggregate portfolio of actively managed U.S. equity mutual funds is close to the market portfolio, but the high costs of active management show up intact as lower returns to investors."

We perform a back-of-the-envelope calculation that assumes that the fees as a percent of AUM remain the same if all EET accounts are converted to TEE. To estimate the size of the U.S. government's implicit account, we multiply the total amount of tax-deferred assets in DC plans and IRAs ($14.4 trillion) by 20%, a reasonable estimate of the average marginal tax rate in retirement, leading to our estimate of $2.9 trillion of retirement assets. Our estimate of assets excludes DB plans, although a parallel argument applies to these plans as well. Including corporate and state and local government DB plans would add $7.1 trillion of tax-deferred money, and thus another $1.4 trillion of an implicit government account. We conservatively estimate asset-weighted fees to be about 80 basis points (bps) based on the lowest asset-weighted estimates available. We assume that 35% of fees paid by the government are recovered via corporate taxation of the asset managers. Multiplying $2.9 trillion by $.80\% \times (1 - .35)$, we reach our estimate of the annual costs of about $15 billion per year. In other words, the government could achieve savings equivalent to $15 billion per year by forcing the conversion of all existing tax-deferred retirement accounts into Roth accounts. This $15 billion, a cost for the government, is an annual subsidy to the asset management industry.

This calculation takes the supply side as given: that is, we rely on the partial equilibrium assumption that investment management fees as a percentage of assets under management are independent of whether retirement accounts are structured as EET or TEE. The extent to which this is true in general equilibrium depends on both the underlying cost structure and the competitive landscape of the asset management industry. If aggregate asset management costs are proportional to aggregate assets under management, then our partial equilibrium result will also hold exactly in general equilibrium.

In order to examine what happens if economies of scale are present, we examine an

extreme alternative by assuming that all asset management costs are fixed and there are no variable costs. Intuitively, these economic assumptions would lead to a monopoly, something obviously at odds with the large observed number of asset management firms. We therefore introduce two additional realistic assumptions. First, consumers are not fully sensitive to the level of fees (see, e.g., Bergstresser et al., 2009; Henderson and Pearson, 2011; Gil-Bazo and Ruiz-Verdú, 2009), and second, entry is free. Specifically, we model competition among funds as spatial competition (Salop, 1979; Tirole, 1988, Ch.7) in a two-period, general equilibrium model in which each fund only needs a fixed amount of labor to operate. A switch from EET to TEE continues to increase assets under management. Despite the assumption that any fund could costlessly expand to manage the additional assets, the model generates increases in (i) the equilibrium number of funds, (ii) employment in the asset management industry, and (iii) the aggregate dollar fees collected.

We next examine whether the larger asset management industry that arises under EET accounts can be optimal. Consider the opposite experiment: starting with a TEE-based system. If TEE were the social optimum, a shift to EET would cause there to be too many funds. Because we are departing from the optimum, however, the total loss would be relatively small, as the social welfare function is flat at the optimum. Only if the equilibrium number of funds increased significantly would social welfare suffer serious consequences. Alternatively, if a TEE-based system started with too few funds, then a shift to EET would get society closer to the optimum, and it would be beneficial. Finally, if a TEE-based system started with too many funds, the shift to EET would be especially pernicious, as it would bring society even farther from the optimum, starting at a point where the social welfare function is already steep. Under a rough calibration, we find that the equilibrium number of funds when all retirement plans are TEE turns out to be about twice as large as a rational planner would set. Since the number of funds in a system is already higher than the social optimum, a shift to an EET system (and the associated higher number of funds) generates a substantial welfare loss. This finding is related to a recent literature on the optimal size of the financial services industry (Philippon and Reshef, 2012; Greenwood and Scharfstein, 2013; Malkiel, 2013; Bolton et al., 2016).

Our results have implications for public policy questions related to retirement saving. The primary question is whether the government should promote a shift towards TEE, and if so how aggressively. Our results point out one advantage of TEE accounts—lower present-value cost to the government. However, before taking a policy stance, one should recognize that there are potentially other important factors not captured by our model, including progressive taxation, behavioral biases, and political economy considerations, that could affect the relative desirability of the two types of account. Our analysis also raises

the question whether the government should act to try to reduce the overall level of fees, either directly, by leveraging the bargaining power arising from its large implicit account, or indirectly, via stricter fiduciary standards for retirement savings accounts. We address these issues in the conclusion.

Our paper is structured as follows. Section 2 derives the basic result that the investor and the government are indifferent between EET and TEE in a benchmark partial-equilibrium model. In section 3, we introduce investment fees, and show that the basic indifference result still holds for the investor, but not for the government. In Section 4 we provide an asset-weighted estimate of total investment fees applicable to retirement accounts. Section 5 examines and calibrates a simple general equilibrium model in which the size of the asset management industry is determined in equilibrium. Section 6 briefly examines the implications for public policy and concludes.

# 2 Benchmark: indifference between front-loaded and back-loaded taxation

In this section we describe the classic result (e.g. Brady, 2012) that, assuming flat taxation and no time variation in the tax rate, optimizing individuals under TEE can and will choose the same consumption allocation (both during work life and during retirement) as they would under EET. In addition, the present value of government revenue is identical under TEE and EET. Together, these imply the economic equilibrium is the same under TEE and EET.

## 2.1 Base assumptions and notation

To begin, we assume a model with no uncertainty. Income tax rates are flat, i.e., they do not vary with the level of income. However, they may differ across the life cycle and across different types of income:[5]

- labor income during the working years is taxed at a rate $\tau_L$;
- retirement income (including principal and returns from EET retirement accounts) is taxed at a rate $\tau_R$;
- investment income is taxed at a rate $\tau_I^i$ that varies depending on the type of account $i$.

[5]In practice, the tax system does not have flat rates, but is instead progressive, with marginal tax rates increasing with income. When coupled with uncertain labor income or asset returns, marginal tax rates become stochastic, introducing certain complications into the analysis. We briefly address these complications in the conclusion.

| Account type $i$ | Type of taxation | Tax on initial contribution | Tax rate on investment returns $\tau^i$ | Tax on retirement payouts |
|---|---|---|---|---|
| Taxable (TTE) | Immediate | $\tau_L$ | $\tau_I^{TTE} > 0$ | 0 |
| EET | Deferred | 0 | $\tau_I^{EET} = 0$ | $\tau_R$ |
| TEE | Immediate | $\tau_L$ | $\tau_I^{TEE} = 0$ | 0 |

Table 1: **Different tax treatment of retirement savings.** Money earned and saved for retirement can be taxed at three points: when earned, when it earns returns on investment, and when paid out of the account in retirement. Each type of account is represented by a three-letter abbreviation. For instance, a common taxable account is "TTE" because earned income is taxable, investment returns are taxable, but account distributions in retirement are exempt.

Table 1 summarizes three possible way of taxing retirement savings:

- *Taxable account*: all labor income is taxed at rate $\tau_L$ when earned. Intermediate investment returns are taxed at a rate $\tau_I^{TTE} > 0$. This is referred to as TTE, because the earned income is taxable, investment returns are taxed, and account distributions in retirement are exempt.

- *EET retirement account*: income tax on retirement account contributions is deferred until the time of retirement $T$, when the account is assumed to be liquidated and all the money is paid out as retirement income, taxed at a rate $\tau_R$. Intermediate investment returns are not taxed, i.e., $\tau_I^{EET} = 0$. This scheme is referred to as EET because the earned income put into the account is exempt, the returns are exempt, and the full amount of the retirement account is taxed on withdrawal.

- *TEE retirement account*: all labor income is taxed at rate $\tau_L$ when earned. Intermediate investment returns are not taxed, i.e., $\tau_I^{TEE} = 0$. This scheme is referred to as TEE because the earned income is taxable, the returns on investment are exempt, and account distributions in retirement are exempt.

We assume that money in the account is invested in the only one asset in positive supply, government bonds, paying a known return of $r$. We abstract for now from details such as contribution limits, withdrawal penalties, and required minimum distributions.

## 2.2 Basic neutrality result: investor final wealth and present value of government revenue

We begin by assuming that individuals' initial consumption and the interest rate they earn on the assets in their retirement account are the same under TEE and EET. Later we will verify that these assumptions hold in the resulting equilibrium. We also assume that the individual has the same pretax labor earnings under either system, and there is no contribution cap regardless of account type.[6] Together, these assumptions imply that the individual directs the same amount of pretax labor earnings to account contributions. We call this amount $S$.

Table 2 shows the initial and future cash flows for both the individual and the government. With an EET account, the government has no revenue upfront, and the individual's account balance is $S$. At time $T$, when the individual retires and the account is liquidated, the balance $(S \cdot e^{rT})$ is paid out and taxed. The individual receives $S \cdot e^{rT}(1-\tau_R)$ and the government receives $S \cdot e^{rT} \cdot \tau_R$. Conversely, with a TEE account, the government taxes the money upfront receiving $S \cdot \tau_L$. The individual's starting balance is thus $S(1-\tau_L)$. No additional taxation happens, and therefore at time $T$ the individual can keep the entire balance $S(1-\tau_L) \cdot e^{rT}$.

It is immediate to see that if $\tau_R = \tau_L$, the individual's ending wealth is the same under both account types, and therefore a consumption plan that is feasible under EET is also feasible under TEE, and vice versa. Moreover, we assumed that initial wealth (i.e., pretax labor earnings) is the same under both accounts, and that the only price in the economy, the interest rate, is also the same. With constant wealth and constant prices, the individual would choose the same consumption plan under EET as she would under TEE, i.e., the same consumption plan is optimal.[7]

The government's cash flow differs across plans—with TEE accounts revenue $S \cdot \tau_L$ is received up front, whereas with EET accounts the revenue is deferred until the future, although it is larger $(S \cdot e^{rT} \cdot \tau_L)$. But assuming that the government discount rate is equal to the interest rate on government bonds, the time-0 present value of revenue under EET is

[6] Our analysis makes a few simplifying assumptions for clarity of exposition. Some are immaterial, such as the assumption that the account is liquidated in a lump sum at time $T$, instead of gradually to provide retirement income. Other simplifying assumptions let us abstract from features that may make TEE more attractive than EET from the individual's perspective. However, these features result in a correspondingly higher cost for the government, and therefore they are immaterial for the purposes of our argument. For example, in the U.S., Roth (TEE) accounts have the same nominal contribution limits as Traditional (EET) accounts (Burman et al., 2001), allowing individuals to contribute a larger amount of their after-tax income. Roth accounts also have fewer restrictions on withdrawals, as contributed principal can be withdrawn penalty-free at any time, and there are no required minimum distributions as long as the account owner is alive.

[7] Even though the EET equilibrium is the same as the TEE equilibrium in a neoclassical sense, behavioral biases elicited by the different choice architecture may cause individuals to choose a suboptimal consumption plan under one or the other account. We discuss some of these factors in the conclusion.

| Account | Individual | | | Government | | |
|---|---|---|---|---|---|---|
| | Initial balance | Future balance | Final payout | Initial revenue | Future revenue | PV @ $r$ |
| EET | $S$ | $S \cdot e^{rT}$ | $S \cdot e^{rT}(1-\tau_R)$ | 0 | $S \cdot e^{rT} \cdot \tau_R$ | $S \cdot \tau_R$ |
| TEE | $S(1-\tau_L)$ | $S(1-\tau_L)e^{rT}$ | $S(1-\tau_L) \cdot e^{rT}$ | $S \cdot \tau_L$ | 0 | $S \cdot \tau_L$ |
| EET - TEE | $S \cdot \tau_L$ | $S \cdot \tau_L \cdot e^{rT}$ | $S \cdot e^{rT}(\tau_L - \tau_R)$ | $-S \cdot \tau_L$ | $S \cdot e^{rT} \cdot \tau_R$ | $-S(\tau_L - \tau_R)$ |
| If $\tau_R = \tau_L$ | | | 0 | | | 0 |

Table 2: **Benchmark cash flows under EET and TEE.** With flat taxes, and assuming that the tax rate on labor income ($\tau_L$) is the same as the tax rate on retirement income ($\tau_R$), the individual has the same retirement wealth both with an EET and a TEE account. Government revenue is also constant in present value, assuming that the government's discount rate is the same as the return on government debt ($r$).

$e^{-rT} \cdot S \cdot e^{rT} \cdot \tau_L = S \cdot \tau_L$, i.e., the same as the immediate revenue under TEE. The government will therefore be indifferent (in a present value sense) between the accounts.

Up until now we have assumed that interest rates are the same under the two systems, but we now show the equilibrium result that a shift from TEE to EET does not affect equilibrium interest rates, because the sum of desired private saving and government saving is constant. Under TEE, the account balance is $S(1-\tau_L)$. Under EET, the account balance is $S$. Since the account is invested in government bonds, this creates additional demand for government bonds equal to $S \cdot \tau_L$. At the same time, the government faces a revenue shortfall (relative to TEE) of $S \cdot \tau_L$. Assuming for simplicity that government expenditure is exogenous, the government must issue an amount $S \cdot \tau_L$ of new bonds, adding to the existing supply. Thus, the increase in desired private saving is exactly offset by the decrease in government saving, and the equilibrium interest rate will remain unchanged.

If $\tau_L \neq \tau_R$, a consumption plan that is optimal under EET would no longer be optimal under TEE. However, if we continue to assume that time-0 consumption is the same across the two types of account, the balance in an EET account at any time $t \in [0, T]$ can be decomposed into three virtual accounts as follows:

$$V_t^{EET} = S \cdot e^{rt} \left[ \underbrace{(1-\tau_L)}_{\substack{\text{Virtual} \\ \text{TEE}}} + \underbrace{(\tau_L - \tau_R)}_{\substack{\text{Transfer} \\ \text{Account}}} + \underbrace{\tau_R}_{\substack{\text{Government} \\ \text{Account}}} \right]. \tag{1}$$

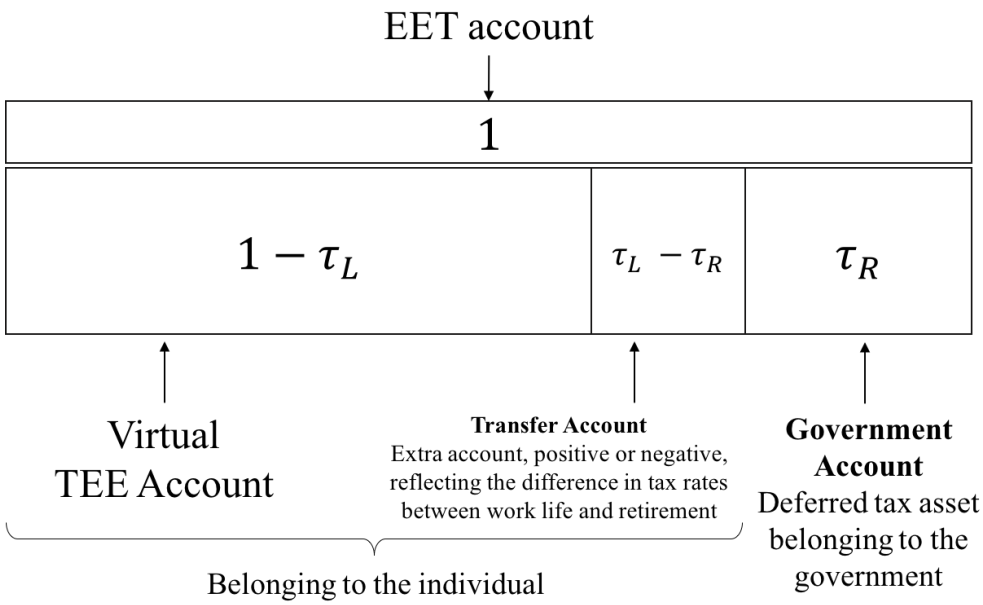


Figure 1: **Decomposition of EET tax-deferred account into three virtual accounts.** Under some simplifying assumptions, the savings in an EET account can be decomposed into a virtual TEE account (the money the investor would have, had she contributed the money to a TEE account); a transfer account (the investor's difference in final wealth reflecting the difference in tax rates between work life ($\tau_L$) and retirement ($\tau_R$); if the $\tau_R > \tau_L$ this account is negative); and a government account (assets necessary to pay future taxes when the investor takes distributions from the account).

This decomposition is depicted in Fig. 1, where the EET account size is normalized to one. The first term is a "virtual TEE" account of size $1 - \tau_L$, belonging to the individual. This account has exactly the same size as the entire account would have under a TEE system. The second term is a "transfer account" of size $\tau_L - \tau_R$, which constitutes the true difference in final wealth between EET and TEE. This account represents a future transfer from the government to the individual (or vice versa, if $\tau_R > \tau_L$). This transfer is not a consequence of inherent differences between EET and TEE, but rather of the difference in tax rates between labor income and retirement income, a separate policy choice. Finally, the third term is a "government account" account of size $\tau_R$. This account represents assets necessary to pay future taxes when the investor takes distributions from the account, i.e., a deferred tax asset belonging to the government. Throughout the paper, we refer to this term as the government's virtual account or implicit account.

## 2.3 Extended neutrality result: adding a risky asset

Now suppose that there are two assets: the government bond yielding $r$ and a risky asset (stocks) with expected return $r_s$. To show the benchmark neutrality result in the presence of stocks, let us assume first that all retirement accounts are held in bonds, as in the basic scenario. Now allow investors to switch from holding bonds to holding a percentage $\alpha$ of their retirement accounts in stocks. From the investors' perspective, if $\tau_L = \tau_R$, it is easy to see that the TEE account would yield an identical outcome to the EET account, as long as investors' expected return for stocks remains $r_s$ under EET and TEE.

Looking at the present value of cash flows, we can also derive the indifference result for the government. Under EET, the *expected* future tax revenue is $S\left[\alpha \cdot e^{r_s T} + (1-\alpha) \cdot e^{rT}\right] \cdot \tau_R$, which is more than in the basic scenario. However, the *actual realized* future tax revenue is riskier than in the benchmark case, because it is proportional to realized stock returns. If the government discounts risky cash flows at a rate $r_s$, the present value of government tax revenue is $S \cdot \tau_R$. Under TEE, the immediate tax revenue is $S \cdot \tau_L$. Thus, if $\tau_L = \tau_R$, the present value of tax revenue is unaffected by the account type.

However, it is not obvious that investors' expected return remains $r_s$ regardless of account type, or that the government's discount rate for risky cash flows should be $r_s$. If investors held a percentage $\alpha$ of their retirement accounts in stocks, the demand for risky assets would be greater under EET than under TEE, because a percentage $\alpha$ of the government account would also be held in stocks. In addition, under EET the government would hold an unhedged risky position, because it has a future claim on risky assets.

If the government is unconstrained in its stock holdings, i.e., if there is nothing preventing it from holding either more or less stocks than it desires, then the benchmark neutrality result would continue to hold. The government could take direct action and adjust its stock holdings. For instance, if under EET its implicit stock holdings are too large, it could hold fewer stocks in an existing portfolio, or short stocks if it doesn't have such a portfolio. In this case, aggregate demand for stocks under EET and TEE would be the same, and so a shift from one to the other would leave equilibrium expected returns unchanged.[8]

In practice, the government may be constrained in its holdings of stocks. For example, if the government would like to hold more stocks, having EET accounts would ease the binding

[8]Even if the government is constrained, the benchmark neutrality result may still continue to hold because of a "Ricardian" equivalence. If, under EET, the government is unable to reduce its stock position, perfectly forward-looking individuals may realize that the government will have to change taxes in the future depending on realized stock returns up to then. Under some assumptions about how the government adjusts taxes in response to stock market returns, this amounts to an implicit long position in stocks because taxes will be lower if stocks have high returns, and vice versa. To offset this implicit position, individuals would reduce their own stock holdings today. Thus, demand and supply of stocks and bonds would remain in balance with no change in interest rates, stock expected returns, or household consumption.

constraint and improve welfare. On the other hand, the government could already have too large an exposure to the stock market (e.g. because future tax revenues are tied directly or indirectly to future stock market performance), and be constrained from reducing its exposure. In this case, having EET accounts worsens the binding constraint by forcing the government to hold even more stocks than it otherwise would under TEE. The arguments here parallel those in the literature on the costs and benefits of the Social Security Trust Fund holding equities (see, e.g. Geanakoplos et al., 1999; Abel, 2001; Diamond and Geanakoplos, 2003).

# 3 The effect of asset management fees

Table 3 extends Table 2 by adding asset management fees. At time $t$, an asset management firm levies fees equal to a fixed proportion of account size, $f$. As in the basic case, we assume that individuals' initial pretax labor earnings, initial consumption, and the effective interest rate they obtain on the assets in their retirement account are the same under TEE and EET. Later we will see whether these assumptions hold in the resulting equilibrium. We begin by assuming $\tau_R = \tau_L$, a necessary condition for the benchmark neutrality result.

Under these assumptions, the individual's final retirement wealth is lower, but still the same across EET and TEE. Both accounts grow at a net-of-fee rate of $r - f$. The left panel of Table 3 calculates the final payouts for the individual. Under EET, the initial balance is $S$, and the final aftertax distribution from the account is $S \cdot e^{(r-f)T}(1-\tau_R) = S \cdot e^{(r-f)T}(1-\tau_L)$. Under TEE, the initial balance is $S(1-\tau_L)$, and the final distribution from the account is $S(1-\tau_L) \cdot e^{(r-f)T}$. Thus, final wealth is the same, and therefore a consumption plan that is feasible under EET is also feasible under TEE. Moreover, with constant wealth and constant prices, a consumption plan that is optimal under EET is also optimal under TEE.

The right panel of Table 3 calculates the present value of tax revenue for the government with fees. To begin, we assume for simplicity that the government does not tax the asset manager's income. Clearly, the stream of tax revenue cash flows is different between EET and TEE. Unlike in the benchmark case, the present value of these cash flows is also different. Under the assumption that $\tau_R = \tau_L$, the individual is still indifferent because a fraction $\left(1 - e^{-fT}\right)$ of her final wealth is equally eroded by fees regardless of account type. On the other hand, the government has unambiguously lower present value of tax revenue under EET:

$$\textbf{PV(Tax Revenue}^{EET}\textbf{)} - \textbf{PV(Tax Revenue}^{TEE}\textbf{)} = -S \cdot \tau_L \left(1 - e^{-fT}\right) < 0. \qquad (2)$$

| Account | Individual | | | Government | | |
|---|---|---|---|---|---|---|
| | Initial balance | Future balance | Final payout | Initial revenue | Future revenue | PV @ $r$ |
| EET | $S$ | $S \cdot e^{(r-f)T}$ | $S \cdot e^{(r-f)T}(1-\tau_R)$ | 0 | $S \cdot e^{(r-f)T} \cdot \tau_R$ | $S \cdot \tau_R$ |
| TEE | $S(1-\tau_L)$ | $S(1-\tau_L) \cdot e^{(r-f)T}$ | $S(1-\tau_L) \cdot e^{(r-f)T}$ | $S \cdot \tau_L$ | 0 | $S \cdot \tau_L$ |
| EET - TEE | $S \cdot \tau_L$ | $S \cdot \tau_L \cdot e^{(r-f)T}$ | $S \cdot e^{(r-f)T}(\tau_L - \tau_R)$ | $-S \cdot \tau_L$ | $S \cdot e^{(r-f)T} \cdot \tau_R$ | $-S\left(\tau_L + -e^{-fT}\tau_R\right)$ |
| If $\tau_R = \tau_L$ | | | 0 | | | $-S \cdot \tau_L\left(1-e^{-fT}\right)$ |

Table 3: **Present value of tax revenue under EET and TEE with fees and no corporate taxes.** An asset manager charges proportional fees $f$ on the account. Assuming that the tax rate on labor income ($\tau_L$) is the same as the tax rate on retirement income ($\tau_R$), the individual has the same retirement wealth both with an EET and a TEE account. However, government revenue is lower with EET, assuming that the government's discount rate is the same as the return on government debt ($r$).

This formula has an intuitive interpretation: $S \cdot \tau_L$ is the initial size of the government's virtual account under EET, and $\left(1-e^{-fT}\right)$ is the fraction of the account that gets eroded by fees.

This result implies that, generally speaking, in the resulting equilibrium the interest rate and the individual's final wealth need *not* remain the same under EET and TEE. For instance, suppose government expenditure follows an exogenous path. To cover the loss of revenue under EET, the government would have to raise tax rates, reducing the individuals' wealth. Thus, a consumption plan that is feasible and optimal under TEE may not be feasible under EET. For the sake of exposition, we do not solve for the equilibrium consumption plan, and instead we continue to assume that the individuals' initial consumption stays constant.

Next, assume that the government levies a corporate tax $\tau_C$ on the asset manager's profits. For simplicity, assume that under EET every additional dollar of fee revenue equals profit for the asset manager.[9] Now the government has not only the initial and final revenue,

[9]If the additional assets under EET result in additional costs, only a fraction of the revenue equals taxable profits directly for the asset manager. However, most additional costs would equal income indirectly for employees or other entities upstream in the supply chain. The overall result would be that a different fraction of asset manager revenue is recovered. For instance, in the U.S., corporate profits are taxed at a top marginal rate of 35%. At one extreme, the firm could use all the extra revenue to fund an additional project done by a freelancer facing a marginal tax rate of 43.3% (personal tax of 28% plus self-employment tax of 15.3%). At the other extreme, the firm's costs could be "royalties" for "intellectual property" paid to

but also a stream of corporate tax revenues that grows at the same rate as the account balance. The algebra is slightly more convoluted, but the end result is still amenable to an intuitive interpretation:[10]

$$\mathbf{PV(Tax\ Revenue}^{EET}) - \mathbf{PV(Tax\ Revenue}^{TEE}) =$$

$$= -S\tau_L \cdot \left(1 - e^{-fT}\right) \cdot (1 - \tau_C) < 0, \tag{3}$$

where the first two terms are the same as in the case with no corporate tax, and $(1 - \tau_C)$ is the fraction of fee income that is *not* recaptured by the government via taxation of the asset manager. From this expression, it is evident that taxing the income of asset management firms can only ameliorate, but not eliminate the difference in present value of government revenue between EET and TEE accounts.

If $\tau_R \neq \tau_L$, the individual is not indifferent between EET and TEE, just as in the benchmark case. The EET account can be still decomposed into three virtual accounts, as shown in Figure 2: a TEE equivalent, a transfer account, and a deferred tax asset belonging to the government. However, the existence of a virtual transfer account due to a difference in the tax treatment of labor income and retirement income does not create any additional inefficiency. The inefficiency is created by the government's leaving an amount $S \cdot \tau_L$ in the account at time 0. How this amount is ultimately split between the government and the individual does not matter. At time $T$, the individual simply receives an additional $S \cdot e^{-fT}(\tau_L - \tau_R)$, and the government's tax revenue drops by the same amount:

$$\mathbf{PV(Tax\ Revenue}^{EET}) - \mathbf{PV(Tax\ Revenue}^{TEE}) =$$

$$= -S \cdot e^{-fT}(\tau_L - \tau_R) - S \cdot \tau_L \left(1 - e^{-fT}\right)(1 - \tau_C). \tag{4}$$

Summarizing, if $\tau_R = \tau_L$, the investor is still indifferent between EET and TEE; if $\tau_R \neq \tau_L$, the individual's relative preference for one or the other account does not change vis-à-vis the case with no fees, because the sign of the transfer depends only on $\tau_L - \tau_R$ regardless of fees. For the individual, both EET and TEE are eroded by fees in equal proportions. For the

a shell corporation situated in a non-U.S. jurisdiction facing a 0% corporate tax.

[10] For an EET account, the present value of corporate tax revenues is equal to $S \cdot f \cdot \tau_C \cdot A(r, r - f, T)$, i.e. the initial account balance, times the percentage fees $f$ (to obtain the asset manager's instantaneous revenue flow) times the corporate tax rate $\tau_C$ (to obtain the government's instantaneous corporate tax revenue flow) times a growing annuity term $A(r, r - f, T)$. For a given growth rate $g$, $A(r, g, T) = \left[1 - e^{-(r-g)T}\right] / (r - g)$ is the present value of a unit flow growing at a rate $g$ until time $T$ discounted at rate $r$. Similarly, the present value of corporate tax revenues for a TEE is $S \cdot (1 - \tau_L) \cdot f \cdot \tau_C \cdot A(r, r - f, T)$, and therefore an EET yields an additional $S \cdot \tau_L \cdot f \cdot \tau_C \cdot A(r, r - f, T)$ in corporate tax revenues. Adding this term to (2), we obtain (3).

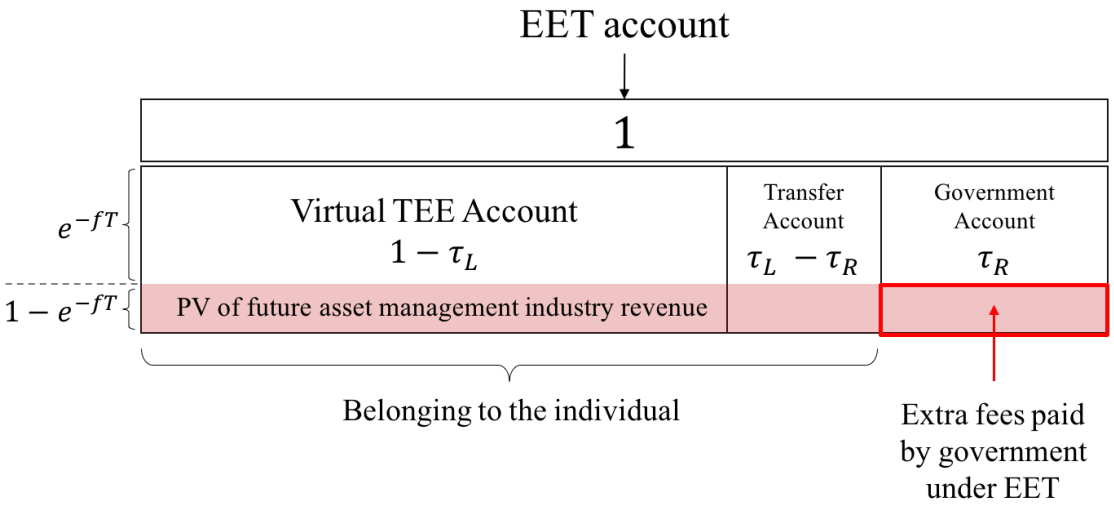


Figure 2: **Decomposition of EET tax-deferred account into three virtual accounts with asset management fees.** The decomposition is the same as in Fig. 1. All three virtual accounts (including the government's deferred tax asset) are eroded by asset management fees equally.

government, however, things are different. With a TEE account, the government receives the tax revenue upfront. With an EET account, the government receives the tax revenue when the money is paid out. As a consequence, a fraction $\tau_R$ of the EET account actually constitutes a virtual account owned by the government.

If the government were to shift from EET to tee it would receive the revenue upfront and pay down some debt. By leaving the money in the account, the government keeps paying an interest rate $r$ on the outstanding debt, but receives a net-of-fees return $r - f$ . In other words, the government keeps money in a *virtual* account, which pays *real* fees. The fees are real because we assume that the same percentage fees are charged on a larger account size.

Two natural questions arise. First, how important are fees? Second, would fees really remain the same if the size of accounts varied? We address these questions in the next two sections.

# 4 Investment fees in retirement accounts

An individual saving for retirement faces at least four types of costs: account fees, advisory fees, asset management fees, and trading costs. Account fees cover the cost of account maintenance. Advisory fees and "wrap" fees cover financial advice that comes with the account. Asset management fees are charged based on what financial products the account money is invested in. These fees cover the operating costs and profits of mutual funds

sponsors and other providers. Finally, trading costs are incurred while buying and selling securities, either by the individual, or by a fund owned by the individual.

In the U.S., tax-deferred retirement accounts include two main components: employer-sponsored defined contribution retirement accounts such as 401(k) and 403(b) plans (DC plans), and individual retirement accounts (IRAs).[11] With roughly $8 trillion of assets each, DC plans and IRAs are both important components of overall retirement assets. Most of the assets in IRAs were initially accumulated in DC plans and later "rolled over" to an individual account.

## 4.1 Defined contribution plans

Assets in DC plans are invested in a menu of investment products chosen by the employer. The menu typically includes mutual funds and other pooled investment products. These products include funds that are not available to the general public, such as collective investment trusts (CITs), and stable value products, such as guaranteed investment certificates (GICs). Henceforth, we refer to collective investment products as "funds". In private-sector DC plans, the menu may include employer stock.[12] Individual participants choose how to allocate contributions across funds in the menu. Individuals may also shift existing balances across funds.

In DC plans, explicit account fees are usually assessed as a function of assets or number of participants. These fees cover account management costs such as recordkeeping and regulatory compliance. On average, roughly one-tenth of these fees is covered by the employer, and the rest is charged directly to the participants (Rosshirt et al., 2014). In what follows, we include employer-paid fees as part of total fees, and we exclude fees for voluntary additional services such as robo-advisors or premium human advisors.

Asset management fees are incurred as a function of the chosen investments. Asset management fees are difficult to separate from account fees, because they often contain recurring, distribution-related fees that are rebated by the asset manager to the account provider. These fees include so-called 12b-1 fees, sub-transfer agent fees, and shareholder servicing fees.

For the purpose of assessing total fees paid by the government on its implicit account, we need an asset-weighted estimate of total fees in DC plans. Asset-weighted estimates are typically substantially lower than equal-weighted estimates because low-fee funds attract

[11] This section focuses on the U.S. retirement landscape, but most of the discussion applies to any country that has employer-based or individual retirement accounts.

[12] In some plans, participants are able to access a brokerage window to buy and sell individual securities outside of the employer-chosen menu, but brokerage assets make up only a tiny fraction of total assets.

more customers, and therefore have larger assets under management (Hubbard et al., 2010). Two recent industry publications have attempted to provide an asset-weighted estimate of total fees for 401(k) accounts, a large fraction of overall DC plans. Deloitte (Rosshirt et al., 2014) estimates the "all-in fee" for 401(k) accounts at 58 bps of assets under management. BrightScope (Brightscope and Investment Company Institute, 2014) estimates "total plan costs" at 39 bps. Both estimates are done in partnership with the industry trade association, the Investment Company Institute. The Deloitte estimate is survey-based, providing a less precise but more representative estimate. The survey excludes plans with less than $1 million in assets and oversamples large plans, claiming representation of roughly 97% of the universe of plans filing Form 5500 with the Department of Labor. The BrightScope estimate is based on filings by audited plans, which generally means plans with 100 or more participants. As a consequence, the BrightScope study excludes about $1 trillion or 27% of total assets held in the smallest, and likely most expensive, plans. We arrive at an estimate of 50 bps for total account costs by simply doing a rough average of these two estimates.[13]

## 4.2 Individual retirement accounts

Individual retirement accounts (IRAs) are characterized by greater flexibility, as individuals may open an IRA with any financial institution. Assets in IRAs are invested in a broad set of investment products made available by the financial institution. This set is usually much broader than the one available to DC plan participants, and it includes publicly available funds, products that are unique to the account provider, and individual securities. Financial institutions offering IRAs include brokerage firms, mutual fund sponsors, banks, and insurance companies.

Based on industry sources (ICI, 2015; Collins et al., 2016; Copeland, 2016), mutual funds make up slightly more than half of IRA holdings (26% equity, 8% bonds, 11% balanced, and 10% money market funds). Another 36% consists of individual stocks and bonds; the remainder is invested in other products. We use these weights, together with information on industry-level expense ratios, to estimate an average asset-weighted expense ratio for all IRA assets. Collins et al. (2016) indicate industry-average expense ratios of 68, 54, 77 and 14 bps respectively for equity, bond, balanced and money market funds. These figures include "sales loads", which are one-time fees incurred upon the purchase or sale of mutual fund shares and structured products. Sales loads typically benefit the account manager and not the mutual fund or other product whose shares are being bought or sold. Individual

[13]These estimates imply that fees on funds held in DC plans are substantially lower than the industry average of all mutual funds. However, there may still be scope for further reduction, as plan managers appear to choose suboptimal menus of funds (Ayres and Curtis, 2015; Pool et al., 2016).

securities have no expense ratio, and we assume an implicit expense ratio of 82 for "other" investments (Brightscope and Investment Company Institute, 2014).[14] The weighted average expense ratio is 39 bps.

Expense ratios are only a part of IRA fees; IRA fees also include account fees, advisory fees, and wrap fees. These fees are potentially an important component of total fees. However, to the best of our knowledge no comprehensive estimates of these costs are available. For this reason, we assume very conservatively that average account fees are 50 bps, the same as for DC plans.

## 4.3 Trading costs

Individuals trading on their own account incur explicit trading commissions. Individuals trading mutual fund shares are usually charged no explicit trading costs, but the funds themselves incur all the same costs. The explicit commissions incurred by the fund are not included in the expense ratio, and therefore they are implicit costs for the individual owner. While index funds do not trade much, active funds have considerable asset turnover. Livingston and Zhou (2015) estimates that equal-weighted average portfolio commissions alone are in the order of 18 bps. Wealthfront (2016) finds a very similar number (20 bps).

Individuals and funds also incur implicit trading costs such as bid-ask spreads, defined as the difference between the buy price and the sell price, and market impact, defined as adverse price moves caused by the fund's trades. Usually, these costs are secondary for individual investors, but they are important for mutual funds. Because of their size, unique disclosure requirements, and liquidity needs, mutual funds' trades are more predictable than those of other investors; mutual funds are "sitting ducks" liable to be front-run (Shive and Yun, 2013), to trade against short-sellers (Arif et al., 2016), and to face adverse price pressure (Ben-Rephael et al., 2011). These costs are not straightforward to assess even for the fund itself, and therefore they are rarely or never disclosed, but they are reflected in net returns.

Perhaps because of the difficulty of quantifying implicit trading costs, the literature reports a wide range of estimates. Moreover, the studies we are aware of only analyze equity mutual funds. Wermers (2000) estimates that commissions, transaction costs and cash drag due to liquidity cause a 230 bps wedge between the average equity mutual fund's returns and the return of the stocks they hold. Edelen et al. (2013) estimate average total trading costs of 144 bps using a sample of over 3,000 U.S. domestic equity funds. In this sample, implicit costs

[14] "Other" investments include structured notes whose payoff bears a complex relation to the performance of the underlying assets. Although the difficulty of understanding the price of a financial product does not necessarily translate to a high price, a growing literature on shrouded prices (Gabaix and Laibson, 2006; Carlin, 2009; Henderson and Pearson, 2011) suggests that this is typically the case.

exceed the average expense ratio (119 bps). However, both estimates are equal-weighted and therefore likely higher than corresponding asset-weighted estimates. The lowest estimate in the literature is Bogle (2014), who estimates the overall impact of commissions and market impact (i.e., both implicit ad explicit costs together) to be around 50 bps for active equity funds, and next to nothing for passive equity funds. According to Morningstar's Fund Flows (Jan 2016), about 70% of assets are held in active funds. Assuming that the ratio is the same for equity funds held in retirement accounts, these two figures combined imply asset-weighted average trading costs for equity funds of roughly 35 bps.

Conservatively, we adopt Bogle's estimate of roughly 35 bps for equity funds. For bond funds, we assume 25 bps, and for money market funds we assume no implicit costs.[15] For individual securities, we assume the same trading costs (35 bps for stocks and 25 bps for bonds). Based on the overall asset allocation in DC plans and IRAs, we estimate total implicit trading costs as 30 bps, the weighted average of these three asset classes, a number that reflects a large asset allocation to equity.[16]

## 4.4 The value of services received in exchange for fees

Investors receive services in exchange for fees. In an EET account, the government owns a virtual account that, under our assumptions, pays the same percentage fees. If the government obtains any benefit in exchange for the fees paid on the virtual account, these benefits should be subtracted from the fee to arrive at a net cost. However, we argue the government does not benefit from most services provided by the asset management industry.

If the government wanted to invest in stocks, we argue it could do so much more efficiently and less expensively (e.g. via a sovereign wealth fund) than it implicitly does in an EET system. Basic portfolio management services are an inexpensive commodity. For instance, in 2016, the U.S. federal government's Thrift Savings Plan (TSP) had a net expense ratio of 3.8 bps (Thrift Savings Plan, 2017). Similarly, Collins (2005) shows that portfolio management

[15] We are not aware of any published estimate of asset-weighted average trading costs for bond funds. However, Bessembinder et al. (2016) estimate that transaction costs on the largest corporate bond trades are roughly 0.20% of trading volume. The asset-weighted average portfolio turnover of bond funds is between 90% and 193%, depending on fund type (Rowley and Dickson, 2012). We use a rough average of 125%. Multiplying these two numbers together, we obtain 25 bps as a lower bound estimate of annual transaction costs incurred as a percentage of assets. For reference, Malkiel (2013) estimates that the average bond fund underperforms the reference index by 82 bps, suggesting that this magnitude is reasonable.

[16] In DC plans, 69% of assets is allocated to equity funds, 20% to bond funds, and 11% to other investments (Collins et al., 2016). In IRAs, 62% is allocated to equity, 19% to bonds, and 19% to other investments (Copeland, 2016). We further combine these weights using the total amount of assets in DC plans and IRAs as of the second quarter of 2017 ($7.5 trillion and $8.4 trillion, respectively) to obtain aggregate weights of 66%, 19% and 15%. Using these weights, weighted average trading costs are 28 bps, which we round to 30 bps.

services for S&P 500 index funds cost between 1 and 5 bps and average of 3 bps across all reporting funds.

### 4.4.1 Distribution and advice

A substantial fraction of the cost of investing in mutual funds and other structured products consists of fees such as 12b-1 fees and sales loads. In large part, these fees cover marketing and other distribution costs. However, Collins (2005) suggests that part of these fees may cover complimentary financial advice bundled with the account. In addition, especially in IRAs, individuals may pay account-level fees that are explicit compensation for advice received (advisory fees, or sometimes "wrap" fees). The government, however, does not benefit from distribution services or any bundled financial advice. Moroever, to the extent that the advice received by the individual causes the government to incur lower fees on its implicit account, in equilibrium the value of this advice is already reflected in the observed level of fees.

### 4.4.2 Asset allocation

Higher fees may be caused by costs associated with creating funds or asset allocations that are customized to a particular group of individuals. For instance, some funds may focus on styles (conservative/aggressive, value/growth), maturity (long/short), sector (small cap/large cap, junk/investment grade) or industry. These funds cater to individuals with particular preferences (e.g., low or high risk tolerance, preference for skewness, etc.), or beliefs (e.g., that the health care industry is about to experience massive growth), or personal situations (e.g., health care industry workers who would like to have no health care stocks in their portfolio). Although individual investors may experience real benefits from holding these funds, these benefits largely cancel out in aggregate, because the government holds a fraction of all these funds. Target date funds are a particularly fitting example. Target funds adjust asset allocations as individuals age and get closer to the target retirement date. Clearly, holding target date funds of *all* target dates does not create value to the government.

In general, because individuals' allocations to specialized funds largely cancel out in aggregate, the fees paid to obtain these allocations in the government's virtual accounts entail zero benefit for the government. However, to the extent that individuals' allocations do not cancel out perfectly, the average asset allocation in tax-deferred accounts may differ slightly from the market portfolio. If this departure from the market portfolio exists and in the unlikely case that it is optimal for the government, even this modest benefit could be obtained at a much lower expense by running a sovereign wealth fund.

| | Performance (bps) | | |
|---|---|---|---|
| Source | Net | Gross | Benchmark |
| Berk and van Binsbergen (2015) | -12 * | | Investable Vanguard funds |
| Malkiel (2013) | -64 | | Large cap active vs. SP500 Index |
| Malkiel (2013) | -82 | | Bond funds vs. Barclays US Agg |
| Fama and French (2010) | -100 | -5 | 3- and 4-factor benchmarks |
| Wermers (2000) | -100 | 130 | Own stock holdings |
| Carhart (1997) | -154 per 100 ** | | 1-, 3- and 4-factor benchmarks |
| Jensen (1968) | -40 | ~0 | 1-factor benchmark (CAPM) |
| Fama (1965) | -60 | +20 | Market |

Table 4: **Estimates of average equity mutual fund underperformance.** "Net" and "Gross" refers to expenses. The definition of "expenses" is typically the expense ratio, but in the case of Wermers (2000) it includes everything including cash drag and trading costs (see text) — Footnotes:
[*] Underperformance with respect to the Vanguard benchmark, which charges fees of 18 bps
[**] 100 bps of expense ratio are associated with underperformance of 154 bps.

### 4.4.3 Fees and alpha

Actively managed funds have significantly higher fees than passive index funds. However, it is possible that actively managed funds also have higher expected returns. On the one hand, skilled managers may be able to generate enough excess returns that even after paying higher fees shareholders are able to come out ahead. On the other hand, active management is at least in part a zero-sum game. Even if some mutual funds show evidence of excess returns, at least some of the gain comes at the expense of other mutual funds. As Fama and French (2010) point out, "the aggregate portfolio of actively managed U.S. equity mutual funds is close to the market portfolio, but the high costs of active management show up intact as lower returns to investors". Unless funds held in tax-deferred accounts are systematically winning the zero-sum game against funds in nonretirement accounts or against non-funds, the aggregate alpha on the implicit government account is likely to be zero or negative.

Measuring mutual fund performance is difficult. First, actual performance net of the benchmark has a large random component, and a reliable estimate of performance requires a long time series. Second, unlike direct estimates of fees, every benchmark-based estimate implies and depends on an asset-pricing model. As a result, the literature on mutual fund performance contains numerous estimates done using different methodologies and benchmarks, a few of which are summarized Table 4.

The literature begins with classics such as Fama (1965) and Jensen (1968). Both studies show no evidence of managers predictably beating the market on a net-of-fee basis; on

average, mutual funds show a small underperformance with respect to the market benchmark; consistent with market efficiency, this underperformance is of the same magnitude of fees and cash drag. More recently, Carhart (1997) compiles a mutual fund database that is comprehensive and free of survivorship bias, and uses it to replicate the basic result that there is no evidence of skilled or informed mutual fund managers.[17] Using four-factor and three-factor benchmarks, Carhart finds that there is manager-specific persistence in performance that is not explained by fees, but only for the worst-performing funds. He estimates that 100 bps of expense ratio are associated with a 154 bps underperformance with respect to the market. Wermers (2000) decomposes mutual fund returns into stock-picking talent, style, transaction costs, and expenses, concluding that mutual funds hold stocks that beat the market by 1.3%, but the funds' returns underperform the market by 1%. He attributes the large discrepancy to cash drag (0.7%) and expenses and transaction costs (1.6%). Based on four-factor and three-factor benchmarks, Fama and French (2010) estimate net-of-fees underperformance of about 1% per year. Malkiel (2013) compares several categories of funds with their indices, finding that active large-cap equity funds underperform the S&P 500 Index by 64 bps, and bond funds underperform the Barclay US Aggregate Bond Index by about 84 bps.

Some recent studies have focused on investable benchmarks. French (2008) estimates a broad measure of the annual cost of active management, including not only costs faced by individual investors but also costs faced by institutions and market-making gains by financial intermediaries over 1980-2006. The cost of active management is 0.67% of the aggregate value of the market, in addition to the approximately 0.10% cost of passive management. As a passive benchmark, French uses the Vanguard Total Stock Market Index. Berk and van Binsbergen (2015) compare active funds' dollar returns (as opposed to percent returns) against the relevant Vanguard benchmarks. They estimate a value weighted net alpha of -12 bps (not statistically different from zero) in addition to the cost of investing in the Vanguard benchmark (18 bps), implying a total cost of active money management of about 30 bps.

## 4.5 Calibration: excess investment costs under an EET scheme

Under an EET scheme, the government owns an implicit account of size $S \cdot \tau_R$, where $S$ is the aggregate amount of tax-deferred retirement savings, and $\tau_R$ is the effective tax rate on retirement payouts. This implicit account pays annual percentage fees at a rate $f$, and we assume that the government recovers a fraction $\tau_C$ of these fees via corporate taxation of

[17] Malkiel (1995) also addresses survivorship bias and extends the sample period of previous studies which claimed to find persistence in returns. Carhart also addresses those studies, explaining their findings as the result of momentum investing.

the asset managers. Thus, a simple estimate of the annual ongoing flow of excess investment fees paid under an EET scheme, compared to a TEE scheme, can be calculated as:

$$\text{Excess investment fees} = S \cdot \tau_R \cdot f \cdot (1 - \tau_C)\,. \tag{5}$$

The aggregate amount of tax-deferred retirement savings, $S$, is estimated using data from the Investment Company Institute's Retirement Market Statistics (2017Q2). It is calculated as the sum of assets in individual retirement accounts and employer plans ($15.3 trillion), minus the fraction of these amounts in TEE plans ($0.9 trillion). This results in an estimated amount $S = \$14.4$ trillion. This figure excludes the federal government's Thrift Savings Plan (TSP), whose fees are negligible.

The effective tax rate on retirement payouts, $\tau_R$, is estimated in a few different ways. An average marginal tax rate range of 20%–30% is deduced by reverse-engineering present-value tax expenditure estimates published by the federal government (Office of Management and Budget, 2014) or its employees (Lurie and Ramnath, 2011). Using data on retirement wealth reported in the Survey of Consumer Finances, we independently estimate that the average dollar paid out of a tax-deferred retirement account is currently taxed at a rate of approximately 26%. We choose 20% as a conservative estimate. The size of the implicit government account is therefore $S \cdot \tau_R = \$2.9$ billion.

Based on asset management and account fees of 50 bps, implicit trading costs of 30 bps, and zero benefit, a conservative, asset-weighted estimate of "all-in" average fees is 80 bps, or $f = 0.8\%$. Finally, $\tau_C$, the corporate tax rate, is simply the top statutory corporate tax rate of 35%. Using all these inputs in Eq. (5), we obtain our estimate:

$$\text{Excess investment fees} = S \cdot \tau_R \cdot f \cdot (1 - \tau_C) =$$

$$= \$14.4 \times 20\% \times 0.8\% \times (1 - 35\%) = \$15.0 \text{ billion}.$$

Our estimate of assets under management is also conservative, as it ignores another $7.1 trillion of tax-deferred assets in state and local government and corporate defined-benefit pension plans. Although these assets do not belong to any individual in particular, they are subject to the exact same tax deferral benefit: the contribution is made with pretax money, and benefits are taxed not when the employee becomes entitled to them, but when they are actually paid out. Therefore, even defined-benefit plan assets can be decomposed into an employees' account and a government account earmarked to pay future taxes. While defined-benefit plans are likely to incur lower asset management costs than defined contribution plans or IRAs, they still incur a positive cost of managing the assets held in the government's

virtual account. Accounting for these assets would increase our estimate of the size of the government's virtual account by another $1.4 trillion. Assuming lower costs for DB plans (50 bps instead of 80 bps), this increases the estimate of total government costs to $19.6 billion.

In Table 5, we carry out the same back-of-the envelope calculation for the seven countries with the largest dollar amounts of tax-deferred assets. For each country, we obtain information on all existing types of tax-advantaged retirement plans and their tax treatment from OECD (2015a; 2015b). For EET plans and other plans with a tax-deferral feature, we obtain an estimate of the total assets in each type of plan from various sources. We then estimate the size of the implicit government account by multiplying total tax-deferred assets by the average tax rate on retirement payouts ($\tau_R$). We obtain information on average retirement income from each country's statistical office, and information on basic deductions and tax brackets from each country's tax authority. With this information, we estimate a lower bound to $\tau_R$ as the average tax rate faced by a person earning the average retirement income with no other income. Fees are estimated as the asset-weighted average of money market, equity and fixed-income mutual fund fees based on overall (not retirement-only) asset allocation in each country. Information on fees is collected from Morningstar (Alpert et al., 2013) and other sources. As before, $\tau_C$, the corporate tax rate, is simply the top statutory corporate tax rate as reported by each country's tax authority.[18] We then aggregate the different types of plans up to the country level and convert to U.S. dollars using current exchange rates.

For consistency with the rest of the paper we exclude defined benefit (DB) pension plans from the calculation. With or without DB plans, the U.S. has the world's largest retirement assets, and therefore leads the list. However, other countries have substantial amounts of DB retirement assets (United Kingdom, Netherlands and Japan), and omitting DB leads to an important underestimate of the size of the implicit government account. In the case of United Kingdom and Netherlands, this underestimate meaningfully affects the estimated subsidy.

The average tax rate on retirement payouts ($\tau_R$) is another important factor. Although Switzerland, Australia and Japan have significant tax-deferred assets, the estimated sub-

[18]Unlike other countries, Australia's Superannuation Guarantee is taxed under a TTE scheme: contributions are taxed at favored flat rates (usually 15%) and returns are also taxed at favored rates (15% for interest and dividends, 10% for capital gains), while usually payouts are tax-exempt. Thus, compared to a pure TEE scheme in which all taxes are levied upfront, the Australian system entails some degree of tax deferral. In this case, we define $\tau_{TEE}$ as the tax rate that under a pure TEE scheme would raise the same present value of tax revenue as the actual revenue raised under the current scheme. Then, since the Australian government does levy a 15% upfront tax on contributions, the size of the government's implicit account is simply the remainder, $\tau_{TEE} - 15\%$.

| | Retirement Assets | | Govt. Acct. Size | | | | Subsidy | |
|---|---|---|---|---|---|---|---|---|
| Country | $b | % Deferred | $\tau_R$ | $b | Fees | $\tau_C$ | $b | % GDP |
| United States | 15,315 | 94% | 20% | 2,878 | 0.80% | 35% | 15.0 | 0.08% |
| Canada | 2,082 | 95% | 15% | 295 | 2.06% | 15% | 5.2 | 0.34% |
| United Kingdom | 950 | 32% | 20% | 41 | 1.45% | 20% | 0.7 | 0.02% |
| Netherlands | 108 | 100% | 39% | 41 | 1.41% | 25% | 0.4 | 0.06% |
| Switzerland | 945 | 100% | 4.0% | 38 | 1.29% | 18% | 0.4 | 0.06% |
| Australia | 1,797 | 55% | 3.4% | 34 | 1.10% | 30% | 0.3 | 0.02% |
| Japan | 112 | 100% | 2.6% | 3 | 1.47% | 30% | 0.0 | 0.00% |

Table 5: **Estimated subsidy to the asset management industry in seven countries with the largest EET retirement assets.** Fees are the asset-weighted average of money market, equity and fixed-income mutual fund fees based on overall (not retirement-only) asset allocation in that country. For each country, $\tau_R$ (the tax rate on retirement income, and therefore the fraction of EET assets that implicitly belong to the government) is calculated as the average tax rate faced by a person earning the average retirement income with no other income. $\tau_C$, the corporate tax rate, is simply the top statutory tax rate. Sources: see text.

sidy is small simply because these countries expect relatively little future tax revenue from retirement accounts.

Finally, the level of fees is obviously an important determinant of the size of the subsidy. Although our non-U.S. fee estimates are not as precise as the U.S. estimates, greater accuracy is unlikely to change our findings. For instance, within these seven countries, Canada has the second-largest subsidy in dollar terms ($5.2 billion) and the largest as a fraction of GDP (0.34%). This is in part driven by the surprisingly large fees charged by Canadian funds (2.06%). We have no evidence that this figure is exaggerated, as two independent sources report numbers in the neighborhood of 2% (Alpert et al., 2013; Investor Economics, 2012). However, Canada would not drop in either dollar or percent rankings even using a fee estimate of 1%.

# 5 A general equilibrium model of retirement savings

In this section we examine the conditions under which a larger amount of retirement assets would truly result in a larger amount of resources dedicated to asset management. Consider first the case in which there are no fixed costs and all asset management costs are a linear function of the amount of assets under management. Under the assumption of proportional costs, there is no need to solve an equilibrium model. Upon a switch from TEE to EET, the government postpones the receipt of an amount $B$ of revenue until later. This revenue

shortfall is covered by issuing bonds by an amount $B$. Thus, aggregate retirement savings and aggregate assets increase (compared to what they were under TEE) by exactly the same amount $B$, directly causing aggregate asset management costs to increase. In addition, if investors' demand for the asset management services of a given firm is not perfectly responsive to price, the switch from TEE to EET would further reduce investors' sensitivity to fees, as the same dollar fees become a smaller percentage of asset, with the end result of increasing asset managers' equilibrium profits.

Of course, the assumption that asset management costs are proportional to assets under management is extreme. The asset management business is likely subject to substantial economies of scale. Therefore, consider the other extreme in which all asset management costs are fixed costs and none are variable. In this case, a switch from TEE to EET can still indirectly bring about an increase in resources devoted to asset management by increasing the number of firms that are viable in equilibrium, and with it, the total number of times fixed costs are incurred. If consumers are not perfectly sensitive to prices, a switch from TEE to EET and the associated increase in assets likely enables asset managers to charge higher dollar fees. If entry is free, higher fees cause an increase in the number of firms, as new competitors enter the market until equilibrium profits are zero again. If no new firms can enter, on the other hand, then the total resources devoted to asset management are constant by assumption. Even in this case, however, existing firms could charge higher dollar fees, enjoying higher profits and a larger government subsidy. Thus, when all asset management costs are fixed costs, upon a switch from TEE to EET the subsidy increases, and the total amount of resources devoted to asset management is likely to increase too.

Finally, suppose that all costs are per-participant. Clearly, a switch from TEE to EET would not increase resources used, because the number of investors remains constant. However, unless investors are perfectly sensitive to prices, dollar fees would likely increase because, once again, investors' sensitivity to dollar fees is decreased.

Summarizing, it is only under strict assumptions that an increase in assets does not cause an increase in resources devoted to asset management or a transfer to the asset management industry. First, investors must be perfectly sensitive to prices; second, neither the amount of assets under management nor the number of asset management firms must enter the aggregate production function for asset management services. Because neither assumption is likely to hold in reality, intuition suggests that in general equilibrium the subsidy would almost certainly result in a transfer to the asset management industry, and this transfer is likely to result in excess real resources devoted to asset management. To be more precise, we need to take a stand on what assumptions best describe the actual asset management industry. Existing literature suggests that entry is essentially free, investors are not perfectly

sensitive to prices, and there are substantial economies of scale.

## 5.1 Empirical evidence on cost structure, market structure and competition

Empirical evidence and casual observation suggest low barriers to entry or expansion in the mutual fund industry (Hubbard et al., 2010; Baumol et al., 1990). In 2014 alone, 654 new mutual funds and 69 new mutual fund sponsors entered the industry, for a net increase of 292 funds and 25 fund sponsors (ICI, 2015). A similar situation is reflected in the non-mutual fund segments of the asset management industry. For instance, in a 2016 survey sent by a leading industry publication to 1,070 known third-party retirement plan administrators, the majority of respondents were established in the past 25 years.

Evidence of economies of scale on the cost side is presented by several studies (Baumol et al., 1990; Latzko, 1999; Coates and Hubbard, 2007; Dyck and Pomorski, 2011). Intuition suggests that the asset management business should have a strong fixed-cost component. For instance, Gao and Livingston (2008) find that the "paperwork" part of mutual funds expenses (custodian fees, recordkeeping fees, etc.) does not vary meaningfully with fund size. Statements by industry insiders also confirm this intuition: Kahn (2002) quotes Jeffrey S. Molitor, then director for portfolio review at Vanguard, as saying that the "marginal cost of managing increasing dollars is minimal." This statement refers specifically to *active* funds whose management Vanguard outsources to subadvisers; for passive funds, the economies of scale are obvious.

However, costs are not fees. Lacking perfect competition, firms can charge more than their marginal cost and therefore fees will not necessarily drop as costs do. Malkiel (2013) notes that "academic research has documented substantial economies of scale in mutual fund administration", but between 1980 and 2010, in spite of a more than 100-fold increase in assets under management, percentage fees stayed relatively flat around 70-80 bps. In particular, the fees charged by active managers rose from 66 to 91 bps. Similarly, Philippon (2015) argues that in aggregate the unit cost of financial intermediation (defined as the ratio of the income of financial intermediaries to the quantity of intermediated assets) is very stable in the long run and it has not dropped over the last century in spite of a large increase in total assets under management—both in absolute terms and as percent of output. Alone, the fact that fees are charged as a percentage of assets under management (as opposed to absolute dollar prices) suggests that either there is some variable cost component, or that clients with more assets have lower price sensitivity and asset managers are engaging in price

discrimination.[19]

The available empirical evidence supports both explanations. On the one hand, although all empirical studies of costs support substantial economies of scale, all studies also find that costs do increase as assets increase (both overall assets, and assets per account). On the other hand, there is also abundant evidence of investors' imperfect sensitivity to price. Although funds with lower fees tend to have higher market shares (Hubbard et al., 2010), many studies point out the continued existence of dominated funds; for instance, Hortacsu and Syverson (2004) document the existence of 82 distinct S&P 500 Index funds with large dispersion in fees (an interquartile range of 89 bps). We update their analysis using 2015 data and find that the large dispersion still persists. Hortacsu and Syverson make sense of this phenomenon by assuming the existence of search costs. Gil-Bazo and Ruiz-Verdú (2009, 2008) find evidence that "underperforming funds and funds faced with less performance-sensitive investors charge higher marketing and nonmarketing fees," as in a theory initially proposed proposed by Christoffersen and Musto (2002). They propose an explanation based on investors' inability to precisely observe the quality of fund management. Other explanations are based on the inability of retail investors to observe shrouded prices of complex financial products (Gabaix and Laibson, 2006; Carlin, 2009; Henderson and Pearson, 2011), or the unwillingness to sever relationships with brokers (Bergstresser et al., 2009) or trusted advisors (Gennaioli et al., 2015). In particular, Bergstresser et al. (2009) find that broker-sold funds are more costly *and* underperform, implying that the broker channel enables the survival of otherwise dominated funds. A report by the Executive Office of the President of the United States (2015) summarizes the academic literature on investment advice, finding that fund distribution channels are able to charge investors who seek for investment advice one or two percentage points. Finally, Gârleanu and Pedersen (2015) features a financial market where finding information is costly (a la Grossman and Stiglitz, 1980), with the additional feature that finding money managers is also costly. Agents can pay a cost and allocate money to a well-managed fund, or decide to stay uninformed and become "noise allocators". Only by introducing the concept of noise allocators are the authors able to reproduce several otherwise puzzling empirical facts, such as the existence of a sizable minority of mutual fund managers

[19] Costs are not fees, and fees are not net-of-fees performance. In Section 4 we discussed the relationship between the level of fees and the level of performance. However, theoretical literature suggests that performance could be subject to diseconomies of scale (Berk and Green 2004; Pástor and Stambaugh 2012). The empirical evidence is mixed. Berk and van Binsbergen (2015) find in favor of this, but Reuter and Zitzewitz (2015) present evidence of insignificant changes in performance using a change in Morningstar rating as a shock to assets under management. Dyck and Pomorski (2011) find that large pension plans are able to obtain superior returns through increased access to alternative investments. Pástor et al. (2015) show strong support of the industry-level decreasing returns to scale hypothesis (Pástor and Stambaugh, 2012): as the size of the active mutual fund industry increases, the ability of any given fund to outperform declines. For simplicity, we abstract from performance-related considerations, a conservative assumption.

who can consistently pick stocks well enough to cover their costs.

Our model of spatial competition is very general and it abstracts from the specific causes of limited price sensitivity. Our contribution is to examine the effect of a government subsidy within this model. Although the subsidy does not affect investors' sensitivity to *percent* fees, it reduces their sensitivity to *dollar* fees. As a result, in the presence of fixed costs, investors are subsidized in seeking more variety. Because the underlying limited price sensitivity already results in too many firms, the social welfare effect of a subsidy is particularly deleterious.

## 5.2 A model

In order to formalize the intuition laid out at the beginning of this section, we examine a general equilibrium model of the asset management industry. In this model we make the most conservative assumptions possible that are compatible with the empirical evidence summarized above. First, we assume that the asset management business is essentially a fixed-cost business. Specifically, we assume that each firm needs a fixed amount of labor to operate. Second, we assume no barriers to entry at all. Fixed costs and free entry imply that the equilibrium profits of this industry are zero. Third and last, we assume that individuals' demand for the services of a given firm is not perfectly elastic. Specifically, we model competition among funds as spatial competition (Salop, 1979; Tirole, 1988, Ch.7).

In our simple model we ignore the existence of multiple layers of financial intermediation. The unit of production of asset management services is the "mutual fund", and individuals give their savings directly to mutual funds who charge explicit fees. For this reason, we use the words "fund" and "firm" interchangeably. Funds do not rebate any of the fee revenue to distribution channels, and do not incur any trading costs.

In equilibrium, funds face a downward-sloping demand function, i.e. if they raise their fees, their demand falls, but it does not fall to zero. Although the existence of dominated or duplicate funds and other evidence point to outright inertia as the cause for this low sensitivity, we take a more optimistic (that is, conservative) view, and we assume that individuals have preferences for convenience; more funds means that the distance between an individual and their chosen fund is on average lower, i.e., utility is higher. A low "distance" can be thought of as literally low physical distance from the nearest branch, but also ease of finding (e.g., the fund is recommended by the account administrator), availability (e.g., the fund is part of a small menu of preselected funds, as it is the case for many retirement plans), trust (Gennaioli et al., 2015), or even a preference for non-portfolio characteristics of funds, such as the level of customer service.

In sum, in this model, mutual funds have only fixed costs, they create value, and they face competition that is imperfect but stiff enough to warrant zero equilibrium profits. In spite of that, we aim to (a) show that under a reasonable calibration there will be too many funds in equilibrium, and (b) using comparative statics, show that a switch from TEE to EET causes the number of funds to become even higher.

The model features a two-period economy. When individuals are young (time $t$), they work, produce, consume, and save for retirement by investing via mutual funds in a mix of government bonds and financial assets. When they are old (time $t'$), they receive passive retirement income, consume, and die without bequest. Mutual funds require a fixed amount of labor to operate, and therefore must pay a competitive salary to individuals who otherwise would work to produce goods.

In this section we present assumptions and the main results. A full discussion of the model is provided in Appendix B.

## 5.3 Individuals

Individual $i \in \mathcal{I}$ lives two periods and leaves no bequest. In period $t$ the individual works, saves, and allocates the savings. The individual starts with a net worth of 0 and is endowed with one unit of labor to spend either producing consumption goods, or managing a mutual fund. The production technology is linear: if the individual allocates $L$ units of labor to the production of goods, the output is $F(L) = \omega L$. In order to attract labor, therefore, mutual funds must pay a competitive wage $\omega$. The individual's total supply of labor is inelastic. Overall, the individual's pretax income is $\omega$ regardless of how labor is allocated. Income from work is taxed at rate $\tau_L$. In period $t'$, the individual retires and depletes all the savings.

The individual draws utility from current consumption ($C$) and from discounted future consumption ($\delta \cdot C'$). To finance future consumption, the individual saves and invests an amount $S$. All investment must be carried out via a chosen mutual fund $j$ charging a proportional fee $f_j S$. The individual derives disutility from the distance between the location of chosen fund $j$ ($\iota_j$) and one's own location ($d_{i,j} \equiv |\iota_j - \iota_i|$). A fraction $a$ of the individual's savings is invested in an exogenous storage technology yielding a return $\rho$, and the remainder in government bonds paying a return $r$. The government grants the individual a deduction for savings at a rate $\tau_S$. Thus, the utility of individual $i$ is:

$$U_i = \max_{C,C',S,a,j} \ln C + \delta \ln C' - \gamma d_{i,j} \tag{6}$$

subject to the budget constraints:

$$C = \omega\left(1-\tau_L\right) - S\left(1-\tau_S\right), \tag{7}$$

$$C' = S\left(1-f_j\right)\left[1+a\rho+\left(1-a\right)r\right]\left(1-\tau_R\right). \tag{8}$$

## 5.4 Mutual funds

Mutual funds can differentiate themselves over one qualitative characteristic, $\iota$, defined on the $[0,1)$ circle. On this circle, the distance between two points is the shortest possible distance—for instance, the distance between 0.1 and 0.9 is 0.2, not 0.8. Individuals are uniformly distributed over this circle, and their utility is decreasing in the distance from their chosen fund. Every fund $j \in \mathcal{J}$ needs a fixed amount of labor $\varphi$ just to be able to operate. Fund profits are equal to revenue minus cost:

$$\pi_j = f_j Q_j - \varphi\omega,$$

where $Q_j$ is the fund's assets under management, and $f_j$ are the percent fees the fund charges, so that $f_j Q_j$ is the fund's total revenue. The fund's problem is to choose $f_j$ and $\iota_j$ to maximize $\pi_j$, taking into account that $Q_j$ depends on the fund's choices of pricing ($f_j$) and location ($\iota_j$), and taking competitors' choices as given.

Mutual fund profits accrue to their managers and are taxed as other income. However, because of free entry, profits will be zero, so we do not keep track of who the profit accrues to.

## 5.5 Government

The government spends an exogenously given amount $G$. This expenditure is inevitable and it does not affect the utility of agents. At time $t$, the government taxes income at a rate $\tau_L$, and grants individuals a deduction for savings at a rate $\tau_S$. The government can also borrow an amount $B$ at the market interest rate $r$ to satisfy the government budget constraint:

$$G = \tau_L\omega - S\tau_S + B \tag{9}$$

To pay off the bonds at time $t'$, the government can tax retirement income at a rate $\tau_R$ so as to satisfy

$$B\left(1+r\right) = \tau_R S\left(1-f\right)\left[1+a\rho+\left(1-a\right)r\right] \tag{10}$$

where $f$ is the equilibrium level of fees.

The government takes $\tau_L$ and $\tau_R$ as given. For $\tau_S$, only two policy options are on the table: EET retirement accounts ($\tau_S = \tau_L$), and TEE accounts ($\tau_S = 0$). Once chosen $\tau_S$, the government chooses $B$ to balance the budget constraint.

## 5.6 Market equilibrium

The model has one symmetric equilibrium in which $N$ funds distribute themselves over the circle at equal distance from one another and set equal fees (Tirole, 1988). Individuals' equilibrium savings are

$$S^* = \omega \frac{1-\tau_L}{1-\tau_S} \frac{\delta}{1+\delta}. \tag{11}$$

Obviously, aggregate savings are higher with an EET scheme than they are with a TEE scheme because in the case of EET retirement accounts, $\tau_S = \tau_L$, and in the case of TEE accounts, $\tau_S = 0$.

The equilibrium number of funds is defined implicitly by the following quadratic equation:

$$\frac{\delta}{\gamma} N^2 + N = \frac{S}{\varphi \omega}. \tag{12}$$

The explicit solution for $N^*$ is unwieldy, but without solving explicitly, it is evident from Eq. (12) that $N$ is an increasing function of $S$. Therefore, the number of funds does increase under an EET scheme as compared to a TEE scheme.

Finally, the equilibrium level of fees is a decreasing function of $N$, i.e., higher competition does translate to lower percent fees:

$$f_j^* = f^* = \frac{1}{1 + \frac{\delta}{\gamma} N}. \tag{13}$$

It is easy to show that, as $S$ increases because of a switch from TEE to EET, $f$ does not drop proportionally, so that total fee revenue $fS$ increases, supporting a larger number of funds.

## 5.7 Planner solution

We compare the market equilibrium with the solution chosen by a benevolent planner who also takes $G$ as exogenously given. The planner chooses savings $S$, and number of funds $N$

directly to maximize social utility:

$$U = \max_{C,C',S,N} \ln(C) + \delta \ln(C') - \gamma \bar{d}, \tag{14}$$

where $\bar{d}$ indicates the *average* distance between an investor and their fund. The planner's budget constraints are simply

$$C = \omega(1 - \varphi N) - S - G, \tag{15}$$

$$C' = S(1 + \rho) = (\omega(1 - \varphi N) - C - G)(1 + \rho). \tag{16}$$

Under these assumptions, we obtain an implicit expression for the socially optimal number of funds:

$$4\frac{1+\delta}{\gamma}N^2 + N = \frac{1 - G/\omega}{\varphi}. \tag{17}$$

## 5.8 Calibration

Next, we turn to the question whether a larger asset management industry (a larger $N$) is optimal. Consider a TEE-based system as the starting point. If that were the social optimum, a shift to EET would imply that there are too many funds. Because we are departing from the optimum, however, the total loss need not be too great, as the social welfare function is not steep at the optimum. Only if such a shift caused a large change in the equilibrium number of funds would social welfare suffer serious consequences. On the other hand, if a TEE-based system were to produce too few funds, then a shift to EET would get society closer to the optimum, and it would be very beneficial. Finally, if a TEE-based system were to produce too many funds, the shift to EET would be particularly pernicious, as it would bring society even farther from the optimum, starting at a point where the social welfare function is already steep. Thus, if the number of funds under the market solution ($N_M$) is higher than the number of funds under the planner solution ($N_P$), TEE is to be preferred because, with EET, $N_M$ would be even higher. Vice versa, if $N_M < N_P$ , EET would be preferred.

It is important to note that in this model we give the asset management industry the benefit of the doubt, because we assume that every additional fund improves the utility of the average individual. It would have been also possible, consistent with the prevailing empirical evidence, to write a model with captive demand (Gil-Bazo and Ruiz-Verdú, 2008, 2009; Bergstresser et al., 2009; Gennaioli et al., 2015) and shrouded fees (Gabaix and Laibson, 2006; Carlin, 2009; Henderson and Pearson, 2011). In such a model, back-loaded taxation would still cause an increase in the resources devoted to asset management, and welfare

| Parameter | Description | TEE | EET |
|---|---|---|---|
| $\delta$ | 30-year discount factor | 0.545 | 0.545 |
| $\gamma$ | Preference for funds | 3500 | 3500 |
| $\tau = G/\omega$ | Tax rate/Government Expenditure | 0.18 | 0.18 |
| $\tau_S$ | Subsidy on savings | 0 | 0.18 |
| $\varphi$ | Labor fraction of one fund | 0.000017% | 0.000017% |
| $\varphi N_P$ | Optimal resources employed in asset management | 0.87% | |
| $\varphi N_M$ | Actual resources employed in asset management | 1.70% | 1.88% |
| $f$ | 30-year level of fees | 5.882% | 5.341% |

Table 6: Calibration of the general equilibrium model

consequences would be undoubtedly more severe.

To simplify the calibration, assume that the tax rate is set to achieve a balanced budget in the absence of tax subsidies: $\tau_L = G/\omega$ where $G$ is the exogenous expenditure and $\omega$ is the output, so the tax rate is simply the ratio of government expenditure / output. Thus, we can rewrite the implicit expression for $N_M$ as:

$$\frac{1+\delta}{\gamma}N_M^2 + \frac{1+\delta}{\delta}N_M = \frac{1}{1-\tau_S}\cdot\frac{1-G/\omega}{\varphi}. \tag{18}$$

This new expression is very similar to the implicit expression for $N_P$, the planner solution:

$$4\frac{1+\delta}{\gamma}N_P^2 + N_P = \frac{1-G/\omega}{\varphi} \tag{19}$$

Under a TEE scheme, $\tau_S = 0$ and therefore the right hand side of the two expressions is the same. Because of the "4" coefficient on the quadratic term, $N_P \approx N_M/2$, and therefore a TEE should be preferred under most parameterizations. However, for very impatient investors ($\delta \approx 0.01$), the linear term prevails, obtaining the opposite result. The question, therefore, is whether under a reasonable calibration we obtain the result we anticipate.

Table 6 shows two possible calibrations, with TEE ($\tau_S = 0$) and with EET ($\tau_S = \tau$). The details of the calibration are explained in Appendix C. The table shows that the excess resources dedicated to the asset management industry in an EET scenario as compared to a TEE scenario are substantial: $\varphi\left(N_M^{Trad} - N_M^{Roth}\right) = 0.18\%$. Multiplying this figure by annual total output ($18 trillion), the amount of excess resources dedicated to the asset management industry is about 32.3 billion dollars per year.

This result is related to a recent literature on the optimal size of the financial services

industry. Greenwood and Scharfstein (2013) note that the U.S. financial services industry (encompassing insurance, securities and credit intermediation) as a share of GDP doubled in size in the last 50 years, going from 4% to 8%. Half of this 4 percentage point increase (1.5–2 percentage points) is due to growth of the asset management industry, which has managed to keep its revenue a relatively stable fraction of the stock market. Malkiel (2013) argues that in spite of a more than 100-fold increase in assets under management, the benefits from the vast economies of scale inherent to the asset management industry have accrued to industry insiders, because fees (as a percentage of assets under management) have not fallen proportionately. French (2008) and Fama and French (2010) attempt to quantify the amount of resources spent in the zero-sum game of attempting to beat the market. Philippon and Reshef (2012) find empirically that financial deregulation is associated with greater skill intensity, increased job complexity, and higher wages for finance employees. Bolton et al. (2016) show in theory that it is possible for the financial industry to extract excessively high rents for the provision of financial services, thus attracting too much talent. Our study features another mechanism by which, because of search frictions, the financial industry attracts too much labor, and points out that under reasonable assumptions this mechanism not only exists, but it is magnified by government policy.

# 6 Conclusion

A standard benchmark model yields a neutrality result between front-loaded (TEE) and back-loaded (EET) taxation of retirement savings. Individuals obtain the same consumption in every period, and the present value of government tax revenues is the same under the two systems. The timing of taxation is different: back-loaded taxation leads to higher outstanding government debt and a correspondingly greater amount of retirement assets. These additional assets represent an implicit government portfolio, i.e., resources earmarked to pay future taxes when the money is distributed from the account. In this paper, we add one crucial bit of realism to the benchmark model: asset management fees. We show that the indifference result breaks down because the government is paying an estimated $15 billion a year in fees on its large implicit portfolio. Under a rough calibration of a general equilibrium model, we also show that back-loading taxation inefficiently increases the amount of resources spent on asset management, thus reducing welfare.

Our results raise the policy issue of whether governments should encourage or possibly mandate wider adoption of TEE. Our model, taken literally, would imply that TEE dominates EET. However, both the benchmark model and our model abstract from other potential drivers of the policy choice between front-loaded and back-loaded taxation of re-

tirement savings.

First, most real-world tax systems are progressive. With progressive taxation, lifetime taxes are more aligned with lifetime income under EET than under TEE. For example, consider two workers with the same lifetime income: one with high annual earning and a short work life (e.g. a firefighter), and another with lower annual earnings but a longer work life (e.g. a municipal clerk). Under TEE, the firefighter would pay more lifetime taxes than the clerk. Under EET, the gap between the lifetime taxes paid by the two workers will shrink and potentially disappear. In addition, if a worker's lifetime income is not known in advance, EET may work as "insurance" because the average tax rate on distributions is higher when the account balance is higher.

The effect of progressive taxation has contingent consequences as well. A switch to TEE for new contributions ("Rothification") is being discussed as part of the current U.S. debate on tax reform. Such a switch would not affect all individuals equally. Individuals with high labor income but no expected income in retirement would be disproportionately affected, as their current marginal tax rate is high and their tax rate in retirement is low.[20]

Second, behavioral biases that cause people to save too little are a frequently-cited motive for the provision of retirement saving incentives. Behavioral arguments exist in favor of either taxation scheme. During an individual's work life, back-loaded taxation could induce individuals to underestimate the future "tax bite" and still save too little (Iwry and John, 2009). However, back-loaded taxation could also provide a more powerful behavioral response because of the "instant gratification" of an immediate tax benefit (Thaler, 1994; Iwry and John, 2007), or simply because individuals don't fully understand that EET balances are subject to taxation in the future. Beshears et al. (2017) find empirically that TEE induces individuals to save more, and argue that this is because individuals focus on nominal contributions and savings and underweight future taxes.[21] In addition, during the individ-

[20] As part of the recent U.S. debate, industry sources argued that a shift to Roth (TEE) would endanger retirement security because EET provides more resources for retirement under progressive taxation. That argument, which focuses on the overall level of taxation, is distinct from ours, which focus on distribution. If that argument is correct, under a TEE system the government could use the additional tax revenue to provide additional resources for retirees. Our model allows for differential tax rates during work life and retirement, which we consider to be a separate policy choice.

[21] Our results rely on the assumption that individuals are rational savers and therefore under our benchmark model contribute enough extra dollars into an EET plan relative to what they would contribute under a TEE plan to ensure that retirement consumption would be the same under the two plans. Beshears et al. (2017) provide evidence that individuals do not adjust their retirement savings in this way, but instead find that contributions under a Roth 401(k) plan (TEE) are similar to those under a traditional 401(k) plan (EET), implying a higher retirement consumption under a TEE plan. If these findings generalized to the policy experiments we consider, they may complicate our welfare analysis, but the gist of our argument would still be valid. TEE is more cost effective than EET. If the total amount of assets is constant under EET and TEE, then TEE delivers a larger savings subsidy for the same cost to the government. At the other extreme, if, as in our paper, the total amount of retirement consumption is constant, then TEE deliver

ual's retirement, under progressive taxation an EET system penalizes bulk withdrawals with higher tax rates. As part of the recent British debate, an Economist editorial claims that this feature "is actually quite useful in that it stops people blowing their pension pot in a spending spree at 65" (Buttonwood, 2015). Of course, the other side of the coin is that EET penalizes individuals who withdraw funds in bulk for legitimate reasons such as hardship or investment. We are not aware of any systematic study of this tradeoff.

Third, there are political economy considerations that are important to the debate over a shift from EET to TEE. U.S. budget rules make it more cumbersome to pass bills that increase the total budget deficit over a five- or ten-year window. A transition from EET to TEE generates more cash flow upfront and less when the relevant workers retire, thus bringing more revenue into the budget window, resulting in a temporary deficit reduction. This may in turn make it easier to pass other legislation that involves lower taxes or higher spending.[22] This additional short-run fiscal flexibility may or may not be considered desirable, but it certainly makes TEE attractive to many real-world policymakers. Indeed, one of the purported motivations for proposing Roth accounts in the U.S. was to help "fund" cuts in the capital gains tax (Pine, 1989).

Our analysis raises some additional policy issues. The $15 billion cost to the government exists because the government owns an implicit account that incurs substantial investment fees. One way to reduce this cost is to shrink or eliminate this account, as described above. An alternative approach would be to leave the size of the account unchanged, but to reduce investment fees to a more acceptable level. For example, the government could explicitly segregate its implicit account and try to use its market power to negotiate lower fees. The U.S. Department of Labor fiduciary rule implemented in June 2017 could also help reduce fees on the government account. One of the stated motivations for the rule was protecting retail investors from aggressive marketing of high-fee products—especially senior investors that prepare to roll over their employer plan savings into an individual retirement account. If the rule has the effect of reducing the fee incurred by individual investors, our results suggest that it would also protect the government's future revenue from being eroded by high fees, providing a possible additional rationale for implementing a fiduciary rule.

---

the same savings subsidy for a lower cost to the government.

[22] The effectiveness of this approach is complicated by the "Byrd rule," which requires a supermajority to approve a deficit increase beyond the period covered by the budget resolution (Committee for a Responsible Federal Budget, 2016).

# References


Abel, A. B., 2001. The effects of investing Social Security funds in the stock market when fixed costs prevent some households from holding stocks. American Economic Review 91 (1), 128–148.

Alpert, B. N., Rekenthaler, J., Suh, S., 2013. Global fund investor experience 2013 report. Tech. rep., Morningstar Fund Research.

Antolín, P., de Serres, A., de la Maisonneuve, C., 2004. Long-term budgetary implications of tax- favoured retirement plans, oECD Economics Department Working Papers, No. 393, OECD Publishing.

Arif, S., Ben-Rephael, A., Lee, C. M., 2016. Mutual funds and short sellers: Why does short-sale volume predict stock returns?, stanford University Graduate School of Business Research Paper No. 14-35, available at http://ssrn.com/abstract=2496990.

Ayres, I., Curtis, Q., 2015. Beyond diversification: The pervasive problem of excessive fees and 'dominated funds' in 401(k) plans. Yale Law Journal 124 (5), 1346–1835.

Baumol, W. J., Goldfeld, S. M., Gordon, L. A., Koehn, M. F., 1990. The Economics of Mutual Fund Markets: Competition Versus Regulation. Rochester Studies in Managerial Economics and Policy. Kluwer Academic Publishers.

Ben-Rephael, A., Kandel, S., Wohl, A., 2011. The price pressure of aggregate mutual fund flows. Journal of Financial and Quantitative Analysis 46 (2), 585–503.

Bergstresser, D., Chalmers, J. M. R., Tufano, P., 2009. Assessing the costs and benefits of brokers in the mutual fund industry. Review of Financial Studies 22 (10), 4129–4156.

Berk, J., van Binsbergen, J., 2015. Measuring skill in the mutual fund industry. Journal of Financial Economics.

Berk, J. B., Green, R. C., 2004. Mutual fund flows and performance in rational markets. Journal of Political Economy 112 (6), 1269–1295.

Beshears, J., Choi, J. J., Laibson, D., Madrian, B. C., 2017. Does front-loading taxation increase savings? evidence from roth 401(k) introductions. Journal of Public Economics 151, 84–95.

Bessembinder, H., Jacobsen, S., Maxwell, W., Venkataraman, K., 2016. Capital commitment and illiquidity in corporate bonds.

Bogle, J. C., 2014. The arithmetic of 'all-in' investment expenses. Financial Analysts Journal 70 (1), 13–21.

Bolton, P., Santos, T., Scheinkman, J. A., 2016. Cream-skimming in financial markets. Journal of Finance 71 (2), 709–736.

Brady, P., 2012. The tax benefits and revenue costs of tax deferral. Tech. rep., Investment Company Institute, Washington, DC.

Brightscope, Investment Company Institute, 2014. The BrightScope/ICI defined contribution plan profile: A close look at 401(k) plans. Tech. rep.

Burman, L., Gale, W. G., Weiner, D., 2001. The taxation of retirement saving: Choosing between front-loaded and back-loaded options. National Tax Journal 54 (3), 689–702.

Buttonwood, 5 Aug. 2015. EET your TEE, George. The Economist.

Carhart, M. M., 1997. On persistence in mutual fund performance. Journal of Finance 52 (1), 57–82.

Carlin, B. I., 2009. Strategic price complexity in retail financial markets. Journal of Financial Economics 91, 278–287.

Christoffersen, S. E., Musto, D. K., 2002. Demand curves and the pricing of money management. Review of Financial Studies 15 (5), 1499–1524.

Coates, J. C., Hubbard, R. G., 2007. Competition in the mutual fund industry: evidence and implications for policy, harvard John M. Olin Discussion Paper No. 592, available at http://ssrn.com/abstract=1005426.

Collins, S., 2005. Are S&P 500 index mutual funds commodities? Investment Company Institute Perspective 11 (3).

Collins, S., Holden, S., Duvall, J., Chism, E. B., 2016. The economics of 401(k) plans: Service, fees and expenses, 2015. ICI Research Perspective 22 (4).

Committee for a Responsible Federal Budget, Dec. 2016. Reconciliation 101. URL `http://www.crfb.org/papers/reconciliation-101`

Committee on Finance of the U.S. Senate, 1997. Expanding IRA's. U.S. Government Printing Office.

Copeland, C., 2016. 2014 update of the EBRI IRA database: Ira balances, contributions, rollovers, withdrawals, and asset allocation. EBRI Issue Brief 424.

Cumings, S., May 2017. 'rothification' uncertainties draw concerns from industry. Tax AnalystsRetrieved on 10/11/2017 at http://www.taxanalysts.org/content/rothification-uncertainties-draw-concerns-industry.

Diamond, P., Geanakoplos, J., 2003. Social security investment in equities. American Economic Review 93 (4), 1047–1074.

Dyck, A., Pomorski, L., 2011. Is bigger better? size and performance in pension plan management, rotman School of Management Working Paper No. 1690724, available at http://ssrn.com/abstract=1690724.

Edelen, R., Evans, R., Kadlec, G., 2013. Shedding light on "invisible" costs: Trading costs and mutual fund performance. Financial Analysts Journal 69 (1), 33–44.

Executive Office of the President of the United States, 2015. The effects of conflicted investment advice on retirement savings. Tech. rep.

Fama, E. F., 1965. The behavior of stock market prices. Journal of Business 38, 34–105.

Fama, E. F., French, K. R., 2010. Luck versus skill in the cross-section of mutual fund returns. The Journal of Finance 65 (5), 1915–1947.

Freeman, J. P., Brown, S. L., Pomerantz, S., 2008. Mutual fund advisory fees: new evidene and a fair fiduciary duty test. Oklahoma Law Review 61, 83–153.

French, K. R., 2008. Presidential address: The cost of investing. The Journal of Finance 53 (4), 1537–1573.

Gabaix, X., Laibson, D., 2006. Shrouded attributes, consumer myopia, and information suppression in competitive markets. Quarterly Journal of Economics 121 (2), 505–540.

Gao, X., Livingston, M., 2008. The components of mutual fund fees. Financial Markets, Institutions and Instruments 17 (3), 197–223.

Gârleanu, N. B., Pedersen, L. H., 2015. Efficiently inefficient markets for assets and asset management, NBER Working Paper 21563, available at http://www.nber.org/papers/w21563.

Geanakoplos, J., Mitchell, O., Zeldes, S. P., 1999. Social Security Money's Worth. Prospects for Social Security Reform. Pension Research Council, University of Pennsylvania Press, Ch. 5, pp. 79–151.

Gennaioli, N., Shleifer, A., Vishny, R., 2015. Money doctors. Journal of Finance 70 (1), 91–114.

Gil-Bazo, J., Ruiz-Verdú, P., 2008. When cheaper is better: Fee determination in the market for equity mutual funds. Journal of Economic Behavior and Organization 67, 871–885.

Gil-Bazo, J., Ruiz-Verdú, P., 2009. The relation between price and performance in the mutual fund industry. The Journal of Finance 64 (5), 2153–2183.

Greenwood, R., Scharfstein, D., 2013. The growth of finance. Journal of Economic Perspectives 27 (2), 3–28.

Grossman, S. J., Stiglitz, J. E., 1980. On the impossibility of informationally efficient markets. American Economic Review 70, 393–408.

Henderson, B. J., Pearson, N. D., 2011. The dark side of financial innovation: A case study of the pricing of a retail financial product. Journal of Financial Economics 100, 227–247.

Holzmann, R., Hinz, R., 2005. Old Age Income Support in the 21st Century. The World Bank.

Hortacsu, A., Syverson, C., May 2004. Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds. Quarterly Journal of Economics, 403–456.

Hubbard, R. G., Koehn, M. F., Ornstein, S. I., Audenrode, M. V., Royer, J., 2010. The Mutual Fund Industry: Competition and Investor Welfare. Columbia Business School Publishing.

ICI, 2015. 2015 Investment Company Fact Book. Investment Company Institute.

Investor Economics, 2012. Mutual fund MERs and cost to customer in Canada: Measurement, trends and changing perspectives.
URL https://www.ific.ca/wp-content/uploads/2013/08/Canadian-Study-Mutual-Fund-MERs-and

Iwry, J. M., John, D. C., 2007. Pursuing universal retirement security through automatic iras. resreport 2007-2, Brookings Institution, Retirement Security Project.

Iwry, J. M., John, D. C., 2009. Pursuing universal retirement security through automatic iras. resreport 2009-3, Brookings Institution, Retirement Security Project.

Jensen, M. C., 1968. Problems in selection of security portfolios. The Journal of Finance 23 (2), 389–416.

Kahn, V. M., 14 July 2002. Investing; mutual fund expertise, for rent. The New York Times.

Latzko, D. A., 1999. Economies of scale in mutual fund administration. Journal of Financial Research 22 (3), 331–339.

Livingston, M., Zhou, L., 2015. Brokerage commissions and mutual fund performance. Journal of Financial Research 38 (3), 283–303.

Lurie, I. Z., Ramnath, S. P., December 2011. Long-run changes in tax expenditures on 401(k)-type retirement plans. National Tax Journal 64 (4), 1025–1038.

Malkiel, B. G., 1995. Returns from investing in equity mutual funds 1971 to 1991. Journal of Finance 50 (2), 549–572.

Malkiel, B. G., 2013. Asset management fees and the growth of finance. Journal of Economic Perspectives 27 (2), 97–108.

OECD, 2015a. Stocktaking of the tax treatment of funded private pension plans in oecd and eu countries. Tech. rep., OECD.
URL http://www.oecd.org/tax/tax-treatment-funded-private-pension-plans-oecd-eu-countri

OECD, 2015b. The tax treatment of funded private pension plans - oecd and eu country profiles. Tech. rep., OECD.
URL http://www.oecd.org/tax/tax-treatment-funded-private-pension-plans-oecd-eu-countri

Office of Management and Budget, February 2014. Budget of the u.s. government. analytical perspectives. fiscal year 2015. Tech. rep., United States Government.

Osborne, G., July 2015. Strengthening the incentive to save: a consultation on pensions tax relief, Her Majesty's Treasury Cm 9102.

Pástor, L., Stambaugh, R. F., 2012. On the size of the active management industry. Journal of Political Economy.

Pástor, L., Stambaugh, R. F., Taylor, L. A., 2015. Scale and skill in active management. Journal of Financial Economics 116, 23–45.

Philippon, T., 2015. Has the U.S. finance industry become less efficient? on the theory and measurement of financial intermediation. American Economic Review 105 (4), 1408–1438.

Philippon, T., Reshef, A., 2012. Wages and human capital in the U.S. financial industry: 1909–2006. Quarterly Journal of Economics 127 (4), 1551–1609.

Pine, A., October 20 1989. GOP senators offer capital gains cut, new type of IRA. Los Angeles Times. Retrieved online on 10/19/2016.

Pool, V. K., Sialm, C., Stefanescu, I., Aug. 2016. It pays to set the menu: Mutual fund investment options in 401(k) plans. Journal of Finance 71 (4), 1779–1812.

Reuter, J., Zitzewitz, E., 2015. How much does size erode mutual fund performance? a regression discontinuity approach, working paper, available at http://ssrn.com/abstract=1661447.

Rosshirt, D. E., Parker, S. A., Pitts, D. A., 2014. Inside the structure of defined contribution/401(k) plan fees, 2013: A study assessing the mechanics of the 'all-in' fee. Tech. rep., Deloitte Consulting LLP.

Rowley, Jr., J. J., Dickson, J. M., 2012. Mutual funds—like ETFs— have trading volume. Tech. rep., Vanguard.

Salop, S. C., 1979. Monopolistic competition with outside goods. Bell Journal of Economics 10 (1), 141–156.

Shive, S., Yun, H., 2013. Are mutual funds sitting ducks? Journal of Financial Economics 107 (1), 220–237.

Thaler, R. H., 1994. Psychology and savings policies. The American Economic Review, Papers and Proceedings of the Hundred and Sixth Annual Meeting of the American Economic Association 84 (2), 186–192.

Thrift Savings Plan, Apr. 2017. Highlights.

Tirole, J., 1988. The Theory of Industrial Organization. MIT Press.

Wealthfront, 2016. Mutual fund fees. Online document. Visited on 11/6/2016.
URL `https://www.wealthfront.com/research/mutual-fund-fees`

Wermers, R., 2000. Mutual fund performance: An empirical decomposition into stock-picking talent, style, transaction costs, and expenses. Journal of Finance 55 (4), 1655–1703.

# A Extra information on fees

## A.1 Account fees

Many investors hold their retirement savings in some kind of wealth management account (including brokerage accounts, managed accounts, or defined-contribution retirement plans). Typically, these accounts pay some kind of "wrap fee" or account management fee (typically a fixed percent of assets in the account). These fees cover basic account administration costs and, in many cases, premium services such as financial advisory. Advisory can also be provided as a separate service from the account, for a separate fee.

Part of the advisory service is covered by mutual fund distribution fees (i.e., 12b-1 fees) that are already included in the expense ratio. However, investors buying shares from mutual fund brokers may incur "load" fees, e.g., upfront fees or fees upon redemption of their shares. These load fees directly pay for the advisory services the broker performs at the time when the mutual fund is selected.[23] Usually, load fees and explicit account management fees do not appear together. In the presence of explicit account fees—including plan management costs paid by the sponsors of defined-contribution employer plans—investors typically have access to no-load funds. Overall, advisory and distribution fees (excluding 12b-1 fees, which are already included in the net expense ratio) average about 50 bps (Bogle, 2014).

## A.2 Mutual fund fees

The net expense ratio includes three types of costs. First, paperwork costs: custodial fees, legal fees, record-keeping fees, etc. These fees typically cover the cost of inevitable services provided by third parties unaffiliated with the mutual fund. The second type of costs are distribution and service fees (so-called 12b-1 fees). 12b-1 fees cover two types of expense: distribution costs, i.e., commissions to the sales force (capped at 75 bps), and shareholder servicing costs, e.g., cost of providing internet access to fund filings, etc. (capped at 25 bps).[24] 12b-1 fees are included in the fund's expense ratio and they are taken from the

[23] Both the distribution costs component of 12b-1 fees and sales loads constitute compensation for the broker, rather than the fund manager. However, the two fees are not overlapping or mutually exclusive. 12b-1 fees are taken year after year out of fund assets, loads are directly paid by the investor to the broker. For instance, with a 5% load, an investor giving $100 to the broker is only investing $95. If the fund has 12b-1 fees in addition to loads these fees will be levied upon the $95. The same fund may have multiple classes of shares. According to Morningstar's Glossary, "In a typical multi-class situation, the class A fund has a front-end load and either a 0.25% distribution fee or a 0.25% service fee. Class B shares usually have a contingent deferred sales charge and a corresponding 0.75% 12b-1 fee, plus a maximum 0.25% service fee. [...] Class C shares customarily charge a level load with the same fee structure found in a class B share."

[24] 12b-1 fees are so called after SEC Rule 12b-1 under the Investment Company Act of 1940. FINRA regulations from 1993 establish the caps on these fees. See SEC > Mutual Funds Fees and Expenses

fund's NAV. Third, the net expense ratio includes asset management advisory fees, i.e., the actual revenue of the money management company that sponsors the fund in the first place.[25]

# B Appendix: Full discussion of the model

The model features a two-period economy. When individuals are young (time $t$), they work, produce, consume, and save for retirement by investing via mutual funds in a mix of government bonds and financial assets. When they are old (time $t'$), they receive passive retirement income, consume, and die without bequest. Mutual funds require a fixed amount of labor to operate, and therefore must pay a competitive salary to individuals who otherwise would work to produce goods.

## B.1 Individuals

Individual $i \in \mathcal{I}$ lives two periods and leaves no bequest. In period $t$ the individual works, saves, and allocates the savings. The individual starts with a net worth of 0 and is endowed with one unit of labor to spend either producing consumption goods, or managing a mutual fund. The production technology is linear: if the individual allocates $L$ units of labor to the production of goods, the output is $F(L) = \omega L$. In order to attract labor, therefore, mutual funds must pay a competitive wage $\omega$. The individual's total supply of labor is inelastic. Overall, the individual's pretax income is $\omega$ regardless of how labor is allocated. Income from work is taxed at rate $\tau_L$. In period $t'$, the individual retires and depletes all the savings.

The individual draws utility from current consumption ($C$) and from discounted future consumption ($\delta \cdot C'$). In addition, the individual prefers a fund of type $\iota_i$, and derives disutility that increases with the distance between the chosen fund type and one's own preference ($d_{i,j} \equiv |\iota_j - \iota_i|$). The individual saves an amount $S$. A fraction $a$ of the individual's savings is invested in financial assets paying a return $\rho$, and the remainder in government bonds

(https://www.sec.gov/answers/mffees.htm).

[25] Typically, advisory fees are not set at arm's length because the fund is a captive customer of the management company. It is generally believed that market forces curb excessive advisory fees, because of the threat of investors withdrawing their money and taking it to a different fund (e.g., Coates and Hubbard, 2007). Others contend that market forces are not sufficient to keep fees in check because no fund's fees are set at arm's length; even if a fund's fees appear "reasonable" with respect to the competition, they need not be reasonable overall (Freeman et al., 2008). The Supreme Court (*Jones et al. v. Harris Associates L.P.*, 2010) rejects the "market" argument, in part because it is conscious of the lack of arm's length prices, arguing instead in favor of the "workable standard" set in the *Gartenberg* case, i.e., that in order for high fees to be evidence of breach of fiduciary duty, they must be so disproportionately high that they bear no resemblance to the services provided and could not be the result of arm's length bargaining. Evidence of breach of fiduciary duty must otherwise be found in the process by which the mutual fund board has reviewed the advisor's fees.

paying a return $r$. The government grants the individual a deduction for savings at a rate $\tau_S$. All investment is carried out via mutual fund $j$ charging a proportional fee $f_j S$. Thus, the utility of individual $i$ is:

$$U_i = \max_{C,C',S,a,j} \ln C + \delta \ln C' - \gamma d_{i,j} \tag{20}$$

subject to the budget constraints:

$$C = \omega (1 - \tau_L) - S (1 - \tau_S), \tag{21}$$

$$C' = S (1 - f_j) [1 + a\rho + (1 - a) r] (1 - \tau_R). \tag{22}$$

### B.1.1 The individual's savings and asset allocation decisions

The individual decides how much to save, $S$, and the fraction to allocate to stocks, $a$. It is simpler to solve for $a$ first.

The individual's first-order condition with respect to $a$ is

$$\frac{\partial}{\partial a} \ln C + \delta \frac{\partial}{\partial a} \ln C' = 0,$$

i.e.

$$r = \rho. \tag{23}$$

This condition permits us to rewrite the time-$t'$ budget constraint (22) as:

$$C' = S (1 - f_j) (1 + \rho) (1 - \tau_R) \tag{24}$$

Next, the individual's first-order condition with respect to savings $S$ is

$$\frac{\partial}{\partial S} u \ln C + \delta \frac{\partial}{\partial S} \ln C' = 0.$$

Rewrite using the budget constraints (21) and (24) and (23):

$$\frac{\partial}{\partial S} \ln (\omega (1 - \tau_L) - S (1 - \tau_S)) + \delta \frac{\partial}{\partial S} \ln (S (1 - f_j) (1 + \rho) (1 - \tau_R)) = 0$$

to obtain the Euler equation:

$$(1 + \rho) \delta \frac{C}{C'} = \frac{1}{1 - f_j} \frac{1 - \tau_S}{1 - \tau_R}. \tag{25}$$

### B.1.2 The individual's choice of a fund

Individuals choose fund $j$ to satisfy the following criterion:

$$j^* = \arg\max_j \ln C + \delta \ln C' - \gamma d_{i,j}$$

Since $C$ does not depend on $f_j$, simplify

$$j^* = \arg \max_{j \in \{1,2,...N\}} \delta \ln C'(f_j) - \gamma d_{i,j} \tag{26}$$

where $C'(f_j)$ is written in that way to underscore that future consumption depends on the fees paid on one's chosen fund.

## B.2 Mutual funds

Mutual funds can differentiate themselves over one qualitative characteristic, $\iota$, defined on the $[0, 1)$ circle. On this circle, the distance between two points is the shortest possible distance—for instance, the distance between 0.1 and 0.9 is 0.2, not 0.8. Individuals are uniformly distributed over this circle, and their utility is decreasing in the distance from their chosen fund. Every fund $j \in \mathcal{J}$ needs a fixed amount of labor $\varphi$ just to be able to operate. Fund profits are equal to revenue minus cost:

$$\pi_j = f_j Q_j - \varphi\omega,$$

where $Q_j$ is the fund's assets under management, so that $f_j Q_j$ is the fund's total revenue. The fund's problem is to choose $f_j$and $\iota_j$ to maximize $\pi_j$, taking into account that $Q_j$ depends on the fund's choices of pricing $(f_j)$ and location $(\iota_j)$, and taking competitors' choices as given.

Mutual fund profits accrue to their managers and are taxed as other income. However, because of free entry, profits will be zero, so we do not keep track of who the profit accrues to.

### B.2.1 Funds' location and pricing decisions

Fund $j$ sets fees $f_j$ and location $\iota_j$ to maximize profits:

$$\max_{f_j, \iota_j} f_j q_j S - \varphi\omega. \tag{27}$$

Because all funds choose simultaneously whether to enter the market, the problem is symmetric for every fund $j$, which implies that funds will distribute at a constant distance over the circle (Tirole, 1988)

$$d_{j,j+1} \equiv |\iota_j - \iota_{j+1}| = 1/N \qquad \forall j \in \mathcal{J}, \tag{28}$$

and all funds will charge the same fees

$$f_j = f \qquad \forall j \in \mathcal{J}. \tag{29}$$

Finally, because entry is free, the number of funds that choose to enter, $N$, will be such that[26]

$$\pi_j = 0 \qquad \forall j \in \mathcal{J}. \tag{30}$$

## B.3 Government

The government spends an exogenously given amount $G$. This expenditure is inevitable and it does not affect the utility of agents. At time $t$, the government taxes income at a rate $\tau_L$, and grants individuals a deduction for savings at a rate $\tau_S$. The government can also borrow an amount $B$ at the market interest rate $R$ to satisfy the government budget constraint:

$$G = \tau_L \omega - S\tau_S + B$$

To pay off the bonds at time $t'$, the government can tax retirement income at a rate $\tau_R$ so as to satisfy

$$B\,(1+r) = \tau_R S\,(1-f)\,[1 + a\rho + (1-a)\,r]$$

where $f$ is the equilibrium level of fees.

## B.4 Market clearing

### B.4.1 Consumption good market

Total output of goods is equal to consumption plus savings plus government expenditure:

$$C + S + G = \omega L \tag{31}$$

[26] We allow $N$ to be noninteger for simplicity. One could see $N$ as the *expected* number of entrants in a mixed-strategy one-shot entry game in which the realized number of entrants can be constrained to be integer. However, modeling this game does not add any useful insights.

### B.4.2 Labor market

Total labor employed in the production of goods ($L$) and by the $N$ mutual funds ($\varphi N$) must be equal to the working population, i.e., equal to one.

$$L + \varphi N = 1. \tag{32}$$

### B.4.3 Asset management market

Total assets under management by funds are equal to total savings by individuals:

$$\sum_{j \in \mathcal{J}} Q_j = \sum_{i \in \mathcal{I}} S_i = S.$$

Define market share $q_j = \frac{Q_j}{S}$, then the market clearing condition becomes

$$\sum_{j \in \mathcal{J}} q_j = 1 \tag{33}$$

### B.4.4 Government bond market

After fees, the capital available to buy government bonds is $(1-a)\,S\,(1-f)$. The government gives a subsidy $\tau_S$ to saving, and borrows to make up the difference:

$$(1-a)\,S\,(1-f) = B. \tag{34}$$

## B.5 Market solution

### B.5.1 Asset management market equilibrium

In a hypothetical equilibrium there is a marginal investor who is indifferent between funds $j$ and $j+1$. For this individual,

$$\delta \ln C'\left(f_j\right) - \gamma d_{i,j} = \delta \ln C'\left(f_{j+1}\right) - \gamma d_{i,j+1}. \tag{35}$$

Equation (29) says that, in equilibrium, all funds charge the same fees $f$ because of symmetry considerations. Moreover, when fund $j$ chooses $f_j$, it takes all other funds' fees as given. Thus, we can write $f_{j+1}$ as simply the aggregate average level of fees, $f$. Because of the definition of marginal investor, and because of the symmetry of the problem, the distance

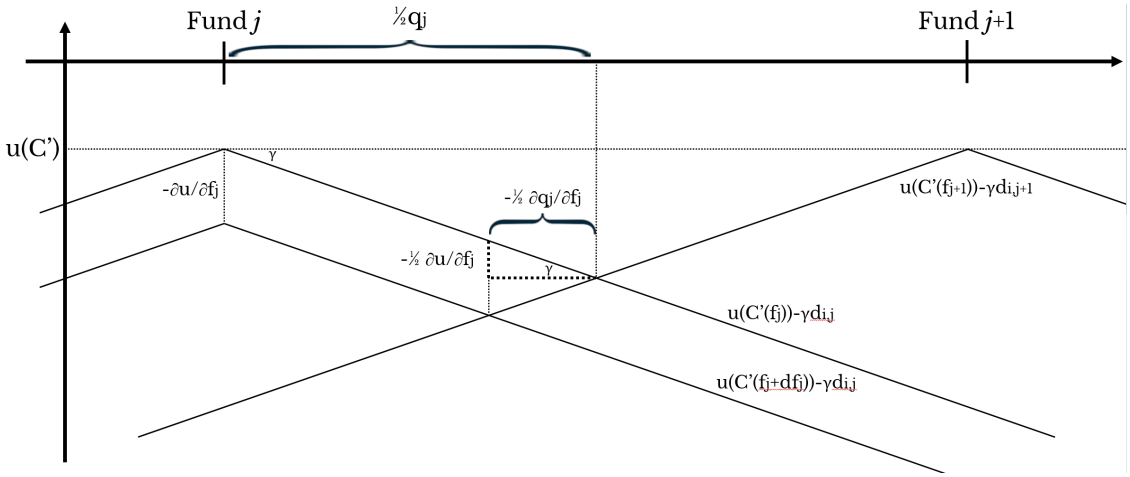


Figure 3: Geometric intuition: competitive fee equilibrium for mutual funds.

between fund $j$ and the marginal investor is equal to one-half $q_j$, the market share of fund $j$:

$$d_{i,j} = \frac{1}{2} q_j.$$

Finally, equation (28) says that the distance between fund $j$ and fund $j+1$ is known to be $1/N$, so

$$d_{i,j+1} = d_{j,j+1} - d_{i,j} = \frac{1}{N} - \frac{1}{2} q_j.$$

Replacing distances in terms of demand within (35) we obtain the following demand function that fund $j$ faces *in equilibrium*:

$$q_j = \frac{1}{N} + \frac{\delta}{\gamma} \left( \ln C' \left( f_j \right) - \ln C' \left( f \right) \right) \tag{36}$$

The geometric intuition behind this result is represented in Figure 3.

Given the fund's objective function (27), we can rewrite the first-order condition for maximization as

$$\frac{1}{N} + \frac{\delta}{\gamma} \left( \ln C' \left( f_j \right) - \ln C' \left( f \right) \right) - f_j \frac{\delta}{\gamma} \frac{1}{1 - f_j} = 0. \tag{37}$$

The optimal level of fees $f_j^*$ solves this condition.

Now, consider the market equilibrium. Because every fund faces the same problem, $f_j^* = f$, so that the second term of (37) cancels out. The equilibrium level of fees of fund $j$, equal to all other funds, is then

$$f_j^* = f = \frac{1}{1 + \frac{\delta}{\gamma} N}. \tag{38}$$

From (30) we know that because of free entry, profits are zero, i.e., revenues are equal to

costs. Thus,

$$fQ = \frac{1}{1 + \frac{\delta}{\gamma}N} \cdot \frac{S}{N} = \varphi\omega \tag{39}$$

### B.5.2 Goods market equilibrium

The individual's budget constraint (21) pins down $S$ as a function of $C$:

$$C = \omega\left(1 - \tau_L\right) - S\left(1 - \tau_S\right)$$

but we also know, from the Euler equation, that

$$C = \left(1 - \tau_S\right)\frac{S}{\delta}$$

and thus we have a fixed, closed-form expression for savings that does not depend on the number of funds $N$:

$$S = \omega\frac{1 - \tau_L}{1 - \tau_S}\frac{\delta}{1 + \delta} \tag{40}$$

Because $S$ does not depend on $N$, total profits of the fund industry are a strictly decreasing function of $N$:

$$\sum_{j=1}^{N}\pi_j = \left(f_j^* Q_j - \varphi\omega\right)N = \frac{1}{1 + \frac{\delta}{\gamma}N}S - \varphi\omega N$$

(the second equality follows from the definition of the equilibrium level of fees $f^*$ and because fund market share $Q_j = S/N$). More funds means more competition with lower fees (first term), and more total fixed costs (second term). Since demand is unvaried, lower revenues and higher costs translates into lower profits. This is not obvious; if industry profits are ever to reach zero for some $N^*$, they have to be a decreasing function of $N$ *past a certain point*, but not monotonically.

### B.5.3 Number of funds

The updated zero-profit condition (39) and the expression for savings (40) help us pin down the parameter of interest: $N$, the equilibrium number of funds under the market solution.

$$\frac{\delta}{\gamma}N^2 + N = \frac{S}{\varphi\omega} = \frac{1}{\varphi} \cdot \frac{\delta}{1 + \delta} \cdot \frac{1 - \tau_L}{1 - \tau_S}. \tag{41}$$

Recall that under the market solution, savings are fixed:

$$S = \omega\frac{1 - \tau_L}{1 - \tau_S}\frac{\delta}{1 + \delta},$$

where $\tau_L$ is the labor income tax rate and $\tau_S$ is the subsidy to saving. In the case of an EET $\tau_S = \tau_L$, and for a TEE $\tau_S = 0$. Therefore, without solving explicitly for $N$, it is already evident that $S$ and $N$, are higher with an EET than they are with a TEE scheme.

## B.6 Planner solution

The planner also takes $G$ as exogenously given, and chooses savings $S$, and number of funds $N$ directly to maximize social utility:

$$U = \max_{C,C',S,N} \ln(C) + \delta \ln(C') - \gamma \bar{d} \tag{42}$$

with $\bar{d}$ the average distance between an investor and their fund. The planner's budget constraints are simply

$$C = \omega(1 - \varphi N) - S - G, \tag{43}$$

$$C' = S(1 + \rho) = (\omega(1 - \varphi N) - C - G)(1 + \rho). \tag{44}$$

The $N$ funds are equally spaced along the circle.[27] Individuals' distance from the nearest fund is uniformly distributed over $[0,\ 1/(2N)]$, hence the density $f(x) = 2N$ for all $x$, and the average distance is (obviously):

$$\bar{d} \equiv \int_{i \in \mathcal{I}} d_{ij} = \int_0^{1/(2N)} x f(x)\, dx = 2N \left[\frac{1}{2}x^2\right]_0^{1/(2N)} = \frac{1}{4N}. \tag{45}$$

Using (43), (44), and (45), the utility function can be rewritten as

$$U = \max_{S,N} \ln(\omega(1 - \varphi N) - G - S) + \delta \ln(S(1 + \rho)) - \frac{\gamma}{4N} \tag{46}$$

The first-order condition with respect to $N$ is

$$C = \frac{4\varphi\omega}{\gamma} N^2 \tag{47}$$

The first-order condition with respect to $S$ is

$$S = C\delta \tag{48}$$

[27] In principle, the planner gets to choose the optimal placement of funds too, but it is easy to show that equal spacing is indeed optimal.

Using (43), rewrite the time-$t$ budget constraint (43) as

$$C = \omega(1 - \varphi N) - C\delta - G$$

$$C = \frac{\omega(1 - \varphi N) - G}{1 + \delta} \tag{49}$$

Finally, using (47) and (49), obtain an implicit expression for the socially optimal number of funds:

$$4\frac{1+\delta}{\gamma}N^2 + N = \frac{1 - G/\omega}{\varphi} \tag{50}$$

# C Appendix: Calibration details

The parameters we need to calibrate are

- $\delta$, the discount rate between working life consumption and retirement consumption.
- $G/\omega$, government spending as a fraction of total output
- $\varphi$, the percentage of the labor force employed by the average mutual fund (or, more intuitively, $\varphi N$, the percentage of labor force employed by the mutual fund industry). Total labor force is normalized to one in the model, so $\varphi$ is expressed as a fraction of total labor force in the economy.
- $\gamma$, i.e. how much people care about convenience in the choice of a mutual fund, vs. consumption.

## C.1 Discount rates

$\delta$ is the one-period discount rate. How long is one period? The first period is working life (~35 years) and the second period is retirement (~25 years). The average number of years between savings and consumption is about 30 years. In order for $\delta$ to be low enough that the equilibrium number of funds is too low, individuals must be unrealistically impatient (e.g. with a 15% per annum discount rate, the 30-year $\delta$ is about 0.01). For our calibration, we adopt a discount rate of 2% per annum, i.e. $\delta = (1.02)^{-30} = .545$ over a 30-year period.

## C.2 Government expenditure

$G/\omega$ is government expenditure as a percent of total output. Note that in this model government expenditure matters only to the extent that it subtracts real resources from the

economy. To exclude transfers, we use the "Real Government Consumption Expenditures and Gross Investments" from the Bureau of Economic Analysis. This expenditure is about 18% of gross domestic product.

## C.3 Mutual fund industry

$\varphi$ can be directly estimated by combining information about the number of funds in the U.S. economy, the total number of workers, and the size of the labor force. According to ICI (2015), in 2013 the asset management industry had employed 166,000 people, and featured 7,713 funds. The Bureau of Labor Statistics reports a total employment of about 130 million people, implying $\varphi = 0.000017\%$.

## C.4 Investors' preferences for convenience

$\gamma$ has no real-world equivalent and thus it can only be calibrated indirectly. However, the level of fees is observable in the real world. In the model, $f$ is the level of fees (per period). We have an equation for fees as a function of $\delta$, $\gamma$, and $N$.

$$f = \frac{1}{1 + \frac{\delta}{\gamma} N}$$

Once again, we need to adjust $f$ to take into account the fact that one period is about 30 years. We choose 0.20% as the minimum cost of money management, based on the expense ratios of large index funds, and we choose a 30-year horizon, implying $f = 1 - (1 - 0.20\%)^{30} = 5.83\%$. We then change $\gamma$ in the model until $f$ is about 6%. This yields $\gamma = 3500$.