What Determines 401(k) Plan Fees?
A Dynamic Model of Transaction Costs and Markups*

Hanbin Yang†

February 13, 2025

Abstract

I study how transaction costs between employer sponsors and 401(k) plan providers affect the level and dispersion of 401(k) plan fees. Despite the introduction of many low-fee funds, providers exploit transaction costs to delay the inclusion of these funds in 401(k) menus. Using a dynamic structural model, I estimate markups based on how employers trade off transaction costs against plan fees. I find that markups constitute 18% of plan fees on average, and the between-employer variation in markups explains 38% of the overall fee dispersion. Policies aimed at reducing transaction costs have limited impact due to providers' invest-and-harvest incentives. Modifying employer preferences or consolidating plans could generate substantial fee savings.

*I am indebted to the guidance and support of John Campbell, Ariel Pakes, Mark Egan, Robin Lee, and Adi Sunderam. I benefited from discussions with Vivek Bhattacharya, Yufeng Huang, Samuel Jäger, Alex MacKay, David Laibson, Pietro Tebaldi, and seminar participants at Boston College Carroll School of Management, Columbia Business School, European Winter Finance Conference, Harvard Finance and IO workshops, International Industrial Organization Conference, London Business School, London School of Economics, and University of Washington Foster School of Business. All errors are my own.

†London Business School. Email: hyang@london.edu

1 Introduction

Employer-sponsored 401(k) plans are a critical component of retirement savings in the US, with over $7 trillion assets and 60 million active participants as of 2021 (ICI, 2022a). Given the importance of 401(k) plans, regulators and policymakers have long been concerned about protecting the interests of employee participants, a concern that dates back to the Employee Retirement Income Security Act of 1974 (ERISA). Under ERISA, employer sponsors of 401(k) plans are required to act as fiduciaries in the best interests of their plan participants.

Using comprehensive data on 401(k) plans, I document a large fee dispersion. In 2019, an employer at the 90th percentile offers a plan with over 70 basis points (bps) higher fees than an employer at the 10th percentile. This fee dispersion has led to lawsuits in recent years, including *Hughes v. Northwestern*, which reached the US Supreme Court in 2022. In these lawsuits, participants cite high fees as evidence that their employers have violated their fiduciary duties, arguing that employers allow providers to impose unreasonably high markups. Conversely, employers attribute high fees to differences in the scope and quality of services, suggesting that high fees may reflect higher production costs associated with more valuable plan features. Without directly observing cost structures or contract terms, separating the cross-section dispersion in fees into markups or production costs can be difficult. Moreover, using this dispersion to assess whether employers fulfill their fiduciary standards is even more challenging.

In this paper, I examine how employers' transaction costs affect 401(k) plan fees and then estimate markups using how employers trade off transaction costs against plan fees. Over the past two decades, plan fees have experienced a secular decline, primarily driven by the introduction of new, low-cost mutual funds. Due to transaction costs associated with selecting and switching providers, providers can delay menu updates to include these funds, thus earning increasingly higher markups relative to the secular decline. Motivated by these patterns, I develop and estimate a structural model that integrates employers' provider choice, transaction costs, and providers' dynamic fee competition. To account for employer heterogeneity, I introduce a fee sensitivity parameter that captures how employers trade off transaction costs against plan fees. This parameter reflects an employer's sophistication or negotiation effort and is inversely related to markups under providers' optimal fee-setting conditions. Based on model estimates, markups constitute 18% of plan fees on average, or 11 bps. The variation in markups between employers with different fee sensitivities explains 38% of the overall fee dispersion.

I begin my analysis by documenting patterns in 401(k) fees using a comprehensive sample of 401(k) plans from 2009 to 2019. The dataset includes annual plan-level information on investment menus, allocations, plan providers, and services offered. As of 2019, average plan fees are 65 bps over plan assets and consist of two components. First, participants pay mutual fund fees, with a plan-level asset-weighted average of 47 bps. Second, participants pay recordkeeping fees to cover administrative and advisory services, which are on average 18 bps over plan assets. Total plan fees vary substantially across employers, with a standard deviation of 24 bps, and 10% of employers having fees above 100 bps. Observed characteristics, such as menu composition, explain less than 25% of the fee dispersion.

Then, I present motivating evidence of how employers' transaction costs influence 401(k) fees. Employers face two main sources of transaction costs when dealing with providers. First, employers incur costs when procuring services from plan providers. Employers use a Request

for Proposal (RFP) process to solicit customized bids from plan providers, which can be time-consuming. Second, employers face costs when switching providers. Transitioning to a new provider requires setting up new infrastructure and may disrupt participants' experience. I find that employers reduce fees by over 10 bps, or 16% relative to the 2019 average, when they switch providers. However, only 3% to 5% of employers switch annually, consistent with transaction costs. In addition to monetary and time costs, transaction costs can also reflect inertia. I do not differentiate between the sources of transaction costs in my analysis.

Providers can exploit transaction costs to increase markups by delaying menu updates. During my sample period, many newly available funds were comparable to existing ones but charged lower fees. I show that mutual fund fees in 401(k) plans decrease by on average 30 bps from 2009 to 2019. Only 30% of this decline comes from fee reductions among existing funds included in 2009 plan menus. The majority of this decline results from menu updates, where new, lower-cost funds replace existing ones. Providers naturally prefer higher fees generated by existing funds and can take advantage of employers' transaction costs to delay menu updates. I find that over 70% of the funds in employers' 2019 menus are available as early as 2009, and these funds could have reduced fees in 2009 by almost 20 bps.

Building on these findings, I develop a dynamic structural model of transaction costs in a negotiated-price market. Since providers offer customized services and negotiate fees separately with each employer, I treat each employer as an independent market and model how multiple providers compete to serve a single employer. The model comprises three key components. First, the employer determines whether to conduct a costly RFP for potential fee reductions, as an optimal stopping problem (Rust, 1987). Second, the employer chooses a provider at an RFP, where the employer pays switching costs if choosing a non-incumbent provider. I model provider choice using a random-coefficient discrete choice model (Berry et al., 1995). Lastly, providers compete in a dynamic fee-setting game (Dubé et al., 2009; Cabral, 2016; Sweeting et al., 2022), where forward-looking providers may initially set lower markups, anticipating future increases due to transaction costs. I jointly solve these three components and compute outcomes under the Markov perfect equilibrium.

In my model, providers set higher markups when the employer has lower fee sensitivity. Conceptually, this fee sensitivity parameter reflects the employer's sophistication or effort, capturing the employer's preference for minimizing fees charged to participants relative to their own transaction costs. To identify this parameter, I face a common challenge in negotiated-price markets that I do not observe fees from non-chosen providers. I overcome this challenge by leveraging dynamic fee variation. Specifically, fee reductions when employers switch providers, along with how fees transition over time, can approximate the differences in proposed fees between incumbent and non-incumbent providers. A small fee reduction from switching indicates that the employer is highly fee-sensitive, preventing the incumbent provider from exploiting switching costs to impose high markups. [1]

I estimate my model separately for employers in each quartile of plan assets, and I recover reasonable estimates of transaction costs and markups. Transaction costs measured in basis points over plan assets decrease in employer size. From small employers in the first quartile to large employers in the fourth quartile, RFP costs range from 10 to 1 bps, while switching costs

[1] I find that employers with more plan assets, more participants, higher employer contribution, higher participation rates, and higher ESG scores have lower fee reductions from switching providers, suggesting this variation is informative of employers' sophistication or effort captured by their fee sensitivities.

range from 46 to 2 bps. Scaled by total plan assets, RFP costs range from $7,000 to $18,000, and switching costs from $30,000 to $57,000. On the other hand, I estimate that markups account for 18% over plan fees on average. Larger employers tend to have higher fee sensitivities and lower markups. Specifically, markups for small employers are 18 bps or 23% over plan fees, decreasing to 6 bps or 13% for large employers. My estimates also suggest economies of scale, with production costs declining from 61 bps to 42 bps from small to large employers, explaining two-thirds of the difference in plan fees across employer sizes.

Using recovered markups, I decompose the dispersion in 401(k) plan fees. Heterogeneous fee sensitivities generate different long-run averages of markups across employers, accounting for 38% of the fee dispersion. Within a cross-section, employers incur transaction costs at different points in time, leading to markup variation that accounts for 13% of the fee dispersion. The remaining 48% is attributed to production costs, including heterogeneous menu designs and services as well as economics of scale. This decomposition indicates that a large fraction of the dispersion in 401(k) plan fees is unrelated to fee sensitivity, which reflects employer sophistication or negotiation effort. As a result, fees may appear high even when employers are plausibly meeting their fiduciary duties.

Having estimated the model, I evaluate counterfactual policies aimed at mitigating different determinants of plan fees. First, I examine the effect of increasing employers' sophistication or effort by setting fee sensitivities at the 90th percentile of their distribution. I find that higher fee sensitivity reduces plan fees by 10 bps for the average employer, saving participants $800 million per year. [2] As employers prioritize fees over transaction costs, they conduct RFP and switch providers more often, resulting in $160 million higher annual transaction costs. The pressure coming from recent lawsuits may encourage employers to increase their fee sensitivities, which can reduce fees for participants but also impose a higher burden on employers.

Second, I consider plan consolidation since larger employers have lower markups, production costs, and transaction costs. Specifically, I consolidate plans of smaller employers to match the average plan assets of employers in the fourth size quartile. Consolidation can reduce fees by on average 16 bps or $400 million annually, through reductions in both markups and production costs. Additionally, employers save $110 million in transaction costs by eliminating duplicated RFPs and provider switches. This counterfactual suggests that recent policies facilitating multiple employer plans (MEPs) under the SECURE Act of 2019 could benefit both participants and employers.

Finally, counterfactual policies that reduce transaction costs have more complicated effects on plan fees. As highlighted in previous literature (Klemperer, 1995; Dubé et al., 2009; Cabral, 2016), transaction costs influence fees through two opposing mechanisms. On one hand, providers exploit the transaction costs of their existing employers by charging higher markups. On the other hand, anticipating higher future markups, providers initially set lower markups when competing for employers. I find that when both transaction cost parameters are set to zero, fees remain roughly the same as the status quo, as providers raise fees set at RFPs in the new equilibrium. Interestingly, moderate transaction costs incentivize providers to compete more aggressively, reducing plan fees by 4 bps or $500 million annually.

I make four contributions in this paper. First, I contribute to the literature on how frictions

[2] In Section 3.2, I discuss several restrictions when constructing my estimation sample and report aggregate fees in my sample only. The overall impact of these counterfactual can be three to five times larger when applied to the full universe of 401(k) plans with comparable plan assets.

affect retirement plan designs and fees by analyzing the impact of transaction costs.[3] Pool et al. (2016, 2021) show how providers steer plan menus toward their proprietary funds and funds with revenue-sharing. Doellman and Sardarli (2016) and Badoer et al. (2020) show how providers trade off direct compensations from sponsors with indirect compensations through revenue sharing. Chalmers and Reuter (2020) and Reuter and Richardson (2022) study the demand for financial advice and the impact of advice on asset allocation in 401(k) plans. Bhattacharya and Illanes (2022) model misaligned incentives of employers and imperfect competition among providers. My empirical framework incorporates many of the frictions analyzed by these studies, and I focus on the role of transaction costs.

Second, I contribute to the literature on fee dispersion in financial markets. Prior studies primarily focus on fee dispersion within the same or homogeneous products, such as index funds (Hortaçsu and Syverson, 2004) and private equities (Begenau and Siriwardane, 2022). In contrast, 401(k) plans are potentially heterogeneous, with both markups and production costs contributing to fee dispersion. To disentangle these components, I apply methods in empirical industrial organization to recover markups from demand estimation.[4] A common challenge in negotiated-price markets is that fees set by non-chosen providers are unobserved. I address this challenge by developing novel estimation and identification strategies.[5]

Third, my counterfactual analysis provides insights into recent regulations in 401(k) plans. My paper is closely related to Bhattacharya and Illanes (2022), who also study how counterfactual regulations can enhance plan quality by addressing different frictions. Similar to their findings, I find that modifying employer preferences can lead to substantial fee savings.[6] Compared to their results, I estimate stronger economies of scale, suggesting that multiple employer plans can achieve larger benefits through reducing production costs, even without changing employer behaviors.

Lastly, this paper contributes to the literature on dynamic competition and switching costs. Klemperer (1995) and Farrell and Klemperer (2007) provide early summaries of this literature. More recently, Dubé et al. (2009) and Cabral (2016) show that the equilibrium impact of switching costs on fees is ambiguous, and small switching costs can reduce equilibrium fees. I contribute to this literature by empirically quantifying the impact of both RFP and switching costs in the 401(k) setting. In addition, I solve a dynamic game with both forward-looking demand and supply, which is rare in empirical dynamic competition models.

The rest of the paper is structured as follows. I discuss institution details in Section 2, data

[3] More broadly, a large literature in household and behavioral finance focuses on employees' participation and contribution. See Benartzi and Thaler (2007) and Choi (2015) for a discussion of this literature. Papers in this literature have studied the effects of plan design such as automatic enrollment (Madrian and Shea, 2001; Beshears et al., 2009; Choi et al., 2007; Carroll et al., 2009) and firm matching (Choi et al., 2002; Duflo et al., 2006; Dworak-Fisher, 2011). Yogo et al. (2025) study access to retirement plans and participation across US households using administrative data.

[4] See Clark et al. (2021) for an overview of the broader literature on the industrial organization of financial markets. Previous works in this literature have studied car loans (Einav et al., 2012; Grunewald et al., 2020), credit cards (Nelson, 2018), insurance (Koijen and Yogo, 2016), mortgages (Allen et al., 2014, 2019; Robles-Garcia, 2019), municipal bonds (Brancaccio et al., 2017), pensions (Luco, 2019; Illanes, 2016; Illanes and Padi, 2019), financial advice (Bhattacharya et al., 2019; Egan, 2019; Guiso et al., 2022), and consumer and student loans (Bachas, 2018; Cuesta and Sepúlveda, 2021). Egan et al. (2022, 2023) estimate investor demand to recover beliefs and risk preferences.

[5] Allen et al. (2019); Allen and Li (2020); Cuesta and Sepúlveda (2021) use results from second-price auctions to infer the distribution of unobserved fees. It is difficult to adopt this approach in my setting because auction results generate expected fees but the value functions in my dynamic model depend on the actual realizations of fees.

[6] My fee sensitivity parameter is similar to their willingness to pay for plan quality parameter.

and summary statistics in Section 3. In Section 4, I present motivating evidence that employers face transaction costs and how transaction costs allow providers to charge markups. I introduce the structural model in Section 5, explain estimation and identification in Section 6, and present estimation results in Section 7. I present policy counterfactuals in Section 8. Lastly, Section 9 concludes.

2 Institutional Background

This paper examines the interaction between employer sponsors of 401(k) plans and plan providers. More broadly, there are four main types of agents in the 401(k) sector: employers, employee participants, plan providers, and mutual fund companies. In the following, I introduce these agents and discuss the rationale for focusing on employer-provider interactions. Then, I discuss allegations in excessive fee lawsuits and sources of transaction costs in the employer-provider interaction.

2.1 Employer Sponsors of 401(k) Plans

Private sector employers sponsor 401(k) plans, allowing employees to make tax-deferred contributions from their salaries into these accounts. Employees then invest these contributions in various investment vehicles to save for retirement.[7]

The Department of Labor regulates employer sponsors under the Employee Retirement Income Security Act of 1974 (ERISA). ERISA imposes fiduciary responsibilities on employer sponsors, requiring them to act solely in the best interest of their plan participants. Employers must exercise prudence in selecting and reviewing investment options in their plan menus, including due diligence concerning risk, performance, fees, and diversification. Fiduciary duties also extend to the selection of plan providers and the assessment of their performance.

2.2 Plan Providers

Often referred to as recordkeepers or trustees, financial service providers administer 401(k) plans and design plan menus. They process employee contributions and distributions, set up online portals for participants to access account details and execute transactions, and may also offer educational resources to help participants with investment decisions. In Figure 1a, I display market shares based on the number of plans and assets for providers as of 2019. Some of these providers are asset managers, such as Fidelity and Vanguard.[8] Empower, ADP, and Principal Financial Group are third-party providers. They do not have proprietary investment products but may offer other bundled services. For example, ADP offers HR payroll services, while Principal Financial Group offers insurance products. Many banks, insurance companies, and asset managers also provide recordkeeping services. Due to their relatively small market shares,

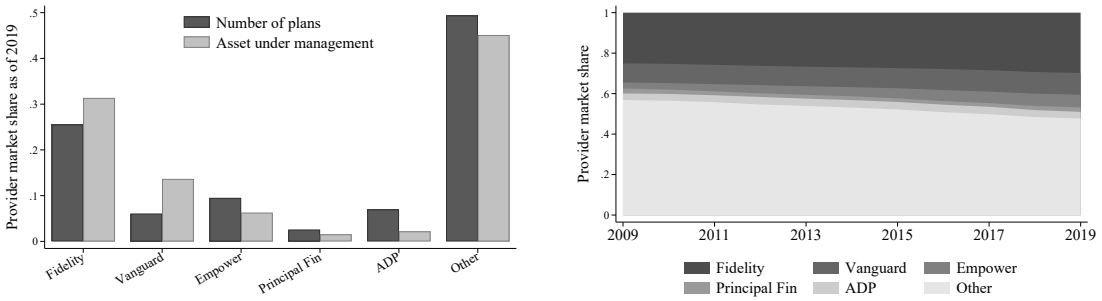
[7] A 401(k) is a type of defined contribution (DC) plan. Other employer-sponsored DC plans include 403(b) plans for employees in education and nonprofit organizations, 457 plans for certain governmental employees, and Thrift Savings Plan (TSP) for federal government employees. My data and analysis include 401(k) and 403(b) plans.

[8] When designing plan menus, recordkeepers who are asset managers typically favor their proprietary funds but may also offer funds from other mutual fund companies. The fraction of proprietary funds among Vanguard and Fidelity plans are around 90% and 50% on average, according to Table A3.

I group them collectively into the "other" category. In Figure 1b, I show that large providers have maintained stable market shares throughout my sample period.[9]

Figure 1: Provider Market Shares

(a) Market share as of 2019 (b) Market share over time



Panel (a) plots market shares of large providers both in number of plans and total assets as of 2019, where smaller providers are grouped into the other category. Panel (b) plots the market share of providers over time, where I restrict to a balanced panel.

2.3 401(k) Plan Fees, Participants, and Fund Managers

401(k) plan fees consist of investment fees and recordkeeping fees. First, employee participants pay investment fees on their selected funds, such as mutual fund expense ratios, to the asset managers who operate those funds.[10] Second, employee participants (and occasionally employers) pay recordkeeping fees to plan providers. Recordkeeping fees can be structured as either a fixed dollar amount per participant or a percentage of invested assets. Deloitte (2013) shows that investment fees make up 82% of total plan fees, with participants covering 87%.

In my analysis, I abstract away from the interaction between employers and their participants, assuming employers choose plan providers and negotiate plan fees on behalf of their participants. Although individual participants may pay different fees based on their fund allocations, their choices are limited by the plan menus. As a result, employers' decisions still play a significant role in determining the fees that participants pay.

On the other hand, since investment fees are paid to fund managers, there is a vertical relationship between recordkeepers and fund managers. I abstract away from analyzing this vertical relationship. Integrated providers that are asset managers themselves (e.g., Fidelity and Vanguard) receive investment fees directly. Third-party providers indirectly obtain a portion of investment fees through "revenue sharing" with fund managers. I assume plan providers maximize joint profit and can use "revenue sharing" or other transfers to align incentives with fund managers.[11]

[9] There are several consolidations among large recordkeepers after the end of my sample. For example, Empower acquired MassMutual's retirement businesses in 2010 and Prudential's in 2022. Consolidations in my sample are between smaller providers and are adjusted accordingly, as detailed in Appendix A.

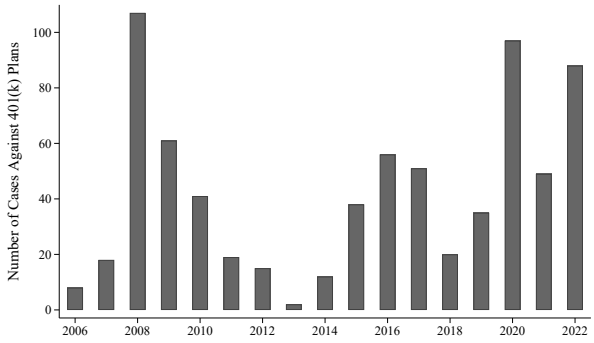
[10] A mutual fund can have multiple share classes with different expense ratios. Investors across different 401(k) plans pay the same expense ratio for the same share class.

[11] Pool et al. (2016, 2021) and Bhattacharya and Illanes (2022) study how vertical integration and revenue sharing create misaligned incentives in menu design.

2.4 Excessive Fee Lawsuits

In Figure 2, I show the number of lawsuits bought against employer sponsors of 401(k) plans each year. The number of lawsuits peaked during the Great Recession when most cases focused on inappropriate investment choices that led to losses in participants' retirement savings. Recently, lawsuits have surged again, focusing on "excessive fees". I discuss three main allegations and highlight the challenges of using fee comparisons to determine whether employers fulfill their fiduciary duties.

Figure 2: Number of Cases Against Employer Sponsors of 401(k) Plans



Statistics based on Mellman and Sanzenbacher (2018); Aronowitz (2023)

Offering retail share classes with higher expense ratios instead of institutional share classes of the same mutual funds. According to the plaintiffs, offering a more expensive share class of the same mutual fund suggests employers are negligent in monitoring their plan menus. However, although retail share classes have higher expense ratios, they may include higher revenue-sharing to offset recordkeeping fees paid by participants. On a net basis, a retail share class could result in lower overall fees compared to the institutional share class of the same fund. In *Forman v. TriHealth*, the court dismissed claims that the employer violated fiduciary duties by selecting a more expensive share class of the same mutual fund. In addition, to satisfy higher minimum investment thresholds of institutional share classes, employers may need to limit the number of funds offered, thereby restricting investment options for participants. [12]

Offering actively managed funds instead of the index funds with lower expense ratios. Similar to the previous claim, plaintiffs argue that selecting more expensive active funds suggests negligence. While active funds may underperform their index counterparts (Gruber, 1996), no regulatory guidance requires employer sponsors to offer index funds. In *Smith v. CommonSpirit Health*, the court ruled that active funds are a common component of retirement plans and employers should offer active funds as an option for risk-tolerant investors among plan participants.

Allowing plan providers to charge higher recordkeeping fees than comparable employers. Similar to claims regarding share classes, plaintiffs argue that higher recordkeeping fees indicate employer negligence in monitoring and negotiating plan fees. Conversely, employer sponsors tend to argue that they offer a different scope or quality of services than the employers used

[12] See detailed argument in the Investment Company Institute's Amicus Brief https://www.supremecourt.gov/DocketPDF/19/19-1401/198014/20211028135053555_19-1401%20ICI%20Amicus%20Brief.pdf

as benchmarks in participants' comparisons. Plaintiffs typically compare their recordkeeping fees with those of five to ten employers of similar sizes and rarely provide direct evidence that another provider would offer a lower fee for identical services. In *Matousek v. MidAmerican Energy Company*, *Smith v. CommonSpirit Health*, and *Albert v. Oshkosh Corp.*, courts ruled that plaintiffs failed to present sufficient evidence that employer sponsors allow higher recordkeeping fees for comparable services.

I do not have direct evidence on the extent of service differentiation across providers or employers. Industry practitioners tend to argue that such differentiation could exist (Aronowitz, 2022b). As indirect evidence, survey results from Deloitte (2019) indicate that when asked to rank top improvements from recordkeepers, 18% of employers cited participant readiness for retirement, 15% mentioned plan sponsor websites or tools, and 11% prioritized participant experience. In contrast, only 10% and 7% identified direct recordkeeping fees and investment fees, respectively, as top concerns. The fact that more employers focus on service quality over fees suggests that they may not view recordkeeping as a homogeneous service.

2.5 How Employers Choose and Monitor Providers

Employers incur transaction costs when choosing providers and when switching to a different provider. First, employers choose providers by running a Request for Proposal (RFP), a standard procedure in private-sector procurement. The RFP process can be complex and time-consuming. Employers may hire external consultants to navigate the RFP process, leveraging their industry knowledge and legal expertise. Potentially due to the time costs of handling RFPs and the monetary costs of hiring consultants, employers typically conduct an RFP every three to five years.[13]

Second, switching providers can create administrative burdens and disrupt participants' experience. Employers also value their overall relationships with their current providers, especially since their providers may offer other services.[14] According to Deloitte (2015, 2019), 45-50% of employers have maintained the same recordkeepers for over 10 years, with the primary reason being "overall relationships". In addition, there might be contractual restrictions or penalties associated with switching providers. In *Hughes v. Northwestern*, a significant portion of Northwestern University's plan asset is invested in TIAA annuity, with TIAA also serving as the recordkeeper. The annuity has specialized administrative requirements and substantial early withdrawal penalties. As a result, it would be highly costly for Northwestern University to switch away from TIAA.

After selecting a provider, employers need to monitor the provider's performance. Employers typically schedule annual or quarterly review meetings with providers, where they may discuss potential changes to plan menus or other services. Employers may also hire external consultants to benchmark their fees against a comparable peer group or conduct Requests for Information (RFIs) to solicit fee quotes from other potential providers. Fee comparison results from

[13] In two polls conducted by the National Association of Plan Advisors (Adams, 2020, 2022), 50% and 61% plan advisors recommend that employers conduct an RFP every 3-5 years. See Figure A7 for a sample RFP timeline. There is no specific legal requirement to conduct RFPs. Although some plaintiffs argue that employers breach their fiduciary duties because they do not regularly solicit competitive bids with RFPs, courts tend to dismiss such claims, including in *Sacerdote v. New York University* and *Albert v. Oshkosh Corporation*.

[14] According to Deloitte (2013), only 56% of employers in their survey reported having no other relationships outside their 401(k) plans.

benchmarking or RFIs can help employers negotiate fees with their current providers. However, since providers offer customized services to each employer, benchmarking or RFIs may not fully capture the fees employers could obtain if they were to conduct a formal RFP.

3 Data and Summary Statistics

3.1 Data

The primary dataset for this study comes from BrightScope Beacon. BrightScope Beacon provides plan and fund level information for ERISA defined contribution plans, covering 84% of total plan assets. BrightScope collects data from plan sponsors and publicly available sources, including the Department of Labor's Form 5500 and the Securities and Exchange Commission (SEC). The dataset covers 70,000 401(k) plans from 2009 to 2019. For each 401(k) plan, BrightScope reports annual data on investment menus and the total assets allocated by participants to each investment option. The data also contain fund asset class classifications (e.g. large cap equities, bonds) and investment styles based on Morningstar categories.

BrightScope collects the identities of service providers, the service codes that allow me to identify the primary recordkeeping provider, and the total dollar amount of direct compensation. However, I do not observe whether recordkeeping fees are paid by employers or employees, or whether recordkeeping fees are structured as a dollar amount per participant or as a percentage of plan assets. To be consistent with investment fees, I measure direct compensation as a fraction of total plan assets. Third-party recordkeepers may also receive indirect compensation from fund managers in exchange for including their mutual funds on plan menus. To avoid double-counting expense ratios, I exclude indirect compensation from my analysis. Occasionally, some plans have multiple recordkeepers. In such cases, I designate the provider with the highest compensation or the longest tenure as the primary recordkeeper. See Appendix A for further discussion. It is worth noting that direct compensation data is noisy and may overestimate recordkeeping fees (Aronowitz, 2022a).

While I observe providers, I do not have data on when employers conduct RFPs to choose providers.

3.2 Sample Construction

I merge investment menu data from BrightScope with historical expense ratios from CRSP. To construct my sample, I focus on employers that have at least five mutual funds on their menus and on average over 70% of assets allocated to mutual funds. This step eliminates some large corporations that typically request providers to customize funds only for their plan participants. While I observe these customized funds in my data, their historical expense ratios are not available. Additionally, I exclude employers with fewer than 100 participants. These employers are more likely to outsource the management of their 401(k) services and also delegate part of their fiduciary duties to providers.

Table A1 shows that my estimation sample represents 17% of 401(k) plan assets. The sample is less representative of the smallest employers due to missing menu information and of the largest employers because they are more likely to offer customized funds. My sample is

more representative of employers in the 25th to 75th percentiles of the size distribution, with plan assets ranging between $10 million and $40 million.

My dataset is an unbalanced panel. Of the 19,037 employers in my estimation sample as of 2019, 2,476 or 13% have data for each year. More employers began reporting Form 5500 to the Department of Labor over time. Among employers who consistently report Form 5500, plan menu and recordkeeping compensation are sometimes missing. I use the full estimation sample for structural estimation and restrict to the balanced panel for certain analyses to avoid selection biases.

3.3 Summary Statistics

In Table 1, I present summary statistics for employers in my estimation sample as of 2019. The median employer has approximately 300 participants with $20 million in total plan assets. After excluding large corporations that offer customized funds, my estimation sample includes $1.1 trillion assets. On average, investment fees are 47 bps. This average is computed by calculating asset-weighted mutual fund expense ratios for each employer and then taking a simple average across employers. Direct recordkeeping fees average 17 bps relative to total plan assets. Combining investment and direct recordkeeping fees, total plan fees are 65 bps on average. The dispersion in plan fees is large, with a standard deviation of 27 bps. Plan fees range from 30 bps at the 10th to 100 bps at the 90th percentile.

Table 1: Summary Statistics of 401(k) Plans in 2019

	Num Obs	Mean	St.Dev.	Pct10	Pct50	Pct90
Plan participants	19,037	875	2,990	130	288	1,600
Total plan assets (million)	19,037	58	236	5	18	109
Average account balance (thousand)	19,037	86	91	17	59	181
Fee						
Plan average expense ratio (bps)	19,037	47	24	17	45	79
Direct recordkeeping fees (bps)	19,037	17	19	1	11	44
Total plan fees (bps)	19,037	65	27	31	62	101
Plan menu						
Number of funds	19,037	26	9	16	26	35
Menu turnover (%)	19,037	13	20	0	5	39
Provider switch						
Switched provider (%)	19,037	4	20	0	0	0

Summary statistics across plans as of 2019.

On average, employers offer 26 funds in their plan menus. To measure menu updates, I use menu turnover, defined as the fraction of funds added or replaced in a given year. On average, 13% of funds in plan menus are updated annually. Additionally, 4% of employers switch their providers in 2019.

In Table A3, I present summary statistics by providers. Fidelity and Vanguard typically serve larger employers, whereas ADP primarily targets smaller employers. Vanguard charges the lowest expense ratios, in part because it offers a higher proportion of index funds. Despite similar index fund allocations, third-party providers such as ADP and Principal Financial Group offer funds with higher expense ratios than Fidelity. Vertical integration allows Fidelity and

Vanguard to offer more proprietary funds and charge lower recordkeeping fees. Based on my conversations with industry practitioners, although integrated providers charge lower plan fees, they tend to promote their brokerage accounts and other products to participants, which can negatively impact participant welfare.

4 Motivating Evidence

I present two sets of motivating evidence to demonstrate that employers face transaction costs and how providers can exploit transaction costs to earn markups.

4.1 Employers Face Transaction Costs

I demonstrate that employers rarely switch providers but can substantially reduce fees when they switch, consistent with transaction costs associated with conducting RFPs and switching providers, as discussed in Section 2.5.

In Table 1, I show that around 4% of employers switch providers each year.[15] I then use a difference-in-differences specification to estimate fee reductions when employers switch providers. Using a balanced panel of employers, I define the treatment group as employers that switched providers once between 2012 and 2016, while the control group consists of employers that never switched providers. The specification is as follows:

$$fee_{ijt} = \sum_{s \neq -1} \theta_s 1\{t = s\} + \Gamma_i + \Gamma_j + \tau_{jt} + X_{it}\beta + \epsilon_{ijt} \tag{1}$$

The dependent variable is the plan fee, which includes both investment and recordkeeping fees, of employer i at year t with provider j. s corresponds to the number of years relative to the year of switch, with $s = -1$ assigned to employers that never switch providers. The coefficients of interest are denoted by θ_s, and I normalize $\theta_s = 0$ for $s = -1$. I control for employer and provider fixed effects Γ_i, Γ_j, provider-specific time trend τ_{jt}, and plan characteristics X_{it} which include menu composition, types of services, and number of plan participants.

I plot the coefficients θ_s in Figure 3. There is an immediate reduction of over 10 bps of plan fees when employers switch providers. This reduction is around 16% relative to average plan fees in 2019 shown in Table 1. The large fee reduction combined with a low switching probability suggests substantial transaction costs. After switching providers, fee reductions are persistent but attenuate by half over three years. Even without switching providers, employers in the control group may reduce their fees by inviting other providers to compete with their existing providers during RFPs. Since employers are unlikely to conduct another RFP immediately after switching providers, plan fees tend to increase relative to the control group.

One concern is that employers may change their plan characteristics when they switch providers. My controls for menu composition and services help mitigate this concern. In Figure A5c, I estimate the same difference-in-differences specification where the outcome variable is computed using average expense ratios across all funds of the same asset class × Morningstar category. This measure reflects changes in menu composition, which has a minor and almost

[15] In Figure A5a, I show a time-series of the fraction of employers who switch providers each year, which is between 3% and 5%.

Figure 3: Difference in Differences When Employers Switch Providers

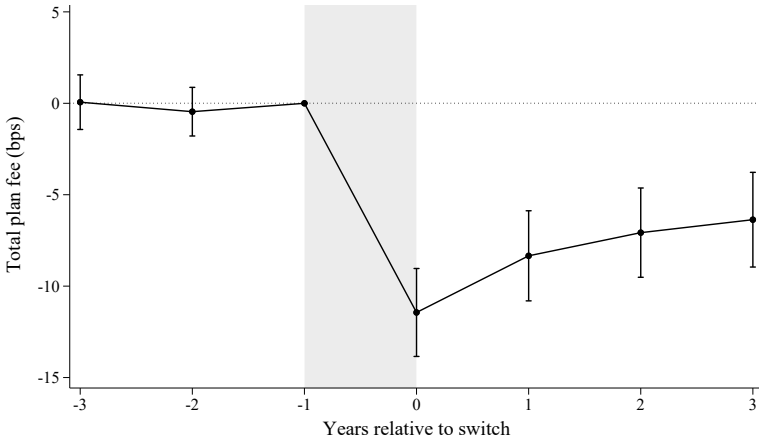


Figure displays diff-in-diffs coefficients θ_s in Equation (1). Vertical bars show 95% confidence intervals. I cluster standard errors at the employer level. Due to staggered provider switches, I follow the procedure in Sun and Abraham (2021) where I estimate θ_s separately for employers who switch each year and take weighted averages. Standard errors are bootstrapped over 200 repetitions.

statistically insignificant change after employers switch providers. Additionally in Figure A5b, the outcome variable corresponds to fees in excess of the cheapest eligible share class of the same mutual funds. Different share classes of the same mutual funds are identical products but charge different expense ratios. There is a statistically significant reduction of over 4 bps when employers switch providers. Thus, fee reductions resulting from provider switches are more likely to capture markup reductions rather than changes in plan characteristics.

4.2 How Transaction Costs Lead to Markups

Next, I discuss the mechanism of how providers can exploit transaction costs to earn markups. There has been a secular decline in mutual fund fees over the past two decades, mostly driven by the introduction of new funds that are similar to some existing funds but have lower expense ratios. 401(k) plan participants cannot access these cheaper new funds unless their employers work with providers to update plan menus. As a result, transaction costs between employers and providers can delay menu updates, allowing providers to sustain higher markups by maintaining outdated plan menus relative to the broader secular decline in investment fees.

The secular decline in fees has been documented by industry sources. For example, according to Morningstar (2021), the asset-weighted expense ratio has declined from 87 to 40 bps from 2001 to 2021. In Appendix B, I discuss factors that contribute to this trend, including the decline in production costs thanks to mutual fund managers' economies of scale.

I show that the secular decline in fees is primarily driven by the introduction of new funds with lower fees rather than reductions in expense ratios among existing funds. To illustrate this, I compare different time series of investment fees in Figure 4. Observed fees (solid black line) decreased by almost 30 bps on average from 2009 to 2019, reflecting changes in fund expense ratios, plan menus, and participant allocations. To construct this time series, I first compute asset-weighted expense ratios for each employer and then take the simple average across employers in a given year. On the other hand, the dashed black line with circles on the top of

the graph only accounts for expense ratio changes among existing funds, using fund menus and allocation as of 2009. When menu updates are not considered, the fee reduction is less than 10 bps, representing less than 30% of the total observed decline. In Appendix E.1, I discuss the construction of these time series in further detail.

Figure 4: Decline in Mutual Fund Fees

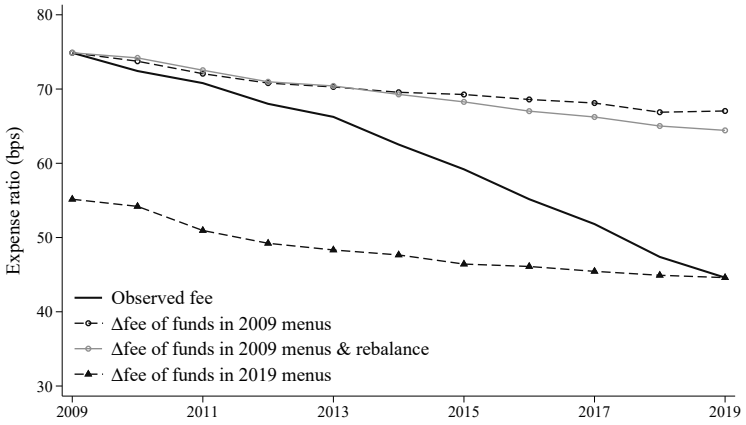


Figure shows average expense ratios of funds in 401(k) plans. I compute asset-weighted expense ratios for each employer and take simple averages across employers. The black solid line uses contemporaneous menus and allocation. The dashed line with circle markers is constructed using funds in 2009 menus, capturing only changes in expense ratios. The gray line with circle markers allows participants to rebalance among funds in their menus as of 2009. The dashed line with triangle markers is constructed using funds in 2019 menus that are available during each calendar year.

While participants can rebalance toward lower-cost funds, they may be constrained by available funds on their plan menus. The gray line in Figure 4 allows rebalancing among funds in 2009 menus, which is close to the dashed line without rebalancing. In addition, the larger fee reductions from menu updates could also be driven by changes in participant preference, such as a shift toward index funds. In Figure A3, I show similar patterns within active and index funds. Menu updates allow participants to access cheaper funds within the active or index universe.

Although menu updates play a crucial role in reducing fees, transaction costs between employers and providers can delay menu updates, allowing providers to earn markups relative to the secular trend. I find that 70% of the funds in 2019 menus were already available in the marketplace as early as 2009. If employers had incorporated those funds into their 2009 menus, they could have reduced fees by around 20 bps, as shown by the dashed line with triangle markers in Figure 4.[16][17]

The 20 bps difference does not necessarily indicate that providers earn 20 bps in markups. Forward-looking providers may temporarily set fees below costs to attract employers. If these providers can delay further menu updates by exploiting transaction costs, they can recoup markups by maintaining the same fee levels while costs decline over time. In the next section,

[16] As another piece of evidence supporting delay in menu updates, in Appendix E.2, I show that when mutual fund companies introduce cheaper share classes of the same mutual funds, providers do not immediately incorporate those into plan menus even though employers are eligible based on their investment size.

[17] The same mechanism also applies to recordkeeping fees. There is also a secular decline in recordkeeping fees thanks to improvements in providers' IT infrastructure. See Appendix B for further details. I focus on mutual fund expense ratios in this analysis, because the patterns of recordkeeping fees are less obvious due to measurement errors.

I introduce a dynamic model that allows for this invest-and-harvest fee-setting incentives.

5 Model

In this section, I introduce a structural model based on institutional details and motivating evidence discussed previously. Since providers customize services and negotiate fees with each employer, each employer represents her own market. I model how multiple providers compete to serve one single employer.[18] I assume the employer determines menu composition and types of services. Providers cater to the employer's demand and compete on fees.[19] I allow both providers and the employer to be forward-looking.

The employer (she) maximizes the utility her employee participants derive from the 401(k) plan, net of her transaction costs. The employer first decides whether to conduct an RFP, trading off the present value of potential fee reductions against RFP transaction costs κ_{rfp}. I model this decision as an optimal stopping problem (Rust, 1987), as potential fee reductions increase over time due to the decline in providers' production costs.

Second, the employer chooses a provider at an RFP. When making provider choices, the employer considers fees proposed by providers and provider quality. The employer also incurs a switching cost κ_{sw} when choosing a non-incumbent provider. I allow for heterogeneous fee sensitivities among employers, capturing how they trade off transaction costs against plan fees. I model the employer's provider choice at an RFP using a random-coefficient discrete choice model (Berry et al., 1995).[20]

The Provider (he) competes with other providers to offer services to the employer. Since providers cater to the employer's preference for menus and services, they compete on fees. While fund expense ratios are set by fund managers, providers can effectively determine plan fees by selecting funds for plan menus and setting recordkeeping fees. In my model, providers compete in a dynamic fee-setting game at each RFP. Due to transaction costs, dynamic incentives can push providers to set fees below static optimal levels, anticipating higher future markups.

Providers may also update menus outside of RFPs. For simplicity, I assume providers make binary decisions of whether to update menus, with a constant and exogenous probability.

State variables: The model is dynamic in discrete time with two state variables: i) the identity of the incumbent provider s, ii) the current level of fee f. I call the state variable f fee for exposition purposes. In this model, f captures markups and does not contain production costs.

Stationarity: My model is stationary for analytical tractability. Although plan fees decline over time, I attribute this decline to reductions in production costs, while the markups providers charge remain stationary across calendar years. As shown in Figure 1b, the market shares of large providers remain stable over time, supporting the stationary assumption.

Figure 5 illustrates that observed plan fees remain roughly constant and decrease following

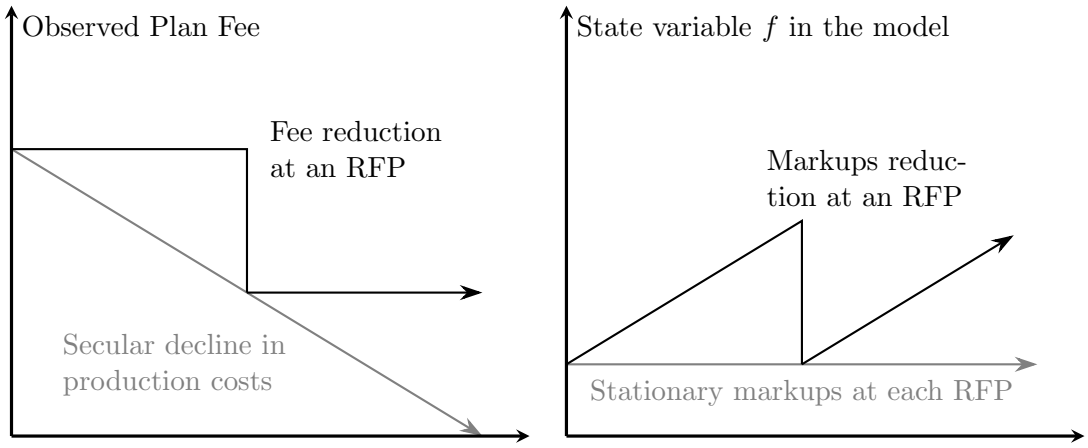
[18] By focusing on a single employer, I abstract away from cross-employer considerations, such as providers using certain employers to signal quality or manage capacity constraints.

[19] I abstract away from how providers steer menu composition, because most providers have access to a wide range of funds and can be relatively indifferent to employers' preferences for funds.

[20] Bhattacharya and Illanes (2022) develop a bargaining model between employers and providers. An employer with a larger bargaining power incurs lower markups, and the bargaining power is analogous to the fee sensitivity parameter in my model.

menu updates in an RFP, whereas the fee state variable increases over time without menu updates. From observed fees to the fee state variable, I effectively rotate the patterns in Figure 4 by normalizing for the declining secular trend. In Section 6.1, I discuss how I detrend observed fees with a hedonic regression before structural estimation using my stationary model.

Figure 5: Detrending Observed Fees to the Fee State Variable



Timing: Within each period (year), the following happens sequentially.

0. The employer and all J providers observe the two state variables: the incumbent provider s and fee f.
1. The employer decides whether to incur RFP cost κ_{rfp} to run an RFP.

2.a. At the RFP, each provider sets his optimal fee $b_j(s)$ in a dynamic fee-setting game.

2.b. The employer realizes her idiosyncratic preference for each provider and chooses the provider j who generates the highest utility. States transition from (s, f) to $(j, b_j(s))$

3. The current provider (the winner of the RFP j or the incumbent from the beginning of the period s) receives fee f from participants of the employer. Fee f transitions to the next period depending on whether the provider updates the menu.

I introduce the employer's provider choice at stage 2 in Section 5.1 and her decision to run an RFP in Section 5.2. Then, I specify providers' fee competition in Section 5.3. Sections 5.4 and 5.5 explain fee transitions and the equilibrium concept.

5.1 Stage 2.b: Employer's Provider Choice at an RFP

I model the employer's provider choice at an RFP as a discrete choice (McFadden, 1973; Berry et al., 1995). The employer's choice depends on the identity of the incumbent provider s due to switching cost κ_{sw} but not on fee f before the RFP.

The employer i derives utility from each provider j according to Equation (2). Each provider proposes a fee $b_{ij}(s)$, which I discuss later in Section 5.3. I allow the fee sensitivity parameter α_i to vary across employers i following a lognormal distribution $\ln \alpha_i \sim \mathcal{N}(\alpha, \sigma)$. Provider quality is captured by fixed effect ν_j. The employer incurs a switching cost κ_{sw} when choosing a non-incumbent provider $j \neq s$. $\mathbb{E}V^E(j, b_{ij}(s))$ represents the employer's value function from

choosing j, where the expectation is taken over fee transitions to the next period. δ is the discount factor, which I assume to be 0.95 for the employer and providers. The employer draws an idiosyncratic preference shock ϵ_{ij} for each provider, which follows the Type I extreme value distribution with the scale normalized to 1.

$$\mathcal{U}_{ij}(s) + \epsilon_{ij} = -\alpha_i b_{ij}(s) + \nu_j - \kappa_{sw} 1\{j \neq s\} + \delta \mathbb{E} V^E(j, b_{ij}(s)) + \epsilon_{ij} \tag{2}$$

The probability that the employer chooses provider j is given by the following. [21]

$$q_{ij}(s) = \frac{e^{\mathcal{U}_{ij}(s)}}{\sum_{k=1}^{J} e^{\mathcal{U}_{ik}(s)}} \tag{3}$$

The employer's expected utility from the RFP follows the standard inclusive value expression.

$$\mathbb{E}\mathcal{U}_i(s) = \ln \sum_{k=1}^{J} e^{\mathcal{U}_{ik}(s)} \tag{4}$$

5.2 Stage 1: Employer's Decision to Run an RFP

In an optimal stopping problem (Rust, 1987), the employer runs an RFP when the expected benefit exceeds the RFP cost κ_{rfp}. Since future fee transitions depend on the provider s, I condition on s in addition to fee f. If the employer runs an RFP, she receives utility following Equation (5), which accounts for the expected utility from an RFP $\mathbb{E}\mathcal{U}_i(s)$ net of the RFP cost κ_{rfp}. If she does not run an RFP, she receives utility following Equation (6), and she remains with her incumbent provider at the current fee f.

The employer's trade-off between the benefits to participants and her own RFP cost κ_{rfp} may differ from her provider choice at RFPs. The employer may have imprecise information about fees before receiving fee proposals from providers or follow a predetermined schedule for conducting RFPs every few years. For simplicity, I introduce an inattention parameter ρ when the employer decides whether to run an RFP. Finally, $\epsilon_{i,1}, \epsilon_{i,0}$ are Type I extreme value shocks, again with the scale normalized to 1. [22]

$$\mathcal{U}_{i1}(s) + \epsilon_{i1} = \rho \mathbb{E}\mathcal{U}_i(s) - \kappa_{rfp} + \epsilon_{i1} \tag{5}$$

$$\mathcal{U}_{i0}(s, f) + \epsilon_{i0} = \rho\Big(\nu_s - \alpha_i f + \delta \mathbb{E} V^E(s, f)\Big) + \epsilon_{i0} \tag{6}$$

The probability of running an RFP is the following:

$$\lambda_i(s, f) = \frac{\exp\big(\mathcal{U}_{i1}(s)\big)}{\exp\big(\mathcal{U}_{i1}(s)\big) + \exp\big(\mathcal{U}_{i0}(s, f)\big)} \tag{7}$$

$V_i^E(s, f)$ below is the employer's start of period value function, which again follows the standard

[21] Since there is no outside option of not offering 401(k) plans, the employer-specific shifter of quality or costs cancels out when comparing the differences in $\mathcal{U}_{ij}(s)$ across providers. Employers with arbitrarily different quality or costs but the same α_i have the same provider choice probabilities.

[22] I have also considered a scale parameter for the Type I extreme value shocks instead of ρ and obtained qualitatively similar estimates. I prefer the inattention parameter ρ because it does not attribute a higher degree of randomness in RFP decisions to expected utility.

inclusive value expression.

$$V_i^E(s,f) = \ln\Big(\exp\big(\mathcal{U}_{i1}(s)\big) + \exp\big(\mathcal{U}_{i0}(s,f)\big)\Big) \tag{8}$$

5.3 Stage 2.a: Providers' Dynamic Fee Competition

The provider's ex-ante value function at an RFP depends on the probability of winning the RFP and the difference in continuation values between winning and losing. Provider j's production cost is captured by cost fixed effect c_j. $V_i^W(j, b_{ij}(s))$ denotes the winner's continuation value, and $V_{ij}^L(k, b_{ik}(s))$ denotes j's continuation value when k wins instead. Expectations $\mathbb{E}V_i^W$ and $\mathbb{E}V_{ij}^L$ are taken over fee transitions.

$$V_{ij}^{P,rfp}(s) = q_{ij}(s)\Big(b_{ij}(s) - c_j + \underbrace{\delta\mathbb{E}V_i^W(j, b_{ij}(s))}_{\substack{\text{Winner}\\ \text{continuation value}}}\Big) + \sum_{k\neq j} q_{ik}(s)\underbrace{\delta\mathbb{E}V_{ij}^L(k, b_{ik}(s))}_{\substack{\text{Loser}\\ \text{continuation value}}} \tag{9}$$

Provider continuation values as the winner or loser of the RFP depend on the probability $\lambda_i(s,f)$ that the employer runs RFPs.

$$\begin{aligned} V_i^W(s,f) &= \lambda_i(s,f)V_{is}^{P,rfp}(s) + \big(1-\lambda_i(s,f)\big)\big(f - c_s + \delta\mathbb{E}V_i^W(s,f)\big) \\ V_{ij}^L(s,f) &= \lambda_i(s,f)V_{ij}^{P,rfp}(s) + (1-\lambda_i(s,f))\delta\mathbb{E}V_{ij}^L(s,f) \end{aligned}$$

If the employer runs an RFP, the previous winner participates as the incumbent $V_{is}^{P,rfp}(s)$ with a higher probability of winning again due to the switching cost, whereas previous losers participate as non-incumbent providers $V_{ij}^{P,rfp}(s)$ with $j \neq s$. If the employer does not run an RFP, the winner earns fee f net of production costs c_s and his continuation value, whereas losers only internalize their continuation values.

To gain further intuition, I can solve these continuation values recursively and rewrite Equation (9) as the following

$$V_{ij}^{P,rfp}(s) = q_{ij}(s)\Big(\overbrace{\gamma_{ij}(s)}^{\substack{\text{expected future}\\ \text{service duration}}}(b_{ij}(s) - c_j) + \overbrace{\eta_{ij}^W(s)}^{\substack{\text{Future gains}\\ \text{as the winner}}}\Big) + (1-q_{ij}(s))\overbrace{\eta_{ij}^L(s)}^{\substack{\text{Future gains}\\ \text{as the loser}}}$$

γ_{ij} corresponds the expected periods that j provides services to employer i. By setting fees at $b_{ij}(s)$, provider j expects to earn $\gamma_{ij}(s)(b_{ij}(s) - c_j)$ without updating menus. In addition, provider j also expects future gains $\eta_{ij}^W(s)$, including the production costs decline that will raise j's markups in the future and the present value of participating in the next RFP as the incumbent. On the other hand, if provider j loses the RFP, he obtains $\eta_{ij}^L(s)$ which only includes the present value of participating in the next RFP as a non-incumbent provider. Note that $\gamma_{ij}(s), \eta_{ij}^W(s), \eta_{ij}^L(s)$ depend on the probabilities of RFPs, which in turns are affected by fees set at the current RFP.

Each provider sets his optimal fee $b_{ij}(s)$ by solving the FOC of $V_{ij}^{P,rfp}(s)$ with respect to $b_{ij}(s)$, as best responses to fees set by other providers $b_{i,-j}(s)$ and before the employer realizes

her idiosyncratic preferences ϵ_{ij} for each provider.

$$b_{ij}(s) = \underbrace{c_j + \frac{1}{\alpha_i(1-q_{ij}(s))}}_{\text{static cost + markups}} - \underbrace{\frac{\eta_{ij}^W(s) - \eta_{ij}^L(s)}{\gamma_{ij}(s)}}_{\text{future gains per period}>0} + \underbrace{\frac{\frac{\partial \gamma_{ij}(s)}{\partial b_{ij}}(b_{ij}(s) - c_j) + \frac{\partial \eta_{ij}^W(s)}{\partial b_{ij}}}{\alpha_i(1-q_{ij}(s))\gamma_{ij}(s)}}_{\text{higher fees raise prob of RFPs }<0} \tag{10}$$

The first term is the standard cost plus markups in static optimal fees. Because $q_{is}(s)$ is higher for the incumbent provider s due to switching cost, the incumbent s charges higher markups. Employers with low fee sensitivities α_i have higher markups and larger markup difference between incumbent and non-incumbent providers.

The second and third terms capture dynamic incentives. The second term is typical in dynamic models with switching costs. It corresponds to higher future gains per period after winning the current RFP. This future gain is competed away at the RFP, prompting providers to reduce their fees $b_{ij}(s)$. The third term captures an additional incentive due to endogenous RFP decisions. Since the employer is more likely to run another RFP if her current fee is high, setting a lower fee will reduce the probabilities of RFPs and raise expected profits. I discuss how I solve the FOC in Appendix F.1.

5.4 Stage 3: Fee Transitions and Menu Updates

I allow providers to update plan menus outside of RFPs. For tractability, I focus on the extensive margin of whether providers update plan menus and abstract away from modeling the magnitude of menu turnover. Without menu updates, providers earn higher markups because their production costs decline over time while plan fees remain unchanged. Since the fee state variable captures markups, it increases between periods when there are no menu updates. I allow the rate of increase g_s to be provider-specific and consider g_s as an exogenous provider attribute.

For simplicity, I assume all providers follow a constant and exogenous menu update probability ϕ. Although I do not allow providers to update menus endogenously, providers internalize this restriction when they set fees at RFPs. Whenever providers update menus, I assume that fee growth is set to $g_s = 0$ for all providers.[23] While my approach is potentially equivalent to a model without menu updates, where fees grow at some average rate between g_s and zero, allowing for menu updates enables me to use menu turnover data to recover the unobserved probabilities of RFPs, which I discuss later in Appendix G.3.

The expected provider continuation values integrate over the probability of menu update ϕ. Again, fee f increases at provider-specific rates g_s without menu updates and stays constant otherwise.

$$\begin{aligned}
\mathbb{E}V_i^W(s,f) &= \phi V_i^W(s,f) + (1-\phi)V_i^W(s,f+g_s) \\
\mathbb{E}V_{ij}^L(s,f) &= \phi V_{ij}^L(s,f) + (1-\phi)V_{ij}^L(s,f+g_s)
\end{aligned}$$

[23] I have estimated a parameter for fee growth when providers update menus and obtained an estimate very close to zero. I have also estimated a version of the model with endogenous menu update and obtained similar results.

The employer's expected value function also integrates over menu update probability ϕ.

$$\mathbb{E}V_i^E(s,f) = \phi V_i^E(s,f) + (1-\phi)V_i^E(s,f+g_s)$$

5.5 Equilibrium

I solve for the Markov Perfect equilibrium, where

1. The employer decides whether to run RFPs given her current incumbent provider and fee (s,f), according to the optimal stopping problem in Section 5.2
2. At RFPs, the employer chooses a provider to maximize expected utility, according to Section 5.1
3. Providers set optimal fees at RFPs as best responses to fees set by other providers following the FOC in Equation (10) of Section 5.3
4. The employer and providers have the same beliefs regarding transitions of fees given each other's optimal policies

The equilibrium corresponds to a fixed point of the employer's provider choices, the employer's RFP decisions, and providers' optimal fees conditional on state variables (s,f). In my model, fees set by providers at RFPs enter the employer's and providers' value functions. Additionally, the employer's probabilities of RFPs affect provider value functions. As a result, I solve the employer's RFP decisions and providers' optimal fees jointly.

6 Estimation and Identification

My estimation follows two steps. First, I use a hedonic regression to control for the variation in fees driven by observed characteristics that are assumed to capture production costs, including the time trend. Second, I estimate structural parameters that affect markups using indirect inference (Gourieroux and Monfort, 1996) to match residualized fees from the first step, along with provider choices and menu turnover. I estimate several parameters offline when they directly correspond to certain empirical moments without having to solve the structural models. In Table 2, I summarize which parameters I estimate in each step and the moments in the data I use for identification.

To allow for heterogeneity across employers of different sizes, I break employers into quartiles based on their total plan assets.[24] I control for size quartiles in the hedonic regression and estimate structural parameters separately for each quartile, allowing transaction costs, fee sensitivities, production costs, and preferences for providers to vary across sizes of employers.

6.1 Step 1: Residualizing Fees with Hedonic Regression

I residualize the observed characteristics that I assume correspond to exogenous variation in production costs and do not affect markups. Specifically, I use the following hedonic regression. The dependent variable fee_{ijt} represents the observed plan fee of

[24] I compute average total plan assets over time for each employer, so employers do not move across size quartiles.

Table 2: Overview of Parameter Estimation

Parameters	Moments
Structural Estimation	
Transaction costs: $\kappa_{sw}, \kappa_{rfp}$	Probability of switching and fee autocorrelation
Employer fee sensitivity and inattention: $\alpha, \sigma_\alpha, \rho$	Fee reductions from switching, sensitivity of switching wrt. fees
Provider propensity to update menu: ϕ	Probability of menu updates
Small providers net quality: $\nu_{other} - \alpha c_{other}$	Probability of switching from small providers
Hedonic Regression & Offline	
Large provider net quality: $\nu_j - \alpha c_j$ Employer specific net quality: $\nu_i - \alpha c_i$	Employer FE & choice probability conditional on switching
Offline	
Markup growth rates: g_j	Increase in residualized fees without menu updates

employer i at time t whose provider is j. Plan fees include asset-weighted expense ratios and direct recordkeeping fees as a fraction of total plan assets.

$$fee_{ijt} = \underbrace{X_{it}\beta}_{\substack{\text{Observed} \\ \text{characteristics}}} + \underbrace{\Gamma_j + \tau_{jt} + \tau_{s(i)t}}_{\substack{\text{Provider FE} \\ \text{and time trend}}} + \underbrace{\Gamma_i + \epsilon_{ijt}}_{\tilde{f}_{ijt}\text{: Residualized fee}} \tag{11}$$

X_{it} captures observed employer and plan characteristics, including menu composition and types of services that provider j offers. I also use time trends $\tau_{jt}, \tau_{s(i)t}$ to residualize provider-specific and employer-size-quartile-specific decline in production costs. Lastly, I use provider fixed effects Γ_j to control for persistent differences in the quality or production costs across providers. Providers can also charge different markups, and I will later include another set of provider fixed effects in the structural estimation step. In Appendix H, I discuss the hedonic regression in greater detail.

Residualized fees $\tilde{f}_{ijt}$ capture instances such as a more expensive fund within an investment category, which are more likely to indicate higher markups. Alternatively, residualized fees can include unobserved services that employers ask their providers to offer, which can correspond to unobserved production costs. Both markups and unobserved production costs can be persistent for a given employer (Γ_i) or time-varying (ϵ_{ijt}), which I disentangle using my structural model.

6.2 Step 2: Structural Estimation with Indirect Inference

The indirect inference estimation procedure relies on simple auxiliary models to capture the variation in the data that allows me to identify structural parameters.[25] I estimate the coefficients of these auxiliary models using observed data $\mathcal{B}^{data}$ and also

[25] Because I do not observe fees proposed by non-chosen providers at RFPs, I cannot use standard demand inversion as in Berry et al. (1995). I also cannot use analytical solutions of a second-price auction (Allen et al., 2019; Cuesta and Sepúlveda, 2021) due to the dynamic feature of the model where fees enter the continuation values. As a result, I adopt this full-solution approach, where I simulate equilibrium outcomes from my structural model and compare with their empirical counterparts.

estimate these coefficients using data simulated by my model $\mathcal{B}^{model}(\Theta)$ at candidate sets of parameters Θ. The estimation procedure searches for structural parameters Θ to minimize the distance between the two sets of auxiliary model coefficients $\hat{\Theta} = \arg\min_{\Theta} ||\mathcal{B}^{data} - \mathcal{B}^{model}(\Theta)||_2$.

I use indirect inference because I do not directly observe the fee state variable f. The closest empirical counterpart is residualized fees $\tilde{f}_{ijt}$. However, residualized fees $\tilde{f}_{ijt}$ still include unobserved production costs, causing an omitted variable bias. For example, an employer may not switch her provider despite high fees because these fees compensate for higher production costs of more valuable services. As a result, I would underestimate the sensitivity of her switching decision with respect to fees.

Estimating a large number of employer fixed effects or allowing for a mixture of latent types is difficult in non-linear structural models with short panels. However, in linear auxiliary models, unobserved production costs can be absorbed by employer fixed effects. These employer fixed effects correspond to employer-specific quality or cost shifters that cancel out when comparing utilities or fees across providers. Therefore, I can estimate comparable auxiliary model coefficients using observed data and data simulated by my model.

I introduce seven auxiliary models to identify seven structural parameters: RFP costs κ_{rfp}, switching costs κ_{sw}, average and standard deviation of fee sensitivity α, σ, inattention in RFP decision ρ, providers' menu update probability ϕ, and lastly the net quality for providers in the other category $\xi_{other} = \nu_{other} - \alpha c_{other}$. I do not separate quality ν from cost c, since only their difference matters after plugging in the expression of optimal fees Equation (10) into employers' utility in Equation (2). Details of the auxiliary models are included in Appendix G.

6.3 Step 2: Identification of Structural Parameters

I focus on the identification of transaction costs and employer fee sensitivity. In Appendix G, I discuss how each auxiliary model identifies the corresponding parameters.

Transaction costs: The probability that employers switch providers identifies the sum of the two transaction costs $\kappa_{rfp} + \kappa_{sw}$. Since switching requires both running an RFP and choosing a non-incumbent provider, higher switching costs and/or RFP costs lead to a lower probability of switching.

Although RFPs are not directly observed in the data, I can infer the probability of RFPs using the autocorrelation of fees. This fee autocorrelation then allows me to separate RFP costs κ_{rfp} from switching costs κ_{sw}. The lower bound of the fee autocorrelation is zero, assuming that providers do not condition on previous fees when setting fees at RFPs. The upper bound can be directly estimated when there is no

menu turnover.[26] For employers who do not switch providers, an autocorrelation close to zero suggests frequent RFPs and low RFP costs, whereas an autocorrelation near the upper bound suggests infrequent RFPs and high RFP costs. See Appendix G.3 for more details.

Employer fee sensitivity: The sensitivity of switching providers with respect to fees identifies the mixture of average fee sensitivity α and the inattention parameter ρ. If higher fees strongly predict switching, employers should be attentive in their RFP decisions and/or put more weight on fees rather than transaction costs.

To separately identify α, I use fee reductions when employers switch providers. In my model, where providers only compete on fees, employers with higher fee sensitivities have smaller fee reductions from switching. Suppose an employer is extremely fee-sensitive, such that providers engage in perfect competition. In this case, her fee reduction from switching providers only captures accumulated growth in markups from the previous RFP. Any additional reductions reflect markup differences between incumbent and non-incumbent providers, which depend on the employer's fee sensitivity. When considering switching providers at an RFP, the employer trades off lower markups against switching costs. If the employer puts less weight on fees, non-incumbent providers have to set lower markups to incentivize switching. As a result, the employer achieves larger fee reductions when choosing non-incumbent providers.[27]

As robustness checks, I examine how fee reductions from switching vary across employer characteristics. Appendix G.6 shows that fee reductions from switching are smaller for larger employers, employers that contribute more to their plans, employers with higher participation rates, and employers with higher ESG scores. These employers are likely more sophisticated and value employee welfare more. It seems reasonable that they place greater emphasis on plan fees paid by their participants rather than their own transaction costs.

Random coefficient in employer fee sensitivity: Since fee reductions from switching identify the average employer fee sensitivity, the variance of fee reductions from switching should capture the variance of employer fee sensitivity σ^2. However, using the variance of fee reductions directly can be problematic, because it may include the variance of changes in measurement errors and unobserved production costs.

To make progress, I assume that, after controlling for provider fixed effects, fee changes from switching in the data consist of two components: strictly negative markup changes (positive markup reductions) related to employer fee sensitivity and zero-mean noise. Any increase in fees from switching providers has to come from noise rather than markups. Thus, the truncated mean of positive fee changes from switching can be used

[26] The upper bound is not necessarily equal to one in the data due to changes in fund expense ratios, reallocation across different funds, and measurement errors in recordkeeping fees.

[27] Figure A6 shows that fee reductions from switching under estimated fee sensitivities are much larger compared to fee reductions assuming perfect competition.

to estimate the variance of noises, allowing me to recover the variance of employer fee sensitivity σ^2.

Limitations: Using the average and variance of fee reductions from switching providers to recover the distribution of fee sensitivity α_i imposes certain assumptions. Employers with the lowest fee sensitivities rarely conduct RFPs or switch providers. Provider switches in my data primarily reflect behaviors of relatively fee-sensitive employers, and I rely on the lognormal functional form to extrapolate.

In addition, I find variation in the data that corresponds to employer fee sensitivities and transaction costs, which determine markups in my model. However, there may be other sources of markups that I do not capture in my structural estimation. Consequently, I may underestimate both the level and dispersion of markups. As an alternative strategy, I directly use the variance of fees to identify the variance of employer fee sensitivity σ^2, assuming no variation in unobserved production costs within each size quartile. Results from this approach likely capture an upper bound of the level and dispersion of markups.

6.4 Parameters Estimated Offline

Other parameters in the model are directly observed or can be estimated offline. First, I assume there are six providers at each RFP, including the five largest providers (Fidelity, Vanguard, Empower, ADP, and Principal Financial Group) and one representative from the other category of smaller providers.[28] Choice probabilities conditional on switching can recover quality net of cost ξ_j for the five large providers, which I explain in Appendix F.2. Second, I measure markup growth rates g_j directly in the data, using median changes in residualized fees $\tilde{f}_{ijt}$ without menu updates to capture how much expense ratios increase relative to the secular trend. In Table A10, I report conditional choice probabilities and markup growth rates.

7 Estimation Results

I present parameter estimates for each employer size quartile in Table 3. To compare magnitudes, I divide transaction cost parameters by average fee sensitivities $\bar{\alpha}$ so they have the same unit as fees. Both RFP and switching costs in terms of basis points over plan assets decrease with employer size. A large component of these costs consists of fixed costs, which naturally become smaller when divided by larger plan assets. From small employers in the first quartile to large employers in the fourth quartile, RFP costs decrease from 10 to 1 bps, and switching costs decrease from 46 to 2 bps. Measured in dollars, both transaction costs increase with employer size except between the third

[28] If the incumbent prior to the RFP is a provider from the other category, I assume there are seven providers to allow for switching between providers in the other category.

and fourth quartile. There can also be some variable components, and these transaction costs become higher for employers with more participants and potentially more complex plan structures. It is reasonable that switching costs are much larger than RFP costs, since changing providers involves handling disruptions to participants' experiences and potentially transitioning of bundled non-401(k) services.

Employer fee sensitivity α_i follows a lognormal distribution, and I report the average $e^{\alpha+\frac{1}{2}\sigma^2}$ and standard deviation $((e^{\sigma^2}-1)e^{2\mu+\sigma^2})^{1/2}$. I estimate that larger employers are more fee sensitive, especially those in the fourth size quartile. In the first quartile, employers' fee elasticity when choosing providers at RFPs is only 1.6, suggesting a relatively low level of fee sensitivity and potentially high markups. On the other hand, employers in the fourth quartile have an elasticity of 50, suggesting that providers engage in near-perfect competition for large employers. The standard deviation of α_i increases along with the mean, while the coefficient of variation (mean divided by standard deviation) is around one for all four quartiles, suggesting similar dispersion in fee sensitivities across employer size.

When deciding whether to conduct RFPs, employers tend to be less attentive with $\rho < 1$. The attention parameter is lower for larger employers, but the multiple of $\rho \times \alpha$ is still higher for larger employers. Outside of RFPs, the probabilities that providers update menus range from 50% for small employers to almost 70% for large employers. Larger employers are in a stronger position to demand more frequent menu updates.

The net equality ξ should be interpreted on a relative basis across providers.[29] Among large providers, Fidelity has the highest net quality, consistent with its large market share shown in Figure 1a. Vanguard has relatively higher ξ among large employers and relatively lower ξ among small employers. Figure 1a shows that Vanguard has a larger market share in plan assets than in number of plans, consistent with higher net quality among large employers. ADP has a higher ξ on a relative basis for small employers in the first quartile. ADP bundles recordkeeping with HR payroll, which is more appealing for small employers.

7.1 Goodness of Fit

In Table A11, I assess the goodness of fit by comparing auxiliary model coefficients estimated using data simulated by the model and coefficients estimated using observed data. The standard errors of coefficients using observed data are shown in parentheses and indicate whether the differences in coefficients are statistically significant. Coefficients are generally not distinguishable at the 5% significance level (two-tailed). The only

[29] Technically, there are two ξ_{other} for other providers in my model. In Table 3, ξ_{other} corresponds to the net quality of one incumbent provider from the other category. At an RFP, the employer may invite multiple small providers from the other category to participate at an RFP. I only model one representative provider of the other category, and the $\tilde{\xi}_{other}$ of this representative also includes a variety effect. I normalize $\tilde{\xi}_{other} = 0$. Without the variety effect, ξ_{other} is negative.

Table 3: Structural Parameter Estimates

Employer quartiles		1st		2nd		3rd		4th	
		Coef	SE	Coef	SE	Coef	SE	Coef	SE
RFP cost	$\kappa_{rfp}/\bar{\alpha}$								
bps		10.2	(0.2)	6.4	(0.4)	4.8	(0.1)	0.8	(0.5)
$ thousand		7		10		17		18	
Switching cost	$\kappa_{sw}/\bar{\alpha}$								
bps		46.0	(0.8)	24.2	(1.3)	15.9	(0.4)	1.7	(1.0)
$ thousand		30		38		57		39	
Fee sensitivity: mean	$\bar{\alpha}$								
Coefficient		0.11	(0.00)	0.24	(0.01)	0.41	(0.01)	4.66	(2.46)
Elasticity		1.6		3.4		5.2		64.4	
Fee sensitivity: st.dev.	σ_α								
Coefficient		0.14	(0.00)	0.32	(0.02)	0.53	(0.02)	15.00	(10.15)
Elasticity		1.6		3.5		5.1		58.9	
RFP attention	ρ								
Coefficient		0.78	(0.01)	0.59	(0.03)	0.47	(0.01)	0.26	(0.15)
Pr menu update	ϕ	0.50	(0.00)	0.62	(0.00)	0.68	(0.00)	0.63	(0.00)
Net quality of other providers	$\xi/\bar{\alpha}$	-14.6	(0.3)	-13.4	(0.7)	-10.5	(0.2)	-1.1	(0.6)
Relative to other providers									
ADP		-0.8		2.0		1.0		0.2	
Principal Fin.		-2.4		0.9		1.7		0.2	
Empower		1.8		3.7		3.3		0.4	
Fidelity		1.6		3.4		3.9		0.4	
Vanguard		-1.7		2.0		2.9		0.4	

Table reports parameter estimates. Standard errors are computed using the asymptotic variance formula, where the variance-covariance matrix of auxiliary model coefficients is computed using the bootstrap method over 250 iterations. In each iteration, I sample employers with replacement and estimate auxiliary model coefficients.

coefficient that I fit slightly poorly is the probability of switching from large providers.

In Table A10, I show that the probabilities that employers run RFPs range from 15% to 30% based on model simulation. This is broadly consistent with the industry common practice of conducting an RFP every three to five years. When some employers claim they run an RFP every three years, they could refer to formal RFPs, informal Requests for Information (RFIs), or benchmarking. RFP decisions in my model should correspond to formal RFPs, and it seems reasonable that the RFP frequency from my model is close to the lower bound of the industry practice. Large firms run RFPs less frequently potentially because they adopt more formal RFP procedures that require coordination across multiple departments.

7.2 Markups

Having estimated my model, I can recover markups. At RFPs, markups can be computed from providers' FOC conditions Equation (10). After RFPs, markups evolve according to growth rates g_j until the next RFPs. I report markups separately for each

size quartile in Figure 6 and show aggregated results in Table 4.

Average level of markups: In the first column of Table 4, I show that markups are on average 18% of fees, among all employers in my sample. Around 12 bps out of the 65 bps average plan fees as of 2019 are markups, while the rest are production costs.

In Figure 6a, I separate plan fees into production costs and markups for each size quartile. Markups are 18 bps or 23% for small employers, decreasing to 6 bps or 13% for large employers. Lower markups for larger employers are consistent with larger employers having higher fee sensitivity. Between the first and fourth size quartiles, the difference in plan fees is 20 bps. Roughly one-third of this difference is due to markups, and two-thirds comes from production costs, which account for economies of scale of large employers.

Next, I decompose markups in Figure 6b to help illustrate the mechanism in my model. First, non-incumbent providers set lower markups at RFPs to incentivize switching. Providers charge 4 bps in markups to small employers in the first quartile and negative markups to large employers. Forward-looking providers are willing to charge negative markups temporarily, anticipating higher future markups taking advantage of employers' transaction costs. Second, incumbent providers set relatively higher markups at RFPs to exploit switching costs. Dark gray bars show that markups from switching costs range from over 10 bps for small employers to 5 bps for large employers. Lastly, because employers do not run RFPs every year, markups increase over time until the next RFPs. Light gray bars show that the last source of markups between RFPs is between 2 and 4 bps across employers with different sizes.

Distribution of markups: In Figure 6c, I plot the distribution of markups simulated by my model across different states. Within each size quartile, markups vary across employers due to differences in fee sensitivity and within employers across state (s, f) because employers incur transaction costs at different points in time. The solid black line indicates that markups center around zero for large employers. Since the market for large employers is nearly perfectly competitive, providers shuffle markups across time but earn a small amount of markups on average. As a comparison, markups are higher on average for smaller employers and also more dispersed. Due to the wide dispersion, there are substantial overlap between the four distributions, suggesting that some large employers can have higher markups than small employers at least temporarily.

Decomposition of fee dispersion: I decompose the dispersion in 401(k) plan fees into different sources of markups and production costs. Using the law of total variance, I separate the variance of markups into the between-employer variation due to different fee sensitivity and the within-employer variation because employers incur transaction

Figure 6: Model Generated Markups

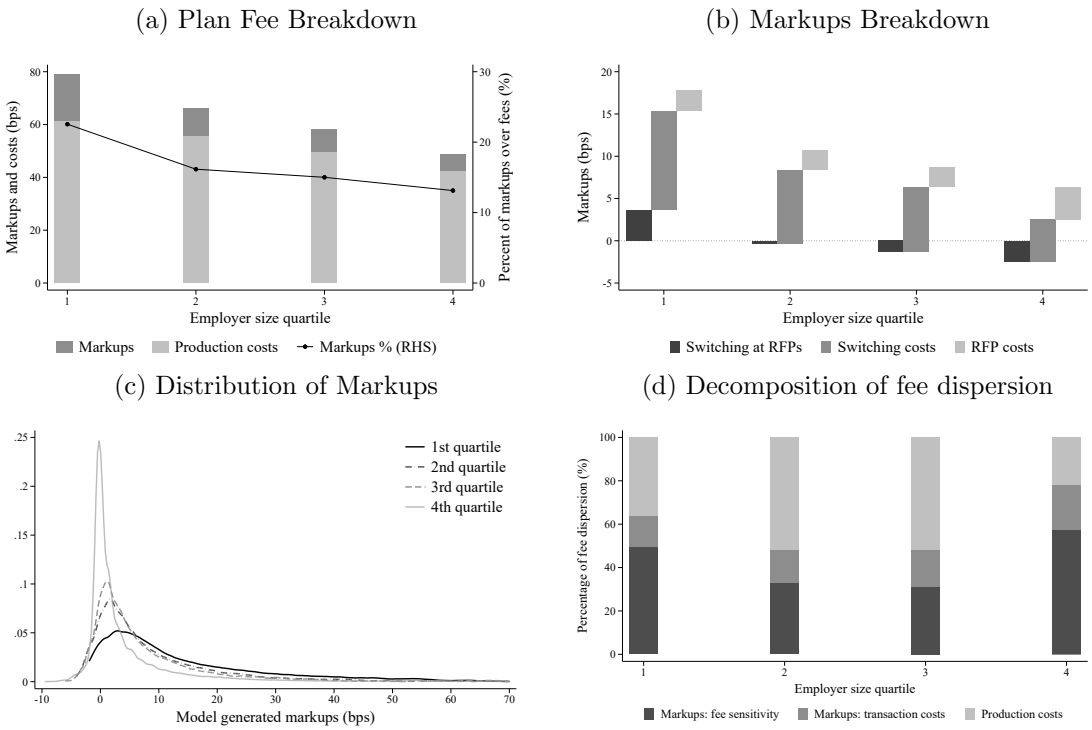


Figure illustrates model generated markups for employers in each size quartile. Panel (a) separates plan fees into markups and production costs. Panel (b) decompose markups into components set by providers at RFPs and growth in markups between RFPs. Panel (c) plots distribution of markups across states. Panel (d) shows the decomposition of fee dispersion.

costs at different points in time.

$$Var(f_i) = \underbrace{Var(\bar{f}_i)}_{\text{between: fee sensitivity}} + \underbrace{E_i[Var(f_i|i)]}_{\text{within: transaction costs}}$$

Further separating markups into these two components is useful. The variance driven by fee sensitivity is more likely to capture how well employers manage their plans. On the other hand, some variance of markups is natural in this type of negotiated-price market with transaction costs. Even homogeneous employers may have different fees in the cross-section, and this component of variance in markups is not informative of employers' behaviors. [30]

I attribute the variation in plan fees not explained by markups to production costs. The variance of production costs includes the component explained by observed plan characteristics in the hedonic regression and the remaining fee variation not explained

[30] I am able to make this decomposition because I allow employers in a size quartile to have different fee sensitivity but face the same transaction costs. In practice, both can vary across employers. In my model, I capture how employers trade off transaction costs against fees based on their fee sensitivity. My transaction costs parameters can be viewed as some common costs of doing business across employers. On top of this common friction, an employer who internalizes higher transaction costs behaves similarly as an employer who is less fee-sensitive.

by either the hedonic regression or the structural model.

In the first column of Table 4, I report the variance decomposition by pooling employers across size quartiles. The variation in markups driven by fee sensitivity explains 38% of the overall fee dispersion. Within-employer variation in markups due to transaction costs explains 13%. The remaining 48% is attributed to production costs.

Across different size quartiles, fee sensitivity explains a greater fraction of the fee dispersion for smallest employers in the first quartile and largest employers in the fourth quartile. There may be greater heterogeneity in sophistication across employers within these two quartiles at the boundaries of the overall size distribution. In Figure A2, I show a wider dispersion of employer size within the first and fourth quartile.

This decomposition can shed light on recent excessive fee lawsuits. While differences in fee sensitivity play an important role, a substantial fraction of the fee dispersion is attributed to transaction costs or production costs. As a result, cross-sectional comparisons of plan fees may not be informative in determining whether employers breach their fiduciary duties.

Table 4: Decomposition of Plan Fees and Fee Dispersion

	Baseline	No unobserved production costs	Myopic
% level of fee			
Markups	18	24	21
Production costs	82	76	79
% variance of fee			
Markups: fee sensitivity	38	63	28
Markups: transaction costs	13	13	14
Production costs	48	23	58

Table reports the decomposition of the level and the dispersion of plan fees into markups and production costs. The first column corresponds to estimates from the baseline specification. The second column corresponds to the alternative assumption without unobserved production costs. The last column assumes providers and employers are myopic.

7.3 Robustness

I estimate my model under two alternative assumptions. First, my overall estimation strategy is to find variation in the data that plausibly captures markups, leaving the residual as production costs. Using this approach, I may underestimate the level and dispersion of markups. To address this concern, I make an alternative assumption that there are no unobserved production costs. Under this assumption, I estimate the random coefficient for fee sensitivity by matching the variance of markups simulated from my structural model to the variance of residualized fees, instead of using the variance of fee changes from switching providers. See Appendix G.7 and Appendix G.8 for a comparison of auxiliary models under these two assumptions.

In Table A12, I report estimates under the assumption of no unobserved production costs. I estimate similar parameters for large employers in the fourth quartile. Even based on my baseline assumptions, I estimate a very small variation in unobserved production costs. It seems plausible that providers incur similar production costs once employers achieve certain economies of scale. For example, providers may always need to designate a dedicated staff member to work with a large employer, despite different plan features or services offered.

For smaller employers in the bottom three quartiles, I estimate that fee sensitivities are around 50% lower. Because the level of markups is a decreasing and convex function of fee sensitivity based on Equation (10), a lower fee sensitivity leads to a higher level of markups and also a larger variance of markups.

In the second column of Table 4, I show that markups make up a slightly higher percentage of plan fees (24% instead of 18% under the baseline assumption). The variation in markups due to fee sensitivity also explains a greater fraction of fee dispersion (63% instead of 38%). Even with potential upper bounds of markups, there is still a meaningful amount of variation in plan fees not explained by employers' sophistication or negotiation effort.

Second, I assume employers and providers are myopic with discount factors equal to zero. Myopic providers do not consider future profits when they set markups at current RFPs. Myopic employers trade off transaction costs against fee savings in the current period only.

In Table A14, I report estimates under the myopic assumption. While I can also rationalize the data (see Table A15 for a comparison of goodness of fit), I obtain different parameter estimates. Since myopic employers only focus on fee differences in the current period, rationalizing infrequent provider switches requires lower transaction costs. I estimate that employers in the first quartile face 2 bps of RFP costs and 13 bps of switching costs. In dollar terms, the costs of running RFPs and switching providers are only \$1,000 and \$9,000, which seem unreasonably small. For large employers in the fourth quartile, the costs of running RFPs are only \$5,000, which seems negligible for employers with on average of nearly 3,000 employees. When comparing the magnitudes of transaction costs, it seems more sensible to assume that employers and providers are forward-looking.

With the myopic assumption, I still estimate a similar level of markups as a percentage of plan fees, according to the last column of Table 4. The variation in markups driven by fee sensitivity explains an even smaller fraction of the fee dispersion (28% instead of 38%).

8 Counterfactual

Using model estimates, I simulate counterfactual policies aimed at mitigating various drivers of plan fees. To evaluate counterfactual outcomes, I focus on plan fees paid by participants and transaction costs incurred by employers. If employers have to take on substantially higher transaction costs, these policies are somewhat unpractical even if they can reduce plan fees for participants. However, I do not add fees and transaction costs together, because my measurement of transaction costs could potentially capture both monetary costs, time costs, and psychological inertia.[31] For fees and transaction costs, I report both averages across employers and also the aggregated dollar amount. Since my estimation sample is a subset of all plans, the total impact of these counterfactual policies could be a factor of three to five according to Table A1.

I report outcomes under the status quo in the first row of Table 5. Pooling employers across size quartiles, the average plan fee is 65 bps or $5.8 billion annually. Employers face 6 bps or $180 million transaction costs per year, where transaction costs are calculated by summing the probability of RFP times RFP costs and the probability of switching times switching costs. In Figure 7, the black square marker captures plan fees (y-axis) and transaction costs (x-axis) under the status quo.

Table 5: Plan Fees and Transaction Costs under Counterfactual Policies

	Plan fees		Transaction costs	
Counterfactual	(bps)	($ billion)	(bps)	($ billion)
Status quo	64.5	5.80	5.9	0.18
High fee sensitivity (90th pct)	54.3	5.02	11.0	0.34
Consolidate plans, production cost only	53.8	5.52	5.9	0.18
Consolidate plans	48.7	5.40	0.6	0.07
Employers ignore transaction costs	64.3	5.41	58.7	2.89
RFP/switch that minimize fees	60.6	5.32	18.7	1.21

Table reports outcomes under the status quo and counterfactual policies. I report both average plan fees or transaction costs measured in basis points across employers and also the aggregated dollar amount of annual plan fees or transaction costs.

8.1 Increasing Employer Fee Sensitivity

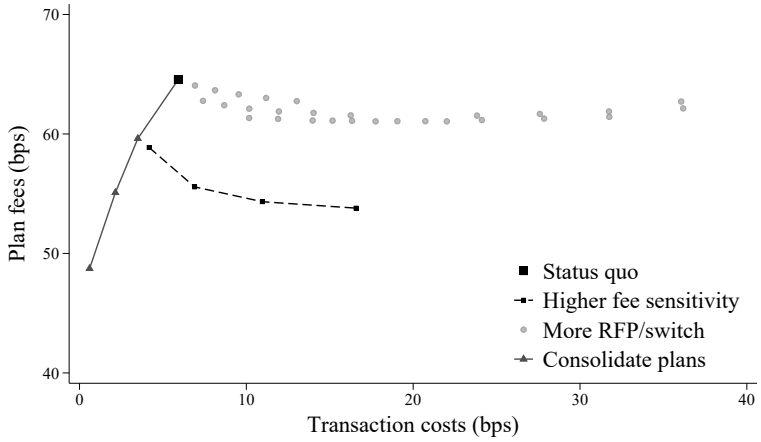
To assess the impact of employer fee sensitivity, I simulate a counterfactual where employers become more fee-sensitive. In Figure 7, the dashed black line with square markers displays outcomes when I set the fee sensitivity parameter α_i for all employers within each size quartile equal to various percentiles based on estimated distributions. From

[31] I also do not compute consumer surplus, which can be hard to interpret in my setting. First, without an outside option of not offering 401(k) plans, the scale of consumer surplus is not identified. Next, I assume employer preferences for provider quality represent their participants' preferences. In practice, employers may request providers to offer certain types of funds or services that their participants do not value.

top left to bottom right, square markers correspond to α_i at the 50th, 75th, 90th, and 95th percentiles. Although higher fee sensitivity can reduce markups, employers incur more transaction costs as they are more willing to conduct RFPs and switch providers to reduce fees. The second row of Table 5 reports outcomes where fee sensitivity is set at the 90th percentile. Plan fees decrease by 10 bps on average or almost $800 million annually. Meanwhile, employers take on 5 bps higher transaction costs or $160 million.

If recent excessive fee lawsuits push employers to become more fee-sensitive, participants may benefit from reduced fees without substantially higher transaction costs for employers. However, I caution against over-interpreting the results of this counterfactual, as changing employer fee sensitivity can be difficult in practice.

Figure 7: Counterfactual Fee Changes



Panels (b) and (c) plot counterfactual plan fees across different transaction costs. The y-axis corresponds to average plan fees, and the x-axis corresponds to the probability that employers run RFPs and switch providers in a given year. Moving from left to right, I reduce both RFP and switching costs from the status quo.

8.2 Consolidating Plans

Next, I consider consolidating 401(k) plans of small employers, motivated by regulations around multiple employer plans (MEPs) in the SECURE Act of 2019. My estimation suggests that having many small employers with limited scale and a lack of sophistication offer 401(k) plans to their respective employees is potentially inefficient. In addition to higher production costs and markups, it is also wasteful for many employers to each incur their own transaction costs.

To quantify the benefits of scale, the solid gray line with triangle markers in Figure 7 illustrates plan fees and transaction costs as I incrementally group employers in smaller size quartiles such that they are as large as employers in the second, third, and fourth quartiles. Note that I simply show different weighted averages of status quo results across different size quartiles, without recomputing model equilibrium. Consolidation

can save both plan fees for participants and transaction costs for employers.

The third and fourth rows of Table 5 display outcomes when I consolidate employers in the bottom three quartiles. In the third row, I only account for the effect of reducing transaction costs, assuming small employers still maintain their fee sensitivity and incur their own transaction costs. Production costs alone can reduce fees by over 10 bps and saving participants $300 million annually. In the fourth row, I also assume employers adopt the fee sensitivity and transaction costs of large employers. As small employers consolidate together their 401(k) plans, they potentially have the resources to hire professionals who are more sophisticated and put more effort in negotiating with providers. Plans fees decrease by 16 bps on average or $400 million in aggregate. Transaction costs also decrease by 5 bps on average or $110 million. Since menu update is important to reduce fees, it would be more efficient to incur necessary transaction costs once and spread the benefit of fee savings to as many participants as possible.

This counterfactual suggests that consolidating small 401(k) plans could achieve substantial savings in both plan fees and transaction costs. The SECURE Act of 2019 made multiple employer plans (MEPs) less restrictive and more attractive for small employers. Previously, employers have to share a common bond (e.g. in the same industry) and the entire MEP could be penalized if one employer within the group violates its fiduciary duty (the "bad apple" rule).

Bhattacharya and Illanes (2022) also study this counterfactual and find plan consolidation leads to minimal fee savings without modifying employer preferences, potentially because we focus on different drivers of markups and make different assumptions to estimate costs across different sizes of employers. In addition, a multiple employer plan could come with a lower fee sensitivity, higher production costs, and higher transaction costs than a large employer with comparable size due to the complexity in its organization structure (Chen and Munnell, 2024). As a result, I may be overestimating the benefits of consolidation. Furthermore, my model is not capturing potential supply size effects. Some providers may rely on small employers to stay profitable. If plan consolidations lead to provider exits, the remaining providers may have higher marker power and can charge higher markups than my estimates.

8.3 More frequent RFPs and Provider Switches

Next, I consider counterfactual scenarios where employers behave as if transaction costs are lower, such that they run RFPs or switch providers more often. I simulate model equilibrium by reducing RFP and/or switching cost parameters but measure outcomes using transaction costs under the status quo. This counterfactual reflects changes in employer behaviors rather than, for example, IT advancements that streamline the RFP and switching process.

Following the literature on switching costs since Klemperer (1987a,b), reducing trans-

action costs has two opposing effects on fees when providers engage in an "invest-harvest" pricing strategy. On one hand, providers can no longer exploit transaction costs to "harvest", by charging markups at RFPs and delaying menu updates until the next RFPs. On the other hand, forward-looking providers anticipate lower future markups and have small incentives to "invest". As a result, they react by setting higher markups during current RFPs. Combining these two effects, the overall impact of lower transaction costs on average plan fees is theoretically ambiguous.

Gray scatters in Figure 7 correspond to counterfactuals where I reduce RFP costs and/or switching costs by different magnitudes. The U-shape curve illustrates the two opposing effects. Initially, small reductions in transaction costs reduce plan fees by mitigating the"harvest" incentive. As transaction costs decrease further, the "invest" effect becomes stronger, leading to higher plan fees.

Table 5 compares outcomes with different transaction costs. When employers behave as if there were no RFP or switching costs in the "Ignore transaction costs" counterfactual, average plan fees remain largely unchanged from the status quo, while aggregate fees decrease by $400 million. This discrepancy is because the "invest" and "harvest" incentives play different roles between small and large employers. Reducing transaction costs tends to reduce fees more for large employers. At the same time, employers incur substantially higher transaction costs.

As a comparison, the "RFP/switch that minimize fees" counterfactual corresponds to the bottom of the U-shape curve of gray scatters in Figure 7. At some modest transaction costs, providers still retain some "invest" incentives that lead to further reductions in fees. Plan fees decrease by 4 bps on average or $ 500 million in aggregate.

Dubé et al. (2009) and Cabral (2016) also show that modest switching costs can reduce prices by incentivizing competition, which is consistent with my findings that fees are minimized at some modest levels of transaction costs. Intuitively, in a static problem, if providers reduce markups from the static optimal level, they increase their probability of being chosen by the employer at the RFP, but conditional on winning their markups are lower. However, with dynamic considerations, providers consider the present value of markups over multiple future periods. Therefore, they will reduce markups below the static optimal level. Hence, some transaction costs can be beneficial to incentivize competition among providers.

In recent excessive fee lawsuits, some participants accuse their employers of not running RFPs regularly or staying with the same providers for a very long period of time. [32] My results suggest that more frequent RFPs or switching may not necessarily achieve as large reductions in plan fees as one anticipates due to providers' dynamic fee-setting incentive.

[32]These allegations are typically not central to the cases but serve as suggestive evidence of employer negligence. See *Singh v. Deloitte LLP* for a recent example.

9 Conclusion

In this paper, I demonstrate that transaction costs are important to understand what determines the level and the dispersion of 401(k) plan fees. Despite the introduction of low-fee mutual funds, providers can exploit employers' transaction costs and delay the inclusion of these funds in 401(k) menus, effectively earning higher markups over time. I show that the majority of the fee decline between 2009 and 2019 comes from updating menus to include cheaper funds. When employers incur transaction costs to switch providers, they can achieve substantial fee reductions.

The descriptive evidence motivates my dynamic model to capture how employers and plan providers interact. Fee variation when employers switch providers allows me to use this model to estimate markups. I find that differences in markups between employers with varying fee sensitivity do not account for most of the observed fee dispersion. Much of the variation in fees can be natural outcomes in a negotiated-price market with transaction costs. In some recent excessive fee lawsuits, plaintiffs rely on simple fee comparisons without conditioning on menu composition and services. My results suggest that much of this fee variation can be natural outcomes in a negotiated-price market with transaction costs.

I use model estimates to quantify whether counterfactual policies can effectively reduce plan fees. I show that consolidating plans could achieve substantial fee savings through both reducing production costs thanks to economies of scale and also higher fee sensitivity of larger employers. The effects of reducing transaction costs on fees are more complex due to providers' dynamic incentives. Ignoring transaction costs does not reduce plan fees for the average employer relative to the status quo, because providers no longer set low markups to incentivize employers to switch. Interestingly, some modest transaction costs lead to more fee reductions than no transaction costs.

The mechanism by which providers maintain sticky prices relative to declining costs to sustain markups extends beyond 401(k) plans. A similar dynamic can be observed in fixed-rate mortgages during periods of declining interest rates. Hence, my empirical results and structural model can be applied more broadly.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High wage workers and high wage firms," *Econometrica*, 67, 251–333.

Adams, N. (2020): "Reader Poll: Has Litigation Led to Change?" *National Association of Plan Advisors*.

——— (2022): "Reader Radar: Revisiting and Revising RFPs," *National Association of Plan Advisors*.

Allen, J., R. Clark, and J.-F. Houde (2014): "The effect of mergers in search markets: Evidence from the Canadian mortgage industry," *American Economic Review*, 104, 3365–96.

——— (2019): "Search frictions and market power in negotiated-price markets," *Journal of Political Economy*, 127, 1550–1598.

Allen, J. and S. Li (2020): "Dynamic competition in negotiated price markets," Tech. rep., Bank of Canada Staff Working Paper.

Arcidiacono, P. and R. A. Miller (2011): "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity," *Econometrica*, 79, 1823–1867.

Aronowitz, D. (2022a): "Debunking Recordkeeping Fee Theories in Excessive Fee Cases," *Euclid Fiduciary Whitepaper*.

——— (2022b): "A Deep Dive into the Hy-Vee Excessive Fee Case," *Euclid Fiduciary*.

——— (2023): "The Key Fiduciary Liability Storylines of 2022," *Euclid Fiduciary*.

Bachas, N. (2018): "The impact of risk-based pricing and refinancing on the student loan market," Tech. rep., SIEPR Working Paper.

Badoer, D. C., C. P. Costello, and C. M. James (2020): "I can see clearly now: The impact of disclosure requirements on 401 (k) fees," *Journal of Financial Economics*, 136, 471–489.

Begenau, J. and E. Siriwardane (2022): "Fee Variation in Private Equity," *Journal of Finance, Forthcoming*.

Benartzi, S. and R. Thaler (2007): "Heuristics and biases in retirement savings behavior," *journal Of Economic Perspectives*, 21, 81–104.

Berry, S., J. Levinsohn, and A. Pakes (1995): "automobile Prices In Market Equilibrium," *econometrica*, 63, 841–90.

Beshears, J., J. J. Choi, D. Laibson, and B. C. Madrian (2009): *The importance of default options for retirement saving outcomes: Evidence from the United States*, University of Chicago Press.

Bhattacharya, V. and G. Illanes (2022): "The design of defined contribution plans," Tech. rep., National Bureau of Economic Research.

Bhattacharya, V., G. Illanes, and M. Padi (2019): "Fiduciary duty and the market for financial advice," Tech. rep., National Bureau of Economic Research.

Brancaccio, G., D. Li, and N. Schürhoff (2017): "Learning by trading: The case of the US market for municipal bonds," *Unpublished paper. Princeton University*.

Cabral, L. (2016): "Dynamic pricing in customer markets with switching costs," *Review of Economic Dynamics*, 20, 43–62.

Carroll, G. D., J. J. Choi, D. Laibson, B. C. Madrian, and A. Metrick (2009): "Optimal defaults and active decisions," *The Quarterly Journal of Economics*, 124, 1639–1674.

Chalmers, J., O. S. Mitchell, J. Reuter, and M. Zhong (2021): "Auto-enrollment retirement plans for the people: Choices and outcomes in oregonsaves," Tech. rep., National Bureau of Economic Research.

Chalmers, J. and J. Reuter (2020): "Is conflicted investment advice better than no advice?" *Journal of Financial Economics*, 138, 366–387.

Chen, A. and A. Munnell (2024): "Will Multiple Employer Plans Help Close the Coverage Gap?" *Center for Retirement Research at Boston College*.

Choi, J. J. (2015): "Contributions to defined contribution pension plans," *Annual Review of Financial Economics*, 7, 161–178.

Choi, J. J., D. Laibson, B. C. Madrian, and A. Metrick (2002): "Defined contribution pensions: Plan rules, participant choices, and the path of least resistance," *Tax Policy and the Economy*, 16, 67–113.

Choi, J. J., D. Laibson, B. C. Madrian, A. Metrick, and J. M. Poterba (2007): *2. For Better or for Worse: Default Effects and 401(k) Savings Behavior*, University of Chicago Press, 81–126.

Clark, R., J.-F. Houde, and J. Kastl (2021): "The industrial organization of financial markets," in *Handbook of Industrial Organization*, Elsevier, vol. 5, 427–520.

Cuesta, J. I. and A. Sepúlveda (2021): "Price regulation in credit markets: A trade-off between consumer protection and credit access," *Available at SSRN 3282910*.

Deloitte (2013): "Inside the Structure of Defined Contribution/401(k) Plan Fees," .

——— (2015): "Annual Defined Contribution Benchmarking Survey. Ese of Use Drives Engagement in Saving for Retirement," .

——— (2019): "The retirement landscape has changed-are plan sponsors ready? 2019 Defined Contribution benchmarking Survey Report," .

Doellman, T. W. and S. H. Sardarli (2016): "Investment fees, net returns, and conflict of interest in 401 (k) plans," *Journal of Financial Research*, 39, 5–33.

Dubé, J.-P., G. J. Hitsch, and P. E. Rossi (2009): "Do switching costs make markets less competitive?" *Journal of Marketing research*, 46, 435–445.

Duflo, E., W. Gale, J. Liebman, P. Orszag, and E. Saez (2006): "Saving incentives for low-and middle-income families: Evidence from a field experiment with H&R Block," *The Quarterly Journal of Economics*, 121, 1311–1346.

Dworak-Fisher, K. (2011): "Matching matters in 401 (k) plan participation," *Industrial Relations: A Journal of Economy and Society*, 50, 713–737.

Egan, M. (2019): "Brokers versus retail investors: Conflicting interests and dominated products," *The Journal of Finance*, 74, 1217–1260.

Egan, M. L., A. MacKay, and H. Yang (2022): "recovering Investor Expectations From Demand For Index Funds," *the Review Of Economic Studies*, 89, 2559–2599.

——— (2023): "What drives variation in investor portfolios? evidence from retirement plans," Tech. rep., National Bureau of Economic Research.

Einav, L., M. Jenkins, and J. Levin (2012): "Contract pricing in consumer credit markets," *Econometrica*, 80, 1387–1432.

Farrell, J. and P. Klemperer (2007): "Coordination and lock-in: Competition with switching costs and network effects," *Handbook of industrial organization*, 3, 1967–2072.

Gourieroux, C. and A. Monfort (1996): *Simulation-based econometric methods*, Oxford university press.

Gruber, M. J. (1996): "Another puzzle: The growth in actively managed mutual funds," *The journal of finance*, 51, 783–810.

Grunewald, A., J. A. Lanning, D. C. Low, and T. Salz (2020): "Auto dealer loan intermediation: Consumer behavior and competitive effects," Tech. rep., National Bureau of Economic Research.

Guiso, L., A. Pozzi, A. Tsoy, L. Gambacorta, and P. E. Mistrulli (2022): "The cost of steering in financial markets: Evidence from the mortgage market," *Journal of Financial Economics*, 143, 1209–1226.

Hortaçsu, A. and C. Syverson (2004): "Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds," *The Quarterly journal of economics*, 119, 403–456.

Hotz, V. J. and R. A. Miller (1993): "Conditional choice probabilities and the estimation of dynamic models," *The Review of Economic Studies*, 60, 497–529.

ICI (2022a): "The BrightScope/ICI Defined Contribution Plan Profile: A Close Look at 401(k) Plans, 2019," .

——— (2022b): "Trends in the Expenses and Fees of Funds, 2021," *Investment Company Institute Research Perspective*, 28.

Illanes, G. (2016): "Switching costs in pension plan choice," *Unpublished manuscript.*

Illanes, G. and M. Padi (2019): "Retirement policy and annuity market equilibria: Evidence from chile," Tech. rep., National Bureau of Economic Research.

Klemperer, P. (1987a): "The competitiveness of markets with switching costs," *The RAND Journal of Economics*, 138–150.

——— (1987b): "Markets with consumer switching costs," *The quarterly journal of economics*, 102, 375–394.

——— (1995): "Competition when consumers have switching costs: An overview with applications to industrial organization, macroeconomics, and international trade," *The review of economic studies*, 62, 515–539.

Koijen, R. S. and M. Yogo (2016): "Shadow insurance," *Econometrica*, 84, 1265–1287.

Luco, F. (2019): "Switching costs and competition in retirement investment," *American economic journal: Microeconomics*, 11, 26–54.

Madrian, B. C. and D. F. Shea (2001): "The power of suggestion: Inertia in 401 (k) participation and savings behavior," *The Quarterly Journal of Economics*, 116, 1149–1187.

McFadden, D. (1973): "Conditional logit analysis of qualitative choice behavior," .

Mellman, G. S. and G. T. Sanzenbacher (2018): "401 (k) lawsuits: What are the causes and consequences?" *Issue in Brief*, 18–8.

Morningstar (2021): "2021 U.S. Fund Fee Study," *Morningstar Manager Research*.

Nelson, S. (2018): "Private information and price regulation in the us credit card market," *Unpublished Working Paper*.

NEPC (2015): "NEPC 2015 Defined Contribution Plan & Fee Survey: What a Difference a Decade Makes," .

Pechter, K. (2022): "Let's Concentrate on Recordkeeping," *Retirement Income Journal*.

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

——— (2021): "Mutual Fund Revenue Sharing in 401 (k) Plans," *working Paper*.

Reuter, J. and D. P. Richardson (2022): "New evidence on the demand for advice within retirement plans," Tech. rep., National Bureau of Economic Research.

Robles-Garcia, C. (2019): "Competition and incentives in mortgage markets: The role of brokers," *Unpublished working paper*.

Rust, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 999–1033.

Sun, L. and S. Abraham (2021): "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 225, 175–199.

Sweeting, A., D. Jia, S. Hui, and X. Yao (2022): "Dynamic Price Competition, Learning-by-Doing, and Strategic Buyers," *American Economic Review*, 112, 1311–33.

Yogo, M., A. Whitten, and N. Cox (2025): "Financial inclusion across the United States," *Journal of Financial Economics*, 166, 104003.

Online Appendix

A Data

A.1 Identifying Providers

According to Form 5500 Schedule C, an employer interacts with multiple providers such as recordkeepers, investment advisors, auditors, and legal consultants. I focus on the main provider with primarily recordkeeping functions. Employers report a list of service codes for each provider in Form 5500 Schedule C, and BrightScope consolidates those service codes to provider functions that help me identify the main recordkeeper.

I use a hierarchy of provider functions to designate the main recordkeeper. I first consider the provider that is "Recordkeeper". If there is no recordkeeper, I move on to "Trustee", "Custodian", and then "Administrator". The vast majority of employers have a recordkeeper following this sequence of functions.[33] My results are relatively insensitive to the precise sequence of services in this hierarchy.

Occasionally, an employer may hire multiple recordkeepers, and I use the following rules to determine the main recordkeeper. These rules err on the side of not introducing provider switches due to adding or removing one of the multiple recordkeepers. First, for each employer $\times$ provider, I define a tenure covering the consecutive years this provider performs recordkeeping functions. I drop a recordkeeper whose tenure is completely covered within the tenure of another recordkeeper. When multiple providers have the same tenures, I select the provider with higher compensation. Second, two recordkeepers overlap during the year of provider transition. In this case, I choose the recordkeeper with higher compensation. Only a few employers still have multiple providers after these two steps, and I drop these employers.

A.2 Identifying Provider Switches

After designating a unique recordkeeper for each provider $\times$ year, I define a provider switch when the employer changed its recordkeeper from the previous year. The precise year of recordkeeper switch may not correspond to the year with menu changes. For example, an employer may hire a new recordkeeper in late 2010 but revise the menu of funds in 2011. I use menu turnover from BrightScope plan menu data to revise switch timing. Around a window of plus and minus one year from the year of switch, I revise the year of switch to be the year with the highest menu turnover.

There has been some consolidation among recordkeepers that can result in recordkeeper changes. To adjust for these cases, I look at pairs of old and new providers

[33] Bhattacharya et al. (2019) used a similar procedure.

with high frequencies in the data and cross check with publicly available information about recordkeeper mergers or acquisitions. These include Great-West's mergers with the defined contribution business of Putnam Investments and JP Morgan Retirement Plan Services and rebranding as Empower in 2014. Other consolidations are between smaller providers in the other categories. When recordkeeper A bought B in year T, I reassign the recordkeeper from B to A prior to T and do not consider this as a provider switch.

A.3 Provider Compensation

BrightScope collects provider direct compensation data from Form 5500 Schedule C. Providers receive both direct compensation from employers and indirect compensation through revenue sharing from mutual fund companies. Because I abstract away from separately modeling integrated versus third party recordkeepers, I consider the total fees paid by employers and employee participants. The total fees include both mutual fund expense ratios and direct compensation for recordkeeping services.

Direct compensation can be a fixed dollar amount per participant or a percentage of invested assets, but Form 5500 only reports the total amount of compensation. I convert the amount of direction compensation to a fraction of total plan assets. If an employer switches to a new provider in the middle of a year, the amount of compensation.

The compensation from Form 5500 may include services other than recordkeeping (Aronowitz, 2022a). Recordkeepers may also offer consulting, advisory, or legal services, and these compensations are not separately reported. In my analysis, I control for other provider functions. But those information are coarse and the control is likely not sufficient. In addition, recordkeeping compensation also includes commissions for individual trades, loan-origination, Qualified Domestic Relations Order (QDRO) related fees that are charged to the individual participants, rather than on a plan-wide basis. For example, in *Matousek v MidAmerican Energy Company*, plaintiffs claim a recordkeeping fee of $326 to $525 per participants computed from Form 5500 Schedule C over five years. But the employer disputed that the recordkeeping fee was only $37.

B Drivers of Mutual Fund Fee Decline

The decline in mutual fund fees over the past 25 years has contributed to the reduction in 401(k) fees. From 1996 to 2021, the average expense ratio of equity mutual funds decreased from 104 to 47 bps. The average expense ratios of hybrid and bond mutual funds fell from 95 to 57 bps and 84 to 39 bps, respectively (ICI, 2022b). This decline can be attributed to several factors (Morningstar, 2021; ICI, 2022b).

First, mutual funds benefit from economies of scale and scope. Certain costs such as transfer agency fees, director fees, accounting and audit fees, are relatively fixed in dollar terms, and these costs become increasingly negligible as funds gather more assets. In addition, the fixed operating costs of an asset manager also spread across a greater number of mutual funds, leading to cost savings.

Second, investors have become more fee-conscious over the past two decades. There has been a growing preference for index funds, which replicate the returns of market indices. Index funds are generally cheaper to manage compared to actively managed funds since they do not require extensive research on underlying companies, and their portfolio turnovers are lower. Even within the respective active and index fund universe, investors tend to gravitate towards the lowest-cost funds, resulting in increased competition and fee reduction.

Third, the popularity of no-load funds has contributed to the decline in expense ratios. Investors increasingly compensate for financial advice separately and prefer funds without sales charges. This explanation is less relevant for 401(k) plans since most funds offered in 401(k) menus are no-load share classes.

In addition to mutual fund fees, recordkeeping fees in 401(k) plans have also declined over time. Average recordkeeping fees have decreased, driven by improvements in IT technology, the scale of recordkeepers, and some standardization of recordkeeping services (Pechter, 2022). Based on its survey, NEPC (2015) shows that the average recordkeeping fees declined from 57 to 46 bps over plan assets or from $118 to $64 per participant from 2006 to 2015.

C Comparison of Plan Fees with Government Sponsored Defined Contribution plans

In this appendix section, I compare plan fees from employer sponsored 401(k) plans with government-sponsored defined contribution plans. Started in 2017, OregonSaves is one of the earliest state-sponsored defined contribution plans for self-employed and employees at small employers that do not offer 401(k) plans. This program charges participants around 50 bps each year. There is another $16 annual fee, which covers higher administrative costs due to low account balances (the average account balance

of $1324 as of 2022) as participants have yet accumulated large amounts of savings in this new program. On the other hand, Federal Thrift Saving Plans (TSP) for federal and civil services employees charge participants 5.7 to 9 bps. TSP is the largest defined contribution plan in the world, and the low fees can reflect its economies of scale.[34]

To compare government sponsored programs with 401(k) plans with comparable scale, I fit a linear relationship between log plan fee and the log of total plan assets using my 401(k) data as of 2019, In Figure A1, I show the estimated plan fee - total asset curve and indicate fees charged by OregonSaves and TSP. The comparison shows that plans with comparable total assets (with the curve extrapolated out of sample to compare with TSP) would charge similar plan fees as these government-sponsored programs. Assuming these government bodies ensured that providers do not charge an unrealistic amount of markup, employers are likely to offer competitive fees to plan participants on average.

Figure A1: Plan fee comparison

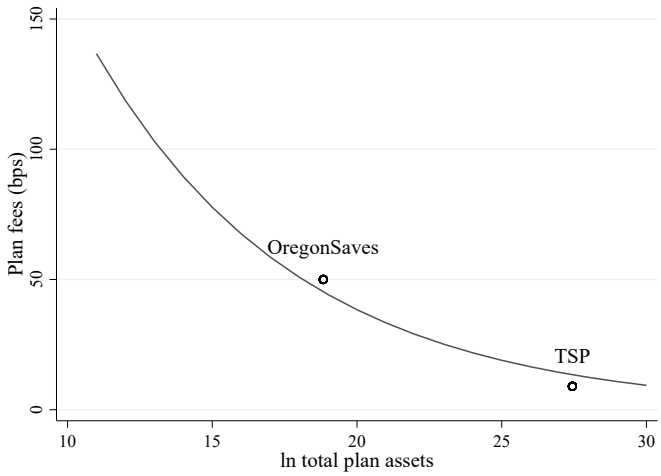


Figure plots curve of total plan fee versus log of total plan assets, estimated from $\ln \text{fee} = \beta_0 + \beta_1 \times \ln \text{plan asset}$. Plan cost and total assets of OregonSaves and TSP are displayed as comparisons.

[34] OregonSaves program detail: https://www.oregonsaves.com/savers/program-details. OregonSaves account statistics are based on https://crr.bc.edu/wp-content/uploads/2022/07/Oregon_July-2022.pdf. Also see Chalmers et al. (2021) for an analysis of OregonSaves. Other state-sponsored defined contribution plans, for example, CalSavers, have similar cost structures. TSP program costs https://www.tsp.gov/tsp-basics/expenses-and-fees/ and account balance statistics from https://www.frtib.gov/pdf/reading-room/FinStmts/TSP-FS-Dec2021.pdf

D Additional Summary Statistics

Table A1: Sample coverage as of 2019

	All plans	1st quartile	2nd quartile	3rd quartile	4th quartile
Number of plans	630,724	473,448	18,055	11,075	9,715
Coverage by number of plans					
Has plan menu (%)	9	5	65	86	94
Estimation sample (%)	3	1	28	41	38
Coverage by plan assets					
Has plan menu (%)	84	18	70	89	95
Estimation sample (%)	17	4	29	42	18

Table A2: Summary Statistics Across Employer Size Quartile

	1st quartile	2nd quartile	3rd quartile	4th quartile
Plan participants	306	398	624	2,703
Total plan assets (million)	6	16	36	224
Average account balance (thousand)	33	67	110	164
Fee				
Plan average expense ratio (bps)	52	48	46	40
Direct recordkeeping fees (bps)	27	18	13	8
Total plan fees (bps)	79	66	58	49
Number of funds	26	27	26	25
Menu turnover (%)	14	14	13	11
Switched provider (%)	5	5	4	3

Figure A2: Employer Plan Size Distributions

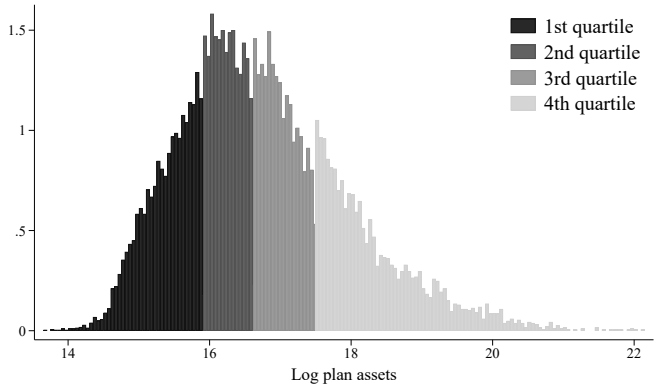


Figure shows the distribution of log average plan assets across employers in each size quartile.

Table A3: Provider Specific Summary Statistics

	Fidelity	Vanguard	Empower	Principal Fin	ADP	Others
Market share (%)	27	6	10	2	7	48
Plan participants	966	1,182	702	763	433	884
Total plan assets (million)	69	138	36	37	17	53
Average account balance (thousand)	90	149	75	83	51	82
Fee						
Plan average expense ratio (bps)	53	18	43	47	57	47
Direct recordkeeping fees (bps)	9	10	18	28	16	23
Total plan fees (bps)	62	29	61	75	73	70
Number of funds	28	27	26	24	27	25
Plan menu						
Fraction of index funds (%)	27	75	38	30	31	33
Menu turnover (%)	13	8	14	15	17	13
Fraction of proprietary funds (%)	50	88	0	16	0	10

E Additional Motivating Evidence

E.1 Decline in Mutual Fund Fees

I explain the mutual fund fee decomposition details in Figure 4. Let p_{ft} denote the expense ratio of fund f in year t. Let w_{ift} denote the fraction of assets of employer i's plan allocated to fund f in year t. Let ι_{ft} denote whether fund f is available in the marketplace as of year t. The fund is not available if its fund manager has not yet introduced this fund or if the fund has been delisted. I keep a balanced panel of employers for this analysis. The observed fee decline (black solid line) and the decline corresponding to changes in expense ratios (dashed line with circle markers) are constructed as follows

$$p_t^{\text{observed}} = \frac{\sum_{if} w_{ift} \times p_{ft}}{\sum_{if} w_{ift}}$$

$$p_t^{\Delta\text{fee in 2009 menu}} = \frac{\sum_{if} w_{if,2009} \times \iota_{ft} \times p_{ft}}{\sum_{if} w_{if,2009} \times \iota_{ft}}$$

$$p_t^{\Delta\text{fee in 2019 menu}} = \frac{\sum_{if} w_{if,2019} \times \iota_{ft} \times p_{ft}}{\sum_{if} w_{if,2019} \times \iota_{ft}}$$

The gray line in Figure 4 allows participants to rebalance among funds in 2009 menus. I consider a simple calculation to approximate rebalancing. I first hold allocation to each asset class fixed. These are nine broad asset classes: large/mid/small cap equities, international equities, bonds, alternatives, allocation funds, and target date funds. I assume, for example, a participant close to retirement will not shift from lower-risk bond funds to higher-risk equity funds because of a few basis point differences in expense ratios. Within each asset class, I allocate assets proportionally based on observed allocation but within funds included in 2009 menus. When all funds in 2009 menus are removed in an asset class, I extrapolate based on the average rate of fee decline within each asset class.

Let $c(f)$ denote the asset class of fund f. The asset class average fee after rebalancing $\hat{p}_{ict}$ is

$$\hat{p}_{ict} = \begin{cases} \frac{\sum_{f:c(f)=c} p_{ft} \times w_{ift} \times \iota_{ft}}{\sum_{f:c(f)=c} w_{ift} \times \iota_{ft}} & \text{if } \sum_{f:c(f)=c} w_{ift} \times \iota_{ft} > 0 \\ \hat{p}_{ic,t-1} + \Delta p_c & \text{if } \sum_{f:c(f)=c} w_{ift} \times \iota_{ft} = 0 \end{cases}$$

Let w_{ict} denote the fraction of total balance that participants in employer i allocate to asset class c at time t. The gray line in Figure A3 is constructed as

$$p_t^{\Delta\text{fee in 2009 menu \& rebalance}} = \frac{\sum_{ic} \hat{w}_{ict} \times \hat{p}_{ict}}{\sum_{ic} \hat{w}_{ict}}$$

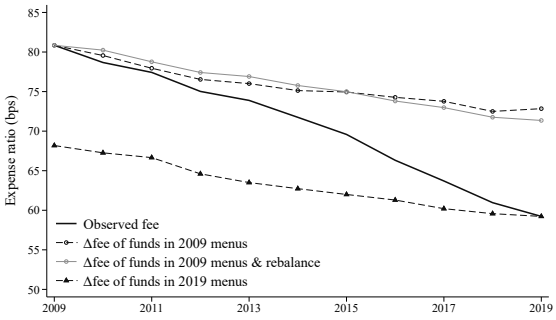
I have also estimated a demand system allowing time-varying fee sensitivities and

used this demand system to predict counterfactual allocation using the menu as of 2009. I obtained similar results to the gray line displayed in Figure 4.

Figure A3 displays the decomposition separately for active and index funds. Employers can reduce fees from updating menus toward cheaper and comparable funds within the active and index universe.

Figure A3: Decline in Mutual Fund Fees

(a) Active funds



(b) Index funds

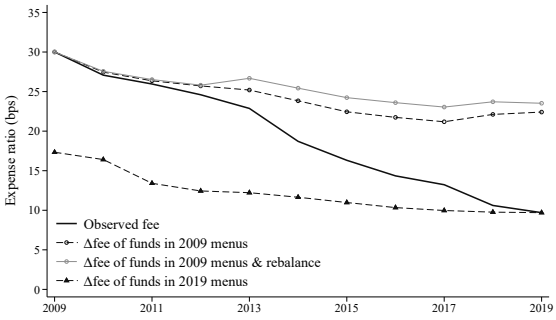


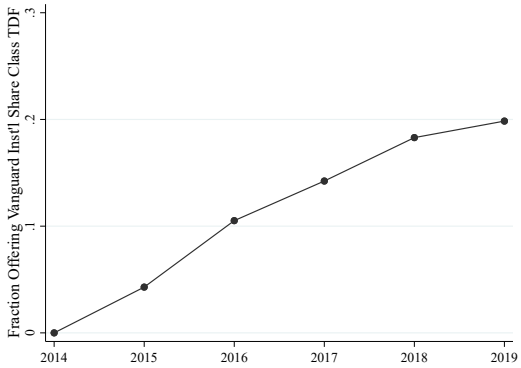
Figure shows the decomposition of mutual fund fee changes separately for active and index funds.

E.2 Providers Do Not Immediately Update Plan Menus

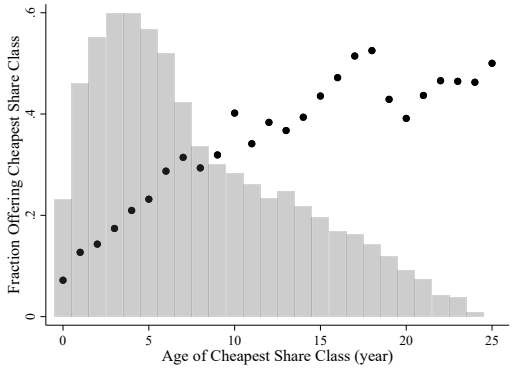
In this section, I show that providers do not immediately include cheaper funds when they become available.

Figure A4: Slow Adoption of Cheapest Share Class

(a) Offer Vanguard Inst'l TDF



(b) Offer Cheapest Share Class



Panel (a) shows the fraction of employers who offer Vanguard Target Date Funds (TDFs) that include cheaper institutional share class since its introduction in 2015. Panel (b) displays the fraction of employers who offer the cheapest share class over the age of the cheapest share class (initial offer year minus the current year).

I start with a case study of when Vanguard introduced institutional share classes for its target date fund (TDF) series in 2015. The institutional share class had an expense ratio of 10 bps, but the pre-existing investor share class had expense ratios

ranging from 14 to 16 bps, depending on the retirement target date. Figure A4a shows that employers who offer Vanguard TDFs adopt the lower-cost institutional share class gradually, reaching 20% in 4 years. The institutional share class comes with higher investment limit. I restrict to employers whose participants invest over $5 million into Vanguard TDFs according to the investment limit[35]

I extend the analysis more broadly in Figure A4b. For each fund × year with multiple share class, I compute the fraction of employers who offer the cheapest share class, by the age of the cheapest share class (current year minus year when fund was first offered). To approximate investment limit, I look at the bottom 10th percentile of investment among plans who offer the cheapest share class in a given year. The gray bars show density for the age of cheapest share class across fund × year.

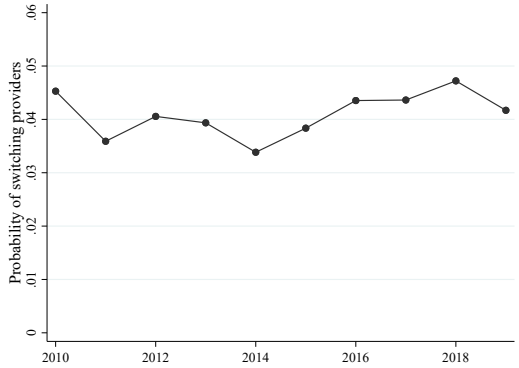
Less then 10% of employers adopt the cheapest share class when they are first introduced. Adoption increases gradually reaching around 50% after 15 years. Due to revenue sharing discussed in Section 2, employers and providers may deliberately offer share classes that have higher expense ratios but lower fees on a net basis, which explains why the fraction may never reaches 100%. The density shown by the gray bars illustrates that it is rather frequent for asset managers to issue a new share class with the lowest expense ratio. But these cheaper share classes are not included to 401(k) plan menus with a timely fashion.

[35] https://investor.vanguard.com/investor-resources-education/mutual-funds/share-classes-of-vanguard-mutual-funds

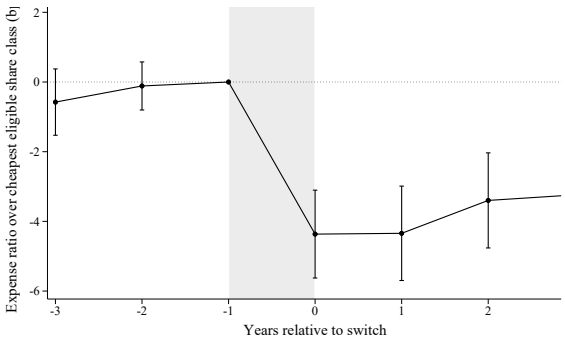
E.3 Additional Difference-in-Differences Results

Figure A5: Difference in Differences around Provider Switch

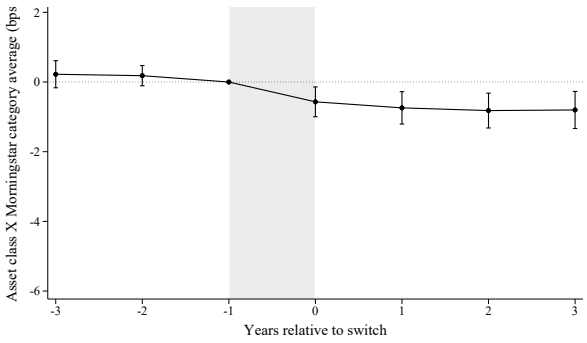
(a) Probability of Switching Providers



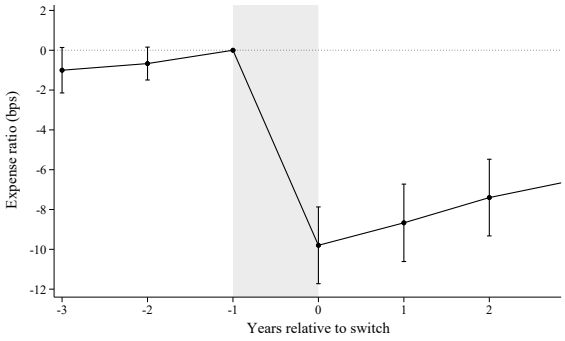
(b) Fee over Cheapest Eligible Share Class



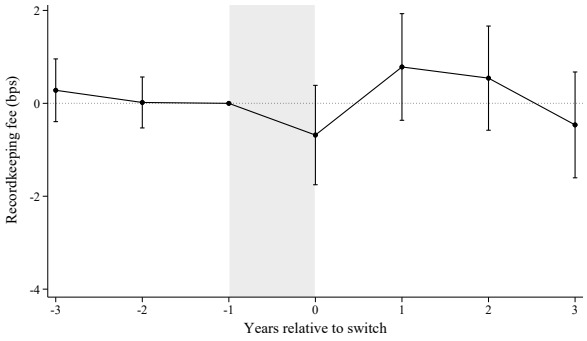
(c) Asset Class × Morningstar Category



(d) Expense ratio



(e) Recordkeeping Fee



Panels (a) and (b) follow the same specification as Figure 3, except that controls $X_{it}\beta$ only include plan number of participants. The outcome variable in (a) is fee over cheapest share class of the same mutual fund. To determine eligibility for the cheapest share class, I compare the asset invested in this mutual fund with the 10th percentile of asset invested across all employers who offer the cheapest share class in the same year. The outcome is unconditional: when the mutual fund does not have multiple share class or when the employer does not qualify, the outcome variable is zero for this employer × fund. The outcome variable in (b) is average fee across all mutual funds in the same asset class × Morningstar category. Panels (c) and (d) separately look at expense ratios and recordkeeping fees, as the two components of total plan fees.

F Model Implementation

F.1 Two-Stage RFP Simplification

I break the RFP into a two-stage process. In the first stage, her incumbent provider proposes a fee. The employer decides whether to accept her incumbent's offer. If she rejects, she incurs the switching cost and enters the second stage, where other providers compete with the incumbent. For simplicity, I assume the employer still incurs the switching costs by choosing the incumbent provider at the second stage. This captures cases when the previous incumbent is willing to match menus and services proposed by competing providers, and so there will be substantial changes even with the same provider.

This two-step process has two computational benefits. The first is dimension reduction. In the dynamic price setting game described in Section 5.3, each provider j sets an optimal fee conditional on the previous incumbent s identity. Suppose there are J providers in total. I need to solve for J^2 optimal RFP fee. With the two-stage RFP simplification, all second-stages of RFPs are homogeneous and no longer depend on the previous incumbent s. Each incumbent provider still sets different fees in the first stage. So there are in total $2J$ optimal fees, instead of J^2. This simplification abstracts away from the variation in fees and transition probabilities across providers. However, since there are relatively few observations of provider switches, this variation may contain too much noise.

Second, all provider switches occur at the second stages of RFPs only. Since I observe provider choice probabilities conditional on switching, I can use them as CCP (Hotz and Miller, 1993) and do not need to estimate provider heterogeneity parameters ξ_j (except for providers in the other categories). I discuss this point later in Appendix F.2.

At the second stage of RFP, each provider sets a fee b_j^{2nd}[36]. These fees no longer depends on the identity of the previous incumbent. Employer i perceives the following utility from provider j:

$$\mathcal{U}_j^{2nd} + \epsilon_{ij} = -\alpha b_j^{2nd} + \nu_j + \delta \mathbb{E}V^E(j, b_j^{2nd}) + \epsilon_{ij} \tag{A1}$$

The probability that the employer chooses provider j is given by

$$q_j^{2nd} = \frac{e^{\mathcal{U}_j^{2nd}}}{\sum_k e^{\mathcal{U}_k^{2nd}}} \tag{A2}$$

Employer's expected utility from the second stage of RFP follows the standard in-

[36] In the main text, my notation conditions on employer i to be more general, but it should become clear that the condition on i is redundant since employer specific quality ν_i and cost c_i shifters are canceled out. So for simplicity in this appendix section, I drop the i subscript.

clusive value expression

$$\mathbb{E}\mathcal{U}^{2nd} = \ln \sum_k e^{\mathcal{U}_k^{2nd}} \tag{A3}$$

Provider's ex-ante value function at the second stage of RFP depends on the probability of winning and the continuation values as the winner providing services or the loser waiting for the next RFP.

$$V_j^{P,2nd} = q_j^{2nd}\Big(b_j^{2nd} - c_j + \underbrace{\delta \mathbb{E}V_j^W(j, b_j^{2nd})}_{\substack{\text{Winner} \\ \text{continuation value}}}\Big) + \sum_{k \neq j} q_k^{2nd} \underbrace{\delta \mathbb{E}V_j^L(k, b_k^{2nd})}_{\substack{\text{Loser} \\ \text{continuation value}}} \tag{A4}$$

Providers set an optimal fee b_j^{2nd} to maximize $V_j^{P,2nd}$. Solve FOC with respect to b_j^{2nd}:

$$b_j^{2nd} = c_j \underbrace{-\delta\left(\mathbb{E}V_j^W(j, b_j^{2nd}) - \frac{\sum_{k \neq j} q_k \mathbb{E}V_j^L(k, b_k^{2nd})}{1 - q_j^{2nd}}\right)}_{\substack{\text{Incremental continuation value} \\ \text{as the RFP winner in 2nd stage}}} + \left(\frac{1 + \delta \overbrace{\frac{\partial \mathbb{E}V_j^W(j, b_j^{2nd})}{\partial b_j^{2nd}}}^{\substack{\text{Effect of } b_j^{2nd} \text{on} \\ \text{continuation value}}}}{\alpha(1 - q_j^{2nd})}\right) \tag{A5}$$

At the first stage of RFP, the incumbent provider j proposes fee b_j^{1st}. The utility from rejecting j's offer and proceeding to the 2nd stage depends on the expected utility from Equation (A3) and switching cost κ_{st}.

$$\begin{aligned} \mathcal{U}_{j1}^{1st} + \epsilon_{i1} &= -\alpha b_j^{1st} + \nu_j + \delta \mathbb{E}V^E(j, b_j^{1st}) + \epsilon_{i1} \\ \mathcal{U}_{j0}^{1st} + \epsilon_{i0} &= \mathbb{E}\mathcal{U}^{2nd} - \kappa_{sw} + \epsilon_{i0} \end{aligned} \tag{A6}$$

The probability of accepting j's offer and the employer's expected utility at the 1st stage of RFP are, respectively:

$$\begin{aligned} q^{1st}(j) &= \frac{e^{\mathcal{U}_{j1}^{1st}}}{e^{\mathcal{U}_{j1}^{1st}} + e^{\mathcal{U}_{j0}^{1st}}} \\ \mathbb{E}\mathcal{U}^{1st}(j) &= \ln\left(e^{\mathcal{U}_{j1}^{1st}} + e^{\mathcal{U}_{j0}^{1st}}\right) \end{aligned}$$

Ex-ante value function for provider k at the first stage depends on whether $k = j$ is the incumbent.

$$V_k^{P,1st}(j) = \begin{cases} q^{1st}(j)\Big(b_j^{1st} - c_j + \delta \mathbb{E}V_j^W(j, b_j^{1st})\Big) + (1 - q^{1st}(j))V_k^{P,2nd} & \text{if } j = k \\ q^{1st}(j)\delta \mathbb{E}V_k^L(j, b_j^{1st}) + (1 - q^{1st}(j))V_k^{P,2nd} & \text{if } j \neq k \end{cases} \tag{A7}$$

Similar to Equation (A5), I solve for the optimal fee at the 1st stage b_j^{1st} with FOC of Equation (A7) when $j = k$:

$$b_j^{1st} = c_j - \delta\Big(\mathbb{E}V_j^W(j, b_j^{1st}) - V_k^{P,2nd}\Big) + \left(\frac{1 + \delta\frac{\partial \mathbb{E}V_j^W(j,b_j^{1st})}{\partial b_j^{1st}}}{\alpha(1 - q^{1st}(j))}\right) \tag{A8}$$

Lastly, when solving employer's dynamic binary choice problems, value function at the 1st stage of RFP $\mathbb{E}\mathcal{U}^{1st}(j)$ replace the value functions in Equation (5).

F.2 Recover Provider Fixed Effects from CCP

I only estimate one provider quality net of cost $\xi_{other} = \nu_{other} - \alpha c_{other}$ for providers in the other category, and use conditional choice probabilities to recover the rest ξ_j.

First, I clarify why I estimate ξ instead of ν and c separately. To solve providers' problem, keeping track of markup $mk_j = b_j - c_j$ is sufficient. Plugging optimal fee, employer's utility in Equation (2) depends on the quality net of cost subtracted by markup. The employer's problem depends on ξ_j rather than ν_j and c_j separately.

$$\underbrace{\nu_j - \alpha c_j}_{\xi_j} - \alpha m k_j$$

Next, I only need to solve for ξ_{other} for providers in the other category. I use choice probabilities q_j^{2nd} in Equation (A2) at the second stage of RFP to recover the rest of ξ_j. I observe choice probabilities conditional on switching in the data, which is close but not exactly q_j^{2nd}, because second stage choice shares also include some non-switch choices.

A simple adjustment step can bridge the gap. Let $s_j(k)$ denote probabilities of choosing provider j conditional on switching from previous provider k. I have $J(J-1)$ of the following equations, with only J unknown q_j^{2nd} for $j \in \{1, ...J\}$. A minimum distance estimation allows me to recover q_j^{2nd}.

$$s_j(k) = \frac{q_j^{2nd}}{1 - q_k^{2nd}}$$

Then, I can use RFP choice probabilities at the second stage to recover $\xi_j = \nu_j - \alpha c_j$. Employer's expected utility from second stage RFP Equation (A3) can be rewritten as the following using the probabilities of choosing any provider in the other category (Arcidiacono and Miller, 2011). I use "Other" with the capitalized initial letter to indicate any provider in the other category with variety effect, where $\xi_{Other} > \xi_{other}$ due to $\nu_{Other} > \nu_{other}$. Cost, markup, and continuation value are the same since I

assume all providers in the other category are homogeneous.

$$\ln\Big(\sum_k \exp(\mathcal{U}_k^{2nd})\Big) = \underbrace{\xi_{Other} - \alpha m k_{other}^{2nd} + \delta \mathbb{E}V^E(other, b_{other}^{2nd})}_{\mathcal{U}_{Other}^{2nd}} - \ln(q_{Other}^{2nd}) \tag{A9}$$

I normalize $\xi_{Other} = 0$. mk_{other}^{2nd} and $V^E(other, b_{other}^{2nd})$ are both endogenized in the model equilibrium. Hence observed q_{Other}^{2nd} is sufficient to pin down the expected utility from second stage RFP for the employer, which I need to solve the rest of the model.

Next, compare the difference in the log of second stage RFP choice probabilities:

$$\begin{aligned}\ln(q_j^{2nd}) - \ln(q_{Other}^{2nd}) &= \mathcal{U}_j^{2nd} - \mathcal{U}_{Other}^{2nd} \\ &= \Big(\xi_j - \alpha m k_j^{2nd} + \delta \mathbb{E}V^E(j, b_j^{2nd})\Big) - \Big(-\alpha m k_{other}^{2nd} + \delta \mathbb{E}V^E(other, b_{other}^{2nd})\Big)\end{aligned}$$

Then, I can recover ξ_j for all providers except those in the other category, which I estimate in the model.

Second stage choice probability $\ln(q_{other}^{2nd})$ is not observed. So I use the probability of switching away from a provider in the other category to identify ξ_{other}, as discussed in Equation (G2).

F.3 State Variables

As discussed in Section 5, my model has two state variables: the identity of the current provider s and fee f. In my implementation, I separate fee f into optimal fee set at RFP b and an excess amount $\tilde{f}$ due to higher markup due to outdated menus.

$$f = b + \tilde{f}$$

Provider s sets different fees at RFP depending on whether s wins the RFP from the first or second stage. I introduce a third state variable $h \in \{0, 1\}$ corresponding to the stage of RFP. Then conditioning on (s, f) is the same as conditioning on $(s, h, \tilde{f})$.

The benefit of conditioning on $(s, h, \tilde{f})$ is computational. Since $\tilde{f} \geq 0$, it is easier to project $\tilde{f}$ on a grid. As a comparison, due to dynamic incentives, markup f can be negative. I would need to have a much larger grid. In practice, I discretize $\tilde{f}$ over an evenly spaced grid with 121 points from 0 to 60 basis points.

F.4 How Optimal Fees Affect Provider Continuation Values

I discuss how I solve for $\frac{\partial \mathbb{E}V^W(j, b_j(s))}{\partial b_j}$ in optimal fee Equations (A5) and (A8). First, as explained in Appendix F.3, I solve for $V^W(j, h, 0)$ in practice where $\tilde{f} = 0$ at RFP by

definition. I can rewrite the derivative of provider continuation values as

$$\frac{\partial \mathbb{E}V^W(j, b_j(s))}{\partial b_j} = \frac{\partial \mathbb{E}V^W(j, h, \tilde{f} = 0)}{\partial \tilde{f}}$$

Let Δ denote the grid size for the grid of $\tilde{f}$. I can then approximate the derivative using the finite difference.

$$\frac{\partial \mathbb{E}V^W(j, h, 0)}{\partial \tilde{f}} \approx \frac{\mathbb{E}V^W(j, h, \Delta) - \mathbb{E}V^W(j, h, 0)}{\Delta}$$

where $V^W(j, h, 0)$ and $V^W(j, h, \Delta)$ are provider continuation values evaluated at $\tilde{f} = \{0, \Delta\}$, which come out of value function iterations.

F.5 Solving for Equilibrium

Conditional on structural parameters, I follow a nested algorithm to solve for the model equilibrium. In the outer loop, I solve for equilibrium markup at 1st and 2nd stage RFP, and provider value function at the 2nd stage of the RFP. This requires solving for $3J$ values, where J is the number of providers. The dimension is relatively low, and I use the MatLab function `fsolve`. In the inner loop, given markup and provider value function at RFP, I solve for employer value functions $V^E(s, h, f)$ using value function iterations.

I rewrite Equations (A5) and (A8) as follows. Note that provider decisions only depend on markup.

$$mk_j^{2nd} = b_j^{2nd} - c_j = -\delta\left(\mathbb{E}V_j^W(j, 2, b_j^{2nd}) - \frac{\sum_{k \neq j} q_k \mathbb{E}V_j^L(k, 2, b_k^{2nd})}{1 - q_j^{2nd}}\right) + \frac{1 + \delta \frac{\partial \mathbb{E}V_j^W(j,2,b_j^{2nd})}{\partial b_j^{2nd}}}{\alpha(1 - q_j^{2nd})}$$

$$mk_j^{1st} = b_j^{1st} - c_j = -\delta\left(\mathbb{E}V_j^W(j, 1, b_j^{1st}) - V_k^{P,2nd}\right) + \frac{1 + \delta \frac{\partial \mathbb{E}V_j^W(j,1,b_j^{1st})}{\partial b_j^{1st}}}{\alpha(1 - q^{1st}(j))}$$

The nested algorithm is as follows

1. Start with an initial guess for $mk_j^{1st,0}, mk_j^{2nd,0}, V_j^{P,2nd,0}$
2. Inner loop value function iterations.
   (a) Start with an initial guess for employer value function $V^{E,0}(s, h, f)$
   (b) Solve for employer's expected utility at second stage RFP $\mathbb{E}\mathcal{U}^{2nd}$ using Equation (A9)
   (c) Solve for first stage RFP provider choice probabilities q_j^{1st} and employer's expected utility $\mathbb{E}\mathcal{U}^{1st}$
   (d) Solve for the probability that the employer runs RFP $\lambda(s, h, f)$ and update employer value function $V^{E,1}(s, h, f)$.

(e) Iteration step (a)-(d) until

$$||V^{E,t}(s,h,f) - V^{E,t+1}(s,h,f)||_\infty < \zeta$$

3. Solve recursively for provider value functions $V_s^W(s,h,f)$ and $V_j^L(s,h,f)$
4. Update provider value function at second stage of RFP $V_j^{P,2nd,1}$ using Equation (A4)
5. Using the finite difference approximation for derivative in markup $\frac{\partial V^W(j,s,b_j)}{\partial b_j}$ discussed in Appendix F.4
6. Update markup $mk_j^{2nd,1}$ and $mk_j^{1st,1}$
7. Iterate step 1 to 7 until

$$\max\left\{||mk_j^{2nd,1} - mk_j^{2nd,0}||_\infty,\ ||mk_j^{1st,1} - mk_j^{1st,0}||_\infty,\ ||V_j^{P,2nd,1} - V_j^{P,2nd,0}||_\infty\right\} < \zeta$$

F.6 Random Coefficient

I use Monte Carlo to numerically integrate over employers with different fee sensitivity parameters α_i with 50 Sobol draws. I have tried 100 Sobol draws and obtained quantitatively similar results. I have also tried evaluating integration using quadrature and obtained similar results. While quadrature method is more computationally efficient when evaluating integration, Monte Carlo simulation generates outcomes of employers at variance percentiles of fee sensitivity α_i.

G Auxiliary Models

The indirect inference estimation minimizes the distance between the vector of auxiliary model coefficients $\mathcal{B}^{data}$ estimated using observed data, and $\mathcal{B}^{model}(\Theta)$ estimated using data generated by my model at structural parameters Θ. When computing $\mathcal{B}^{model}(\Theta)$, I use one observation per state (s, f), weighted according to the stationary distribution across states. To compute the stationary distribution, I start from a given distribution and iteratively apply the transition matrix across all states (s, f) until convergence. Transition probabilities between states are based on optimal decisions of the employer and providers at the model equilibrium.

The distance between $\mathcal{B}^{model}(\Theta)$ and $\mathcal{B}^{data}$ is the sum of squared differences between the two vectors of coefficients, where coefficients are weighted equally. I have seven auxiliary models for seven parameters, so the weighting matrix should not matter and I use the identity matrix for simplicity.

I estimate the following seven structural parameters: RFP cost κ_{rfp}, switching cost κ_{sw}, the average α and standard deviation σ of fee sensitivity, inattention in RFP decision ρ, providers' menu update probability ϕ, and lastly the quality net of cost for providers in the other category ξ_{other}.

$$\Theta = \{\kappa_{rfp}, \kappa_{sw}, \alpha, \sigma, \rho, \phi, \xi_{other}\}$$

I use the following seven auxiliary models to identify these structural parameters. Let $(y_{it}, \tilde{f}_{ijt}, t_{ijt})$ collect observed provider choice y_{it}, residualized plan fee $\tilde{f}_{ijt}$, and menu turnover t_{ijt} for each employer i in year t with provider $y_{it} = j$.

G.1 Probability of Switching Providers: $\kappa_{rfp} + \kappa_{sw}$

The probability that employers switch providers identifies the sum of two transaction costs $\kappa_{rfp} + \kappa_{sw}$. Employers are less likely to switch providers if either RFP cost or switching cost is high. I do not observe RFPs and use Equation (G3) to separate κ_{rfp} and κ_{sw}. I also impose a condition that the incumbent provider prior to switching is not one from the other category, and later use the probability of switching away from a provider from the other category to recover ξ_{other}.

$$1\{y_{it} \neq y_{i,t-1}\} = \beta_1 + \epsilon_{ijt} \quad \text{if } y_{i,t-1} \neq other \tag{G1}$$

G.2 Probability of Switching From Smaller Providers: ξ_{other}

In Equation (G2), I estimate the probability that employers switch away from a provider in the other category. Compared with β_1 above, this probability identifies quality net of costs for providers in the other category: $\xi_{other} = \nu_{other} - \alpha c_{other}$. Providers in the

other category have a lower net quality if $\beta_2 > \beta_1$.

$$1\{y_{it} \neq y_{i,t-1}\} = \beta_2 + \epsilon_{ijt} \quad \text{if } y_{i,t-1} = other \tag{G2}$$

I do not separate quality ν from cost c, since only their difference matters after substituting the expression of optimal fees Equation (10) into employers' utility in Equation (2). I recover ξ_j for large providers (e.g. Fidelity and Vanguard) offline from conditional choice probabilities when employers switch providers, discussed in Section 6.4.

I cannot use the conditional probabilities of choosing any small providers in the data because employers may invite multiple providers from the other category to participate in their RFPs. Thus, I use Equation (G2) to recover ξ_{other} and leave the conditional probabilities of choosing small providers as a free moment to account for the variety from multiple small providers.[37]

G.3 Fee Autocorrelation Separates κ_{rfp} vs. κ_{sw}

Both RFP and switching costs affect the probability of switching providers in Equation (G1). To separate these two transaction costs, I use the autocorrelation of fees to infer the probability of RFPs. In my model, providers do not condition on previous fees when they set fees at RFPs. Under this assumption, the autocorrelation of fees at RFPs should be zero, which is the lower bound. Then, I estimate an upper bound of the autocorrelation when there is no menu turnover, which I can observe in the data. The upper bound is less than one in the data due to changes in expense ratios, reallocation across different funds, and measurement errors in recordkeeping fees.

When employers do not switch providers, an autocorrelation close to zero suggests frequent RFPs, low RFP costs, and high switching costs. Alternatively, an autocorrelation close to the upper bound suggests infrequent RFPs, high RFP costs, and low switching costs.

In my auxiliary model, I regress residualized fees on lag residualized fees interacted with an indicator for any menu update. I condition on observations without provider switches. The ratio of β_{3t}/β_{3n} corresponds to the relative frequency of RFPs.

$$\tilde{f}_{ijt} = \beta_{3t}1\{t_{ijt} > 0\}\tilde{f}_{ij,t-1} + \beta_{3n}1\{t_{ijt} = 0\}\tilde{f}_{ij,t-1} + \gamma 1\{t_{ijt} = 0\} + \Gamma_i + \epsilon_{ijt} \quad \text{if } y_{it} = y_{i,t-1} \tag{G3}$$

In Table A4, I report estimates of autocorrelation in residualized fees. In the first column, I show that the autocorrelation in residualized fees when employers switch providers is statistically indistinguishable from zero, consistent with the assumption in my model. In the second column, the upper bound of fee autocorrelation when employers do not switch providers and have no menu updates is 0.42. The ratio of

[37] $\xi_{other} = \nu_{other} - \alpha c_{other}$ depends on employers' fee sensitivities, which vary across employers. I assume ν_{other} and c_{other} also vary across employers such that ξ_{other} is constant across employers.

the autocorrelation with menu updates to the autocorrelation without menu updates is 0.76, which is the empirical counterpart to the autocorrelation in the fee state variable of my structural model.

Table A4: Fee Autocorrelation

VARIABLES	(1) Full sample	(2) Not switch providers
Switch providers	-0.04***	
	(0.00)	
Lag residualized fee × switch providers	0.02	
	(0.02)	
Lag residualized fee × not switch providers	0.33***	
	(0.01)	
Lag residualized fee × menu update		0.32***
		(0.01)
Lag residualized fee × no menu update		0.42***
		(0.01)
Menu update		-0.01***
		(0.00)
Observations	95,538	91,550
Employer FE	X	X

Table shows regression of residualized fees on lag residualized fees, by whether employers switch providers, not switch providers and have menu updates. I control for employer fixed effects and cluster standard errors at the employer level.

G.4 Probability that Providers Do Not Update Menus: ϕ

In Equation (G4), I estimate the probability that providers do not update menus, which identifies the constant and exogenous probability that providers update menus ϕ outside RFPs.

$$1\{t_{ijt} = 0\} = \beta_4 + \epsilon_{ijt} \tag{G4}$$

In the model, I assume providers always update menus at RFPs. So β_4 corresponds to $(1-\phi) \times (1-\lambda)$ rather than $1-\phi$, and needs to be estimated with other structural parameters.

G.5 Sensitivity of Switching With Respect to Fees: $\alpha \times \rho$

In Equation (G5), I estimate the sensitivity of switching providers with respect to lag residualized fees, after absorbing employer fixed effects. The coefficient β_5 corresponds to a mixture of employer fee sensitivity α and inattention ρ when they decide whether to conduct RFPs. I separate these two parameters using Equation (G6).

$$1\{y_{it} \neq y_{i,t-1}\} = \beta_5 \times \tilde{f}_{ij,t-1} + \Gamma_i + \epsilon_{ijt} \tag{G5}$$

This auxiliary model highlights the importance of controlling for employer fixed effects Γ_i. Some employers may have higher $\tilde{f}_{ij,t-1}$ potentially because of higher Γ_i, if they offer more services in their 401(k) plans. When considering whether to run RFPs, employers focus on fees relative to their production costs, or $\tilde{f}_{ij,t-1} - \Gamma_i$. When estimating Equation (G6), if we only use $\tilde{f}_{ij,t-1}$, we would underestimate β_5 due to an attenuation bias.[38] Table A5 reports β_5 with and without employer fixed effects for the full sample. After controlling for employers fixed effects, β_5 becomes three times larger.

Table A5: Sensitivity of Switching With Respect to Fees

	(1)	(2)
VARIABLES		
Lag residualized fee	0.03***	0.09***
	(0.00)	(0.01)
Observations	95,538	95,538
Employer FE		X

Regression coefficient of switching on residualized fees.

G.6 Fee Reductions When Employers Switch Providers: α

I use the average fee reductions when employers switch providers to separate the average fee sensitivity α from inattention ρ. Fee reductions when employers switch providers include two components. The first component is the accumulated fee growth relative to the trend since the previous RFP. After recovering the probability of RFP using Equation (G3) and measuring fee growth g_j in the data, this component is known. The second component captures higher markups charged by incumbent providers at RFPs. Comparing incremental markups with the probability that employers choose their incumbents at RFPs (which is known after Equation (G3)) identifies the average fee sensitivity α. Suppose α_i is high on average. Even though employers face switching costs, their incumbents cannot substantially raise markups, and fee reductions from switching providers are thus lower.

$$\Delta \tilde{f}_{ijt} = \beta_6 + \epsilon_{ijt} \quad \text{if } y_{it} \neq y_{i,t-1} \tag{G6}$$

To provide additional evidence, I examine how fee reductions from switching providers vary across employer characteristics. I regress fee changes on whether employers switch providers, interacted with a set of employer covariates.

$$\Delta \tilde{f}_{ijt} = \beta^0 + \beta^{sw} Switch_{it} + X'_{it}\beta + Switch_{it} \times X'_{it}\beta^{sw,x} + \epsilon_{ijt}$$

[38] One caveat is that this method cannot address time-varying unobserved production costs, so I potentially still underestimate employer sensitivity of switching providers with respect to fees.

All covariates are standardized. Table A6 reports coefficient $\beta^{sw,x}$. Employers reduce fees on average when they switch. Less negative $\Delta\tilde{f}_{ijt}$ (positive $\beta^{sw,x}$) suggests employers are more fee sensitive. Employer covariates include average total plan assets, which I also use to group employers into size quartiles, and the average number of plan participants. Larger employers should be more sophisticated and have higher fee sensitivities. I indeed estimate positive $\beta^{sw,x}$ in columns (1) and (2). In addition, I include

Table A6: Fee Reduction When Switching Providers

	(1)	(2)	(3)
VARIABLES			
Provider switch	-9.67***	-9.78***	-5.33***
	(0.24)	(0.25)	(0.96)
× plan asset	4.70***		
	(0.27)		
× num participants		2.22***	
		(0.26)	
× employer contribution		1.27***	
		(0.26)	
× participation rate		1.14***	
		(0.29)	
× ESG score			5.11***
			(0.91)
Observations	95,538	88,291	3,622

Regression coefficient of changes in fees on provider switches interacted with a set of employer characteristics

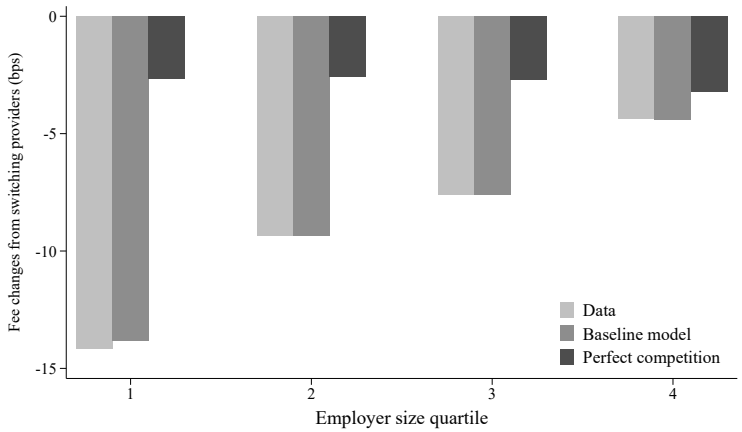
the fraction of employer contribution and the rate of employee participation in 401(k) plans. Employers can choose to match all or a fraction of employees' contributions to their 401(k) accounts. A higher employer contribution suggests employers value their employees' retirement benefits more. Not all employees participate in 401(k) plans, and a higher participation rate suggests that the employers put more effort into educating their employees about saving for retirement. In both cases, these employers should be more fee sensitive, and I estimate positive $\beta^{sw,x}$ on these covariates in column (2).

Lastly, I look at employers' ESG scores from S&P Global. ESG scores are only available in the later part of the sample, and I use the average score for each employer for all periods. Employers with higher ESG scores also have positive $\beta^{sw,x}$, suggesting they are more fee-sensitive. It is reasonable that employers who value environmental, social and governance objectives also prioritize their employees' benefits in their 401(k) plans.

As another way to see how β_6 captures fee sensitivities, I estimate a version of my model where I set α_i to be a very high value such that providers essentially engage in perfect competition. In Equation (10), as α_i approaches infinity, providers set fees at RFPs equal to costs offset by future gains per period, and the average markups would be zero. I estimate structural parameters $(\kappa_{sc}, \kappa_{rfpc}, \rho, \phi, \xi_{other})$ other than fee sensitivities

to match the rest of auxiliary model coefficients. Figure A6 shows that fee reductions from switching providers are much smaller under perfect competition.

Figure A6: Fee changes when switching providers



G.7 Variance of Fee Reductions From Switching: σ

Since the average fee reductions when employers switch providers identifies the average fee sensitivity α, the variance of fee reductions from switching should reflect the variance of fee sensitivity σ^2.

$$Var(\Delta \tilde{f}_{ijt}) = \beta_7 \quad \text{if } y_{it} \neq y_{i,t-1} \tag{G7}$$

However, the variance of fee reductions also includes the variance of changes in unobserved production costs as well as measurement errors in fees. To see this, I expand residualized fees $\tilde{f}_{ijt}$ into the state variable capturing markups f_{ijt}, an employer fixed effect for constant production costs Γ_i, and a time-varying component of production costs which can also include measurement errors in fees ϵ_{ijt}. Although fee changes difference out Γ_i, the variance of fee changes still captures $Var(\Delta\epsilon_{ijt})$, which I assume to be independent from markups.

$$\begin{aligned}
\tilde{f}_{ijt} &= f_{ijt} + \Gamma_i + \epsilon_{ijt} \\
\Delta\tilde{f}_{ijt} &= \Delta f_{ijt} + \Delta\epsilon_{ijt} \\
Var(\Delta\tilde{f}_{ijt}) &= Var(\Delta f_{ijt}) + Var(\Delta\epsilon_{ijt}) \quad \text{if } y_{it} \neq y_{i,t-1}
\end{aligned}$$

To make progress, I introduce another auxiliary model to account for $Var(\Delta\epsilon_{ijt})$. Note that markup changes when employers switch providers Δf_{ijt} are negative in my model, after controlling for provider quality and cost fixed effects. Any positive $\Delta\tilde{f}_{ijt}$ has to come from ϵ_{ijt}. If $Var(\Delta\epsilon_{ijt})$ is small, it is unlikely to see large positive changes in

residualized fees. Hence, I add the eighth auxiliary model to recover $Var(\Delta\epsilon_{ijt})$.

$$E[\Delta\tilde{f}_{ijt}|\Delta\tilde{f}_{ijt} > 0, Switch_{it}] = \beta_8 \tag{G8}$$

My structural model does not have $\Delta\epsilon$, so I simulate markups and analytically incorporate noise terms assuming $\Delta\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$ and that $\Delta\epsilon_{ijt}$ is independent from markups.

I expand β_8 as the following. I can use my structural model to simulate the distribution of markup changes conditional on switching $dF(\Delta f_{ijt}|Switch_{it})$, and evaluate the standard normal CDF Φ and PDF ϕ at $\frac{-\Delta f_{ijt}}{\sigma_\sigma}$ to compute the structural counterpart of β_8

$$\begin{aligned}
E[\Delta\tilde{f}_{ijt}|\Delta\tilde{f}_{ijt} > 0, Switch_{it}] &= \frac{E[(\Delta f_{ijt} + \Delta\epsilon_{ijt})1\{\Delta f_{ijt} + \Delta\epsilon_{ijt} > 0\}|Switch_{it}]}{E[\Delta f_{ijt} + \Delta\epsilon_{ijt} > 0|Switch_{it}]} \\
&= \frac{\int_{\Delta f} \Pr(\Delta\epsilon_{ijt} > -\Delta f_{ijt}) E[\Delta f_{ijt} + \Delta\epsilon_{ijt}|\Delta\epsilon_{ijt} > -\Delta f_{ijt}] dF(\Delta f_{ijt}|Switch_{it})}{\int_{\Delta f} \Pr(\Delta\epsilon_{ijt} > -\Delta f_{ijt}) dF(\Delta f_{ijt}|Switch_{it})} \\
&= \frac{\int_{\Delta f} \left(1 - \Phi(\frac{-\Delta f_{ijt}}{\sigma_\sigma})\right)\Delta f_{ijt} + \sigma_\epsilon\phi(\frac{-\Delta f_{ijt}}{\sigma_\sigma}) dF(\Delta f_{ijt}|Switch_{it})}{\int_{\Delta f} 1 - \Phi(\frac{-\Delta f_{ijt}}{\sigma_\sigma}) dF(\Delta f_{ijt}|Switch_{it})}
\end{aligned}$$

G.8 Alternative Assumption: Variance of Residualized Fees: σ

For robustness, I consider an alternative auxiliary model where I directly fit the variance of residualized fees, ruling out variation in unobserved production costs $Var(\Delta\tilde{f}_{ijt}) = Var(\Delta f_{ijt})$. Because I condition on employers size quartiles, this assumption is treating employers with similar size as having the same economies of scale. Any unobserved services or plan features do not lead to differences in costs.

$$Var(\tilde{f}_{ijt}) = \beta_{7b} \tag{G7b}$$

H Hedonic Regression

To residualize fees, I use a two-way fixed effect regression in Equation (11), which I repeat below

$$fee_{ijt} = \underbrace{X_{it}\beta}_{\substack{\text{Observed} \\ \text{characteristics}}} + \underbrace{\Gamma_j + \tau_{jt} + \tau_{s(i)t}}_{\substack{\text{Provider FE} \\ \text{and time trend}}} + \underbrace{\Gamma_i + \epsilon_{ijt}}_{\tilde{f}_{ijt}\text{: Residualized fee}} \tag{A10}$$

I use a balanced panel of employers so time trends do not reflect sample selections. I rely on employer switches to separate employer and provider fixed effects (Abowd et al., 1999). Due to limited number of observations in 2009 when many employers have not started reporting Form 5500, I include year 2010 to 2019.

First, X_{it} captures observed plan characteristics. I assume employers determine these characteristics as menu composition and types of services from the providers, and abstract away from how providers steer menu composition. Most providers have access to funds across all major asset classes and investment styles, and can be relatively indifferent to employers' requests. In Table A9, I show that provider fixed effects explain generally less than 20% of the variation in menu compositions, whereas provider and employer fixed effects together explain around 70%.

Specially, I include compositions of funds from broad asset classes, such as bonds and large, mid, and small cap equities. To further control for menu composition within broad asset classes, I also include the relative fee of the Morningstar category within each asset class to account for whether participants have more specific investment requests that come with higher expense ratios. To construct this relative fee measure, I regress the Morningstar category average fees on asset class fixed effects and obtain the residuals. Thees residuals measure whether funds in this Morningstar category are on average more expensive within the asset class. To aggregate at plan $\times$ year level, I compute the average of the relative fee of each fund.

I do not control for the fraction of index funds, because offering index funds with lower expense ratios can be driven by both markups and heterogeneous services. It is reasonable for employers to offer more active funds catering to participants' preferences, and higher expense ratios of active funds suggest higher production costs. Alternatively, employers may allow providers to charge high markups by designing plan menus with expensive active funds. Also, Table A9 shows that provider fixed effects explain a larger fraction of variation in fraction of index funds on plan menus, suggesting that employers main jointly choose providers and the fraction of index funds on their plan menus.

X_{it} also includes whether the main provider offers additional services such as actuary, advisory, insurance, and asset management. The data includes limited information about services. I only observe recordkeepers' service codes such as advisory, asset management, and insurance.

I also control for the across time variation of employer's size in X_{it} to reflect how changes in size affect fees. The employer fixed effects Γ_i absorb the variation in fees due to persistent differences in economies of scale or fee sensitivities across employers.

Second, I residualize time trends to control for the decline in production costs. Time trends are provider specific τ_{jt} to allow for different rates of decline in production costs, and also specific to each employer size quartile $\tau_{s(i)t}$ to allow for different rates of decline across different sizes. For example, Vanguard might have already reached low production cost at the beginning of the sample due to large asset under management in its mutual funds. So incremental reduction can be smaller compared to other providers. With a stationary model, I assume that the employer does not have a higher preference for Vanguard in the early part of my sample (higher $\nu_{Vanguard}$). To relax this assumption requires a non-stationary model with calendar time as another state variable. This extended version is much more computationally demanding, where additional economic insights can be limited. The stable market share across main providers shown in Figure 1b suggests that it is reasonable to assume employers have relatively constant preferences across providers.

Third, I residualize provider fixed effects Γ_j to control for persistent differences in quality or production costs across providers. However, in my structural model, providers with different ν_j and c_j can also charge different markups. This is not a major concern since I do not use my model to target any average fee moments. I tried including provider fixed effects in residualized fees in previous versions. I obtained similar estimates for structural parameters, but have to introduce two sets of fixed effects for provider quality ν_j and production costs c_j.

In Table A7 and Table A8, I report coefficients of regression estimates. I report two specifications, and use column (2) to residualize fees in my main analysis. All variables are standardized except for provider fixed effects and time trends. Fees tend to be higher if the employers offer more mid/small cap equity funds, international equity funds, and alternative funds. Fees are lower with more cash funds and large-cap equities. Within each asset class, fees are higher if employers offer Morningstar categories where funds typically have higher expense ratios.

Comparing across providers, Vanguard has the lowest fees, followed by Empower and Fidelity. Principal Financial Group has the highest fees. In column (1), I only include time trend without interaction, which suggests that fees decline by 2.7 bps each year on average. In column (2), I interact time trends with provider fixed effects and employer size quartile fixed effects. Fees decline by 1 to 4 bps per year across different providers. Relative to the largest providers in the 4th quartile, smaller providers see greater fee reductions.

The hedonic regressions have R-squared over 80%, after including employer fixed effects. But these employer fixed effects can capture both markups and production

costs. Hence, I only include employer fixed effects to help estimate provider fixed effects and I still keep them as part of the residualized fees.

Table A7: Two Way Fixed Effects Regression

VARIABLES	(1)	(2)
Frac international	40.65***	35.05***
	(2.68)	(2.66)
Frac alternative	52.84***	49.89***
	(3.85)	(3.81)
Frac cash	-43.22***	-40.95***
	(3.02)	(3.01)
Frac bond	3.14	4.24*
	(2.40)	(2.38)
Frac small cap equity	31.15***	34.01***
	(3.42)	(3.39)
Frac mid cap equity	22.95***	22.80***
	(2.78)	(2.75)
Frac large cap equity	-4.67**	-5.91***
	(1.93)	(1.91)
Frac target date fund	-5.81***	-6.63***
	(1.60)	(1.58)
Category fee within asset class	2.05***	2.05***
	(0.06)	(0.06)
Actuary service	0.78	1.42
	(3.43)	(3.38)
Advisory service	1.16***	0.17
	(0.38)	(0.37)
Insurance service	-3.57***	-3.02***
	(0.51)	(0.51)
Asset management service	-0.21	-0.11
	(0.34)	(0.33)
Ln number of participants	-1.81***	-1.67***
	(0.34)	(0.34)
Observations	30,535	30,535
R-squared	0.82	0.83
Employer FE	X	X

Two way fixed effects regression of total fees including investment and recordkeeping fees on asset category composition, number of participants, provider specific time trends, provider and employer fixed effects. Coefficients are other variables are reported in the second part of in Table A8 I use a balanced panel of plans from 2010-2019.

Table A8: Two Way Fixed Effects Regression (Continued)

VARIABLES	(1)	(2)
ADP	-0.89	-0.91
	(1.03)	(1.02)
Empower	-13.05***	-12.62***
	(0.80)	(0.79)
Fidelity	-6.71***	-6.49***
	(0.60)	(0.60)
Principal Fin	8.62***	7.47***
	(1.12)	(1.12)
Vanguard	-21.23***	-21.09***
	(1.06)	(1.04)
Year	-2.71***	
	(0.03)	
Year × Other		-2.48***
		(0.05)
Year × ADP		-2.90***
		(0.15)
Year × Empower		-2.36***
		(0.13)
Year × Fidelity		-2.69***
		(0.06)
Year × Principal Fin		-3.97***
		(0.19)
Year × Vanguard		-0.95***
		(0.08)
Year × employer 1st quartile		-0.90***
		(0.13)
Year × employer 2nd quartile		-0.64***
		(0.08)
Year × employer 3rd quartile		-0.69***
		(0.06)
Observations	30,535	30,535
R-squared	0.82	0.83
Employer FE	X	X

Two way fixed effects regression of total fees including investment and recordkeeping fees on asset category composition, number of participants, provider specific time trends, provider and employer fixed effects. Coefficients are other variables are reported in the first part of in Table A7 I use a balanced panel of plans from 2010-2019.

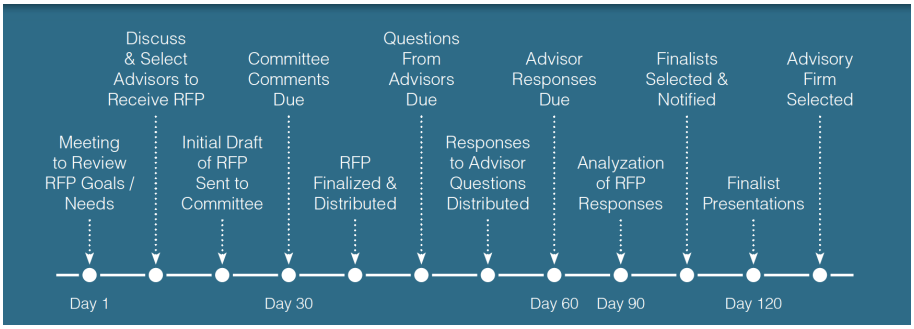
Table A9: Variation of Employer Characteristics Explained by Provider and Employer Fixed Effects

	Provider FE	Provider & employer FE
Frac large cap equity	0.076	0.814
Frac mid cap equity	0.081	0.796
Frac small cap equity	0.089	0.771
Frac bond	0.191	0.791
Frac cash	0.079	0.666
Frac international	0.254	0.830
Frac alternative	0.049	0.757
Frac target date fund	0.252	0.860
Morningstar category average fee within asset class	0.150	0.769
Fund fee within category	0.322	0.817
Frac index funds	0.321	0.835

Table reports R^2 of regressing employer $\times$ year level menu characteristics on provider fixed effects and both provider and employer fixed effects.

I Appendix Tables and Figures

Figure A7: Sample RFP Timeline



Source: SageView, "401(k) RFP Guide: A step-by-step resource to selecting the right advisor for your retirement program"

Table A10: Probability of RFPs and Offline Parameter Estimates

Employer quartiles	1st	2nd	3rd	4th
Prob of RFP	0.30	0.23	0.18	0.15
Conditional choice probability (%)				
ADP	0.08	0.08	0.02	0.02
Empower	0.17	0.17	0.14	0.14
Fidelity	0.17	0.17	0.23	0.23
Principal Fin	0.04	0.04	0.06	0.06
Vanguard	0.05	0.05	0.06	0.06
Other	0.49	0.49	0.48	0.48
Proudction cost decline / markups growth rates (bps)				
ADP	2.91	2.76	2.62	2.39
Empower	1.94	1.79	1.65	1.42
Fidelity	2.44	2.29	2.15	1.92
Principal Fin	3.93	3.78	3.64	3.41
Vanguard	1.16	1.01	0.87	0.64
Other	2.24	2.09	1.95	1.72

Table displays the average probability of running RFPs in the first row. The middle panel displays conditional choice probabilities at RFPs. The bottom panel displays production cost decline for each provider g_j measured using the data.

Table A11: Fits of Auxiliary Model Coefficients

Employer quartiles		1st		2nd		3rd		4th	
		Model	Data	Model	Data	Model	Data	Model	Data
Pr of switching from large providers	eq. (G1)	0.054	0.030	0.032	0.024	0.022	0.018	0.019	0.014
			(0.002)		(0.001)		(0.001)		(0.001)
Pr of switching from other providers	eq. (G2)	0.059	0.065	0.067	0.067	0.066	0.067	0.049	0.049
			(0.003)		(0.002)		(0.002)		(0.002)
Fee autocorrelation ratio	eq. (G3)	0.597	0.599	0.721	0.721	0.788	0.788	0.818	0.819
			(0.042)		(0.025)		(0.024)		(0.024)
Pr of no menu update	eq. (G4)	0.334	0.333	0.272	0.272	0.247	0.247	0.270	0.270
			(0.004)		(0.004)		(0.003)		(0.003)
Sensitivity of switching wrt fee	eq. (G5)	0.090	0.119	0.096	0.098	0.091	0.091	0.057	0.057
			(0.016)		(0.015)		(0.015)		(0.013)
Fee changes from switching	eq. (G6)	-0.138	-0.142	-0.093	-0.094	-0.076	-0.076	-0.044	-0.044
			(0.009)		(0.008)		(0.007)		(0.006)
Variance of fee changes from switching	eq. (G7)	0.087	0.079	0.065	0.066	0.045	0.048	0.032	0.034
			(0.003)		(0.002)		(0.002)		(0.002)
Positive fee changes from switching	eq. (G8)	0.174	0.179	0.164	0.163	0.137	0.136	0.120	0.118
			(0.008)		(0.006)		(0.006)		(0.006)

Table displays auxiliary model coefficients estimated using data generated by the structural model and coefficients estimated using observed data. Standard errors of coefficients estimated using observed data are reported in parenthesis. These standard errors are bootstrapped using 250 samples with replacement.

Table A12: Structural Parameter Estimates, No Unobserved Production Costs

Employer quartiles		1st		2nd		3rd		4th	
		Coef	SE	Coef	SE	Coef	SE	Coef	SE
RFP cost	$\kappa_{rfp}/\bar{\alpha}$								
bps		16.0	(0.1)	14.4	(0.1)	8.6	(0.1)	0.8	(0.0)
$ thousand		10		22		31		18	
Switching cost	$\kappa_{sw}/\bar{\alpha}$								
bps		68.6	(0.3)	51.6	(0.3)	25.5	(0.2)	1.9	(0.0)
$ thousand		44		80		91		42	
Fee sensitivity: mean	$\bar{\alpha}$								
Coefficient		0.07	(0.00)	0.11	(0.00)	0.22	(0.00)	4.92	(0.08)
Elasticity		1.0		1.4		2.9		62.2	
Fee sensitivity: st.dev.	σ_α								
Coefficient		0.09	(0.00)	0.14	(0.00)	0.29	(0.00)	14.35	(0.33)
Elasticity		1.0		1.4		2.8		53.5	
RFP attention	ρ								
Coefficient		0.94	(0.00)	0.95	(0.00)	0.69	(0.00)	0.26	(0.00)
Pr menu update	ϕ	0.50	(0.00)	0.62	(0.00)	0.68	(0.00)	0.63	(0.00)
Net quality of other providers	$\xi/\bar{\alpha}$	-20.8	(0.1)	-23.5	(0.2)	-16.1	(0.2)	-1.2	(0.0)
Relative to other providers									
ADP		-1.3		1.2		0.5		0.2	
Principal Fin.		-2.8		0.3		1.7		0.3	
Empower		1.4		3.1		3.7		0.4	
Fidelity		1.4		2.9		4.7		0.4	
Vanguard		-2.7		0.5		2.7		0.5	

Table reports parameter estimates when I directly fit variance of residualized fees. Standard errors are computed using the asymptotic variance formula, where the variance-covariance matrix of auxiliary model coefficients is computed using the bootstrap method over 250 iterations. In each iteration, I sample employers with replacement and estimate auxiliary model coefficients.

Table A13: Fits of Auxiliary Model Coefficients, No Unobserved Production Costs

Employer quartiles		1st		2nd		3rd		4th	
		Model	Data	Model	Data	Model	Data	Model	Data
Pr of switching from large providers	eq. (G1)	0.057	0.030	0.035	0.024	0.026	0.018	0.016	0.014
			(0.002)		(0.001)		(0.001)		(0.001)
Pr of switching from other providers	eq. (G2)	0.058	0.065	0.056	0.067	0.067	0.067	0.046	0.049
			(0.003)		(0.002)		(0.002)		(0.002)
Fee autocorrelation ratio	eq. (G3)	0.604	0.599	0.716	0.721	0.788	0.788	0.819	0.819
			(0.042)		(0.025)		(0.024)		(0.024)
Pr of no menu update	eq. (G4)	0.336	0.333	0.272	0.272	0.245	0.247	0.270	0.270
			(0.004)		(0.004)		(0.003)		(0.003)
Sensitivity of switching wrt fee	eq. (G5)	0.095	0.119	0.104	0.098	0.086	0.091	0.059	0.057
			(0.016)		(0.015)		(0.015)		(0.013)
Fee changes from switching	eq. (G6)	-0.143	-0.142	-0.091	-0.094	-0.076	-0.076	-0.044	-0.044
			(0.009)		(0.008)		(0.007)		(0.006)
Variance of fee	eq. (G7b)	0.087	0.088	0.053	0.053	0.037	0.037	0.025	0.025
			(0.002)		(0.001)		(0.001)		(0.001)

Table displays auxiliary model coefficients estimated using data generated by the structural model and coefficients estimated using observed data, when I directly fit variance of residualized fees. Standard errors of coefficients estimated using observed data are reported in parenthesis. These standard errors are bootstrapped using 250 samples with replacement.

Table A14: Structural Parameter Estimates, Myopic

Employer quartiles		1st		2nd		3rd		4th	
		Coef	SE	Coef	SE	Coef	SE	Coef	SE
RFP cost	$\kappa_{rfp}/\bar{\alpha}$								
bps		2.1	(0.0)	1.9	(0.0)	1.9	(0.2)	0.2	(0.0)
$ thousand		1		3		7		5	
Switching cost	$\kappa_{sw}/\bar{\alpha}$								
bps		13.4	(0.0)	10.1	(0.1)	8.4	(0.3)	1.4	(0.1)
$ thousand		9		16		30		31	
Fee sensitivity: mean	$\bar{\alpha}$								
Coefficient		0.57	(0.00)	0.94	(0.01)	1.21	(0.06)	8.43	(0.82)
Elasticity		8.4		12.2		13.7		60.0	
Fee sensitivity: st.dev.	σ_α								
Coefficient		0.74	(0.00)	1.23	(0.01)	1.59	(0.06)	11.05	(2.04)
Elasticity		6.5		9.8		10.7		62.9	
RFP attention	ρ								
Coefficient		0.71	(0.01)	0.50	(0.02)	0.49	(0.10)	0.01	(0.00)
Pr menu update	ϕ	0.49	(0.00)	0.60	(0.00)	0.66	(0.00)	0.68	(0.00)
Net quality of other providers	$\xi/\bar{\alpha}$	-2.8	(0.0)	-6.3	(0.0)	-6.1	(0.2)	-1.0	(0.1)
Relative to other providers									
ADP		-0.3		4.5		3.6		0.7	
Principal Fin.		-1.5		3.8		4.5		0.8	
Empower		1.1		5.4		5.3		0.9	
Fidelity		1.1		5.4		5.8		1.0	
Vanguard		-1.2		4.0		4.5		0.8	

Table reports parameter estimates when I assume providers and employers are myopic. Standard errors are computed using the asymptotic variance formula, where the variance-covariance matrix of auxiliary model coefficients is computed using the bootstrap method over 250 iterations. In each iteration, I sample employers with replacement and estimate auxiliary model coefficients.

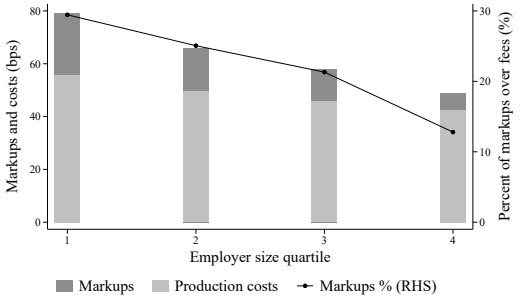
Table A15: Fits of Auxiliary Model Coefficients, Myopic

Employer quartiles		1st		2nd		3rd		4th	
		Model	Data	Model	Data	Model	Data	Model	Data
Pr of switching from large providers	eq. (G1)	0.044	0.030	0.025	0.024	0.018	0.018	0.014	0.014
			(0.002)		(0.001)		(0.001)		(0.001)
Pr of switching from other providers	eq. (G2)	0.048	0.065	0.068	0.067	0.067	0.067	0.049	0.049
			(0.003)		(0.002)		(0.002)		(0.002)
Fee autocorrelation ratio	eq. (G3)	0.599	0.599	0.721	0.721	0.787	0.788	0.819	0.819
			(0.042)		(0.025)		(0.024)		(0.024)
Pr of no menu update	eq. (G4)	0.333	0.333	0.272	0.272	0.246	0.247	0.270	0.270
			(0.004)		(0.004)		(0.003)		(0.003)
Sensitivity of switching wrt fee	eq. (G5)	0.121	0.119	0.097	0.098	0.092	0.091	0.057	0.057
			(0.016)		(0.015)		(0.015)		(0.013)
Fee changes from switching	eq. (G6)	-0.141	-0.142	-0.094	-0.094	-0.076	-0.076	-0.044	-0.044
			(0.009)		(0.008)		(0.007)		(0.006)
Variance of fee changes from switching	eq. (G7)	0.090	0.079	0.066	0.066	0.047	0.048	0.029	0.034
			(0.003)		(0.002)		(0.002)		(0.002)
Positive fee changes from switching	eq. (G8)	0.172	0.179	0.162	0.163	0.136	0.136	0.121	0.118
			(0.008)		(0.006)		(0.006)		(0.006)

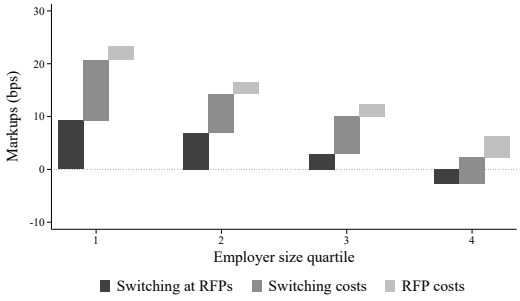
Table displays auxiliary model coefficients estimated using data generated by the structural model and coefficients estimated using observed data, where I assume providers and employers are myopic. Standard errors of coefficients estimated using observed data are reported in parenthesis. These standard errors are bootstrapped using 250 samples with replacement.

Figure A8: Model Generated Markups, No Unobserved Production Costs

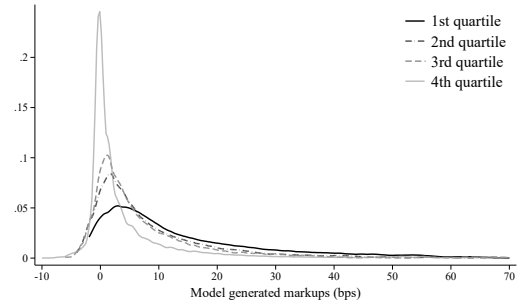
(a) Fee Breakdown



(b) Markups Breakdown



(c) Distribution of Markups



(d) Decomposition of fee dispersion

