The Fiscal Cost of Quantitative Easing*

Adrien d'Avernas† Antoine Hubert de Fraisse‡ Liming Ning§
Quentin Vandeweyer¶

First draft: October 2024 This draft: July 2026

Abstract

Quantitative easing (QE) shortens the duration of the consolidated public balance sheet, swapping long-term government bonds for short, floating-rate liabilities and shifting interest-rate risk onto taxpayers. In segmented bond markets this transfer can support real activity, but the state-contingent losses must be financed with distortionary taxes. We quantify the ex ante fiscal-efficiency cost by mapping forecasted QE-portfolio returns into expected tax deadweight losses. Across U.S. QE programs, the expected cost is 0.35% of GDP—below published output-effect estimates—with a conservative upper bound of 1.35%. Yet the distribution is wide: across programs, 95th-percentile losses sum to 6.65% of GDP.

Keywords: Large-Scale Asset Purchase Programs, Central Bank Losses, Rollover Risk, Interest Rate Risk, Optimal Maturity of Government Debt

JEL Classifications: E5, E58, E6, E63, G10, G12

---

*We would like to thank Max Croce, Doug Diamond, William Diamond, Valentin Haddad, Sam Hanson, Ralph Koijen, Moritz Lenel, Thomas Mertens, Stefan Nagel, Luboš Pastor, Raghu Rajan, Walker Ray, Dmitriy Sergeyev, and Paolo Varraso for valuable comments, as well as participants in seminars and conferences at the University of Chicago Booth School of Business, the Stockholm School of Economics, the Wharton School, HEC Paris, the European Central Bank, the Sveriges Riksbank, the Federal Reserve Bank of New York, the Federal Reserve Bank of Chicago, London Business School, MIT Sloan, Sciences Po, the University of Lausanne, USC, SITE, the SaMMF Seminar, the SFS Cavalcade 2025, the Western Finance Association 2025, FMA Europe 2025, the 2025 Adam Smith Workshop, the CESifo Conference on Macro, Money, and International Finance, the Gerzensee Evening Sessions, the Sovereign Bond Market Conference (Venice), the Oxford SAID–VU SBE Macro-Finance Conference, and the USC Macro-Finance Conference.

†Stockholm School of Economics, adrien.davernas@hhs.se

‡London School of Economics & National Bureau of Economic Research

§Northwestern University Kellogg School of Management

¶University of Chicago Booth School of Business

Conflict-of-Interest Disclosure Statement

Adrien d'Avernas: I have nothing to disclose.

Antoine Hubert de Fraisse: I have nothing to disclose.

Liming Ning: I have nothing to disclose.

Quentin Vandeweyer: I have nothing to disclose.

Since the Great Financial Crisis, central banks have implemented large-scale asset purchase programs known as quantitative easing (QE). In its canonical form, QE purchases long-maturity government securities and finances them by issuing short-maturity interest-bearing liabilities (reserve balances in modern operating frameworks). This maturity transformation exposes the public sector to interest-rate (duration) risk: when short-term interest rates rise, the interest expense on floating-rate liabilities increases quickly, while the cash flows on long-term assets remain largely fixed. Under standard no-arbitrage pricing, the present value of the public sector's future net-interest income varies closely with the market value of this long-duration position.

This exposure became salient during the rapid monetary tightening from 2021 to 2023. Using the Federal Reserve's System Open Market Account (SOMA) holdings and observed yield changes, we estimate that the market value of the Fed's QE portfolio declined by roughly 4 percent of U.S. GDP during 2022.[1] A natural question follows: when QE generates large balance-sheet losses in some states of the world, how should those losses enter an economic evaluation of QE, and how should they be weighed against QE's macroeconomic benefits?

A key premise of this paper is that balance-sheet losses become economically meaningful when they translate into distortionary fiscal adjustments. With fiscal backing of the central bank,[2] central bank losses have real fiscal incidence: they reduce remittances to the Treasury and, in present value, must be offset by future primary surpluses (higher taxes or lower spending) to satisfy the consolidated public sector's intertemporal budget constraint (Jiang et al., 2026). When financed with distortionary taxation, the welfare-relevant object is therefore not the ex post accounting loss per se, but the ex ante expected deadweight losses generated by financing QE-induced gains and losses across states. This perspective links the recent debate on central bank operating losses to a classic public-finance question: how costly is it for the government to shorten the maturity of its liabilities and thereby increase refinancing risk?

[1] Authors' calculations based on SOMA holdings and Treasury yield changes. Begenau, Piazzesi, and Schneider (2025) document how interest-rate risk exposures in U.S. banks' fixed-income positions similarly materialized during this episode.

[2] By fiscal backing we mean that the Treasury covers any shortfall in the central bank's net income, so that central-bank solvency is never in question and losses map into the consolidated budget constraint. This is the standard assumption in the literature on central-bank balance-sheet risk (Hall and Reis, 2015; Del Negro and Sims, 2015; Reis, 2015); absent such backing, large losses could instead force a loss of monetary control.

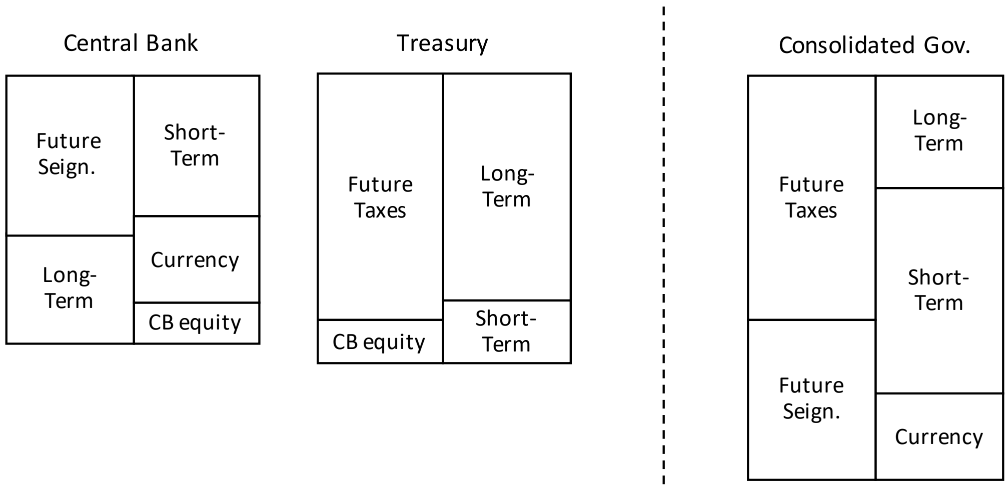


Figure 1: Sketch of Consolidated Government Balance Sheets. The figure sketches the balance sheet of the government in our model, both unconsolidated (left side) and consolidated (right side). The Treasury issues both long-term and short-term debt against agents' future tax liabilities and central bank equity. The central bank issues currency and short-term liabilities, holds long-term government debt, and generates seigniorage revenues from currency.

We study these fiscal-efficiency implications of QE through a unified risk-benefit lens that interprets QE as a maturity choice of the consolidated government. When consolidating the Treasury and the central bank, intra-governmental positions—such as Treasury securities held by the central bank—net out (see Figure 1). QE then corresponds to a maturity-shortening operation of the consolidated public sector: long-term claims held by the private sector are effectively replaced with short-maturity interest-bearing liabilities.[3] The consolidated perspective highlights the dual macro-fiscal implications of QE: by reallocating duration risk, QE can stimulate aggregate demand when constrained at the zero lower bound (ZLB); however, it simultaneously increases refinancing risks for the government, necessitating state-contingent distortionary fiscal adjustments.

To formalize this trade-off, we develop a tractable model with nominal rigidities, a ZLB constraint, and distortionary taxation, featuring two types of households: (i) optimizing bondholders who trade long-term government bonds, and (ii) hand-to-mouth households

[3]An equivalent maturity shortening could be implemented by the Treasury by increasing the share of bills relative to bonds—an operation sometimes described as "stealth QE" or "activist Treasury issuance" (Roubini and Miran, 2024).

who do not participate in financial markets. In this environment, QE operates through an interest-rate-risk-extraction channel akin to preferred-habitat and portfolio-balance mechanisms (Ray, 2019; Caballero and Simsek, 2020; Vayanos and Vila, 2021; Caballero and Simsek, 2021; Kekre, Lenel, and Mainardi, 2024). By reducing the duration risk borne by bondholders, QE can reduce their precautionary saving motive and stimulate aggregate demand when monetary policy is constrained at the ZLB. Under consolidated government accounting, the aggregate risk does not disappear when it moves onto the central bank's balance sheet: it is ultimately shifted, through future tax obligations, onto non-participating households, who do not offset this reallocation by assumption.

The same maturity shortening, however, increases fiscal refinancing risk (Greenwood, Hanson, and Stein, 2015b). By shortening the consolidated maturity structure, QE requires the government to refinance a larger share of its liabilities at uncertain future short rates. This greater exposure increases the dispersion of the government's financing needs and, with fiscal backing, the dispersion of distortionary tax rates across states. With deadweight losses convex in the tax rate, the efficiency gains in lower-tax states are smaller than the efficiency losses in higher-tax states. As a result, expanding QE increases the expected efficiency cost of taxation, even when the government's expected financing need is unchanged. We characterize the optimal scale of QE at the ZLB as an interior condition that equates the marginal output benefit from additional risk extraction to the marginal increase in expected deadweight losses induced by additional refinancing risk.

To take this fiscal-efficiency cost to the data, we introduce an implementable estimand: the *ex ante* expected change in the present value of tax deadweight losses induced by a QE program under a conservative terminal financing rule. Operationally, we isolate the incremental maturity transformation created by QE while holding fixed the primary fiscal stance and the Treasury's baseline debt-management policy. The estimand is thus a partial, ceteris-paribus object: it prices the financing consequences of QE's duration exposure while also holding fixed the discount factor, output, and the term-structure pricing environment (invariant to QE under no price impact), and it does not require taking a stand on QE's general-equilibrium transmission. Under the terminal financing rule, any realized gain or loss on the incremental QE portfolio is financed by a one-time adjustment in distortionary taxes at the unwind date, with no intertemporal smoothing. This assumption is conservative: in practice, fiscal adjustments and central-bank remittances are typically smoothed over time (including through deferred-transfer accounting), which attenuates deadweight losses for a given realized loss.[4] Our cost metric should therefore be interpreted as an upper bound for the fiscal-efficiency consequences of QE-induced duration exposure.

A central implication of the exercise is that QE does not generate systematic fiscal gains from "carry" when evaluated under the pricing kernel that values Treasury cash flows: the incremental QE portfolio is constructed to have zero net worth at inception, and under no-arbitrage, a risk-neutral expectation of zero. Expected fiscal-efficiency costs arise instead from second moments: (i) the volatility of QE-induced tax adjustments, and (ii) the interaction between those adjustments and the fiscal state (losses realized in high-tax states are disproportionately costly). We return to this choice of pricing kernel, and to the expected carry that QE earns under the physical measure, in Section II.

We quantify these fiscal-efficiency costs for the five major U.S. maturity transformation episodes since 2008 (QE1–QE4 and the Maturity Extension Program) using granular SOMA holdings data and an arbitrage-free affine term structure model fitted to observed yield curves. In its simulation design, our exercise is closest to Christensen, Lopez, and Rudebusch (2015), who pioneered probability-based stress tests of the Federal Reserve's balance sheet, income, and remittances using term-structure Monte Carlo methods; we take the further step of mapping the full *ex ante* distribution of QE portfolio outcomes into the expected present value of tax deadweight losses, evaluated under the no-arbitrage pricing kernel rather than as a cash-flow projection. We provide two complementary measures of the fiscal efficiency cost that share the same risk-neutral valuation benchmark but differ in how much structure they impose on long-horizon comovements. The first is an exact $\mathbb{Q}$-measure estimate evaluated under the stochastic discount factor implied by this model. The second is a conservative Cauchy–Schwarz upper bound that avoids pinning down the state-by-state long-horizon comovement between pricing weights and fiscal outcomes beyond an envelope restriction, and therefore depends only on $\mathbb{Q}$-measure second moments.

Under our baseline horizon, we obtain a cumulative risk-neutral expected cost of 0.35% of GDP and an upper bound of 1.35% of GDP. As an external benchmark for economic magnitudes, the cumulative output effects reported in the empirical QE literature (from 1.17% to 4.22% of GDP in Fabo et al., 2021) exceed our expected fiscal-efficiency cost

[4] See, for example, the "deferred asset" discussion in Faria-e-Castro and Jordan-Wood (2023).

of 0.35% of GDP. These qualitative conclusions are unchanged under alternative holding horizons, term-structure parameterizations, and financing rules that spread taxes over time.

Three features discipline these magnitudes. First, the valuation is anchored to market prices: the pricing kernel is implied by the observed Treasury yield curve under no-arbitrage rather than by preference parameters. Second, the ex ante cost is exactly linear in the tax-convexity parameter α, the one input taken from outside asset markets, so readers can rescale the headline estimates one-for-one under any preferred calibration. For the pricing-kernel cost to reach even the most conservative output benchmark in the empirical literature, α would need to exceed 7—more than three times our baseline of 2.17 and, expressed in terms of the underlying elasticity of taxable income, above the entire range of estimates that Saez, Slemrod, and Giertz (2012) consider most reliable. Third, the remaining implementation choices—the one-shot terminal financing rule and the worst-case comovement construction behind the upper bound—are selected to push the cost estimate up, and the discounting of output benefits makes the cost–benefit comparison conservative in the same direction.

At the same time, large marked-to-market QE portfolio losses—such as those observed during 2022—underscore that the distribution of outcomes matters. In addition to expected costs, we therefore conduct a scenario analysis in the spirit of stress testing, which illustrates that adverse interest-rate paths can generate sizable realized fiscal losses even when expected deadweight-loss costs are moderate. For example, summing program-level 95th-percentile losses yields total domestic losses of 6.65% of GDP. This distinction suggests that modest expected fiscal-efficiency costs can coexist with large realized losses in adverse states, helping explain why QE may look fiscally inexpensive ex ante yet become politically contentious ex post.

Related Literature Our paper sits at the intersection of (i) preferred-habitat/portfolio-balance views of QE and term premia, (ii) optimal government debt maturity with distortionary taxation, and (iii) the fiscal implications of large central-bank balance sheets. Since open-market operations are neutral in frictionless benchmarks (Wallace, 1981), QE requires limits to arbitrage or market segmentation. Our mechanism builds on modern preferred-habitat and "duration-risk extraction" frameworks (Greenwood and Vayanos, 2014; Vayanos and Vila, 2021; Ray, 2019; Caballero and Simsek, 2020, 2021; Kekre, Lenel, and Mainardi, 2024), with classic foundations in term-structure and portfolio-balance models (Culbertson, 1957; Modigliani and Sutch, 1966; Tobin, 1969).[5] Empirical work documents effects on long-term yields and term premia through supply and related channels (Gagnon et al., 2011; Krishnamurthy and Vissing-Jorgensen, 2011; D'Amico and King, 2013; Swanson, 2011; Hamilton and Wu, 2012a); we use the program-level survey of Fabo et al. (2021) to summarize estimates of QE output effects.

Recent work interprets post-2022 central bank losses as fiscal consequences to weigh against QE's macroeconomic benefits. Cecchetti and Hilscher (2024) document the accounting and remittance consequences of central bank losses. Most closely related is the contemporaneous work of Adrian et al. (2025), who evaluate QE within an estimated New Keynesian model. Their general-equilibrium framework captures endogenous output, inflation, and policy-rate responses that our partial estimand holds fixed, whereas we price the full *ex ante* distribution of QE portfolio outcomes under a no-arbitrage kernel and implement it program by program from security-level holdings; we view the two approaches as complementary.

On the theory side, our closest antecedents study QE with segmented markets and/or a binding ZLB. Silva (2016) studies optimal QE and risk sharing with financial segmentation but abstracts from distortionary taxation; Abadi (2023) studies segmented markets without tax distortions; and Bhattarai, Eggertsson, and Gafarov (2022) studies tax distortions without market segmentation, emphasizing state-contingent central-bank losses as a commitment device. We combine segmentation, a ZLB constraint, and distortionary taxation to study QE-induced duration exposure when the central bank's rule does not respond to losses. Related work studies large balance sheets beyond the ZLB (Vissing-Jorgensen, 2023) and their macroprudential role through bank duration risk (Eren, Jackson, and Lombardo, 2024).

Our paper also relates to optimal government debt maturity (Barro, 1979; Lucas Jr. and Stokey, 1983; Bohn, 1990). Angeletos (2002) and Buera and Nicolini (2004) show how maturity can substitute for state-contingent securities to smooth taxes with convex costs, while Lustig, Sleet, and Yeltekin (2008) and Bhandari et al. (2017) study fiscal hedging under incomplete markets. Greenwood, Hanson, and Stein (2015b) add demand for short-term safe claims and crowding out of private liquidity transformation, showing why shorter duration may be optimal; we add a production economy with a ZLB constraint and interpret QE as

[5] A complementary behavioral route is Iovino and Sergeyev (2023), where balance-sheet policies affect asset prices and output under level-k thinking.

maturity shortening that trades off output stabilization against refinancing risk. Related work studies maturity-operation effects on inflation and output (Corhay et al., 2023), maturity management and unconventional policy (Greenwood et al., 2015a; Garbade, 2015), and QE–Treasury issuance and auctions (Ray, Droste, and Gorodnichenko, 2024). Bigio, Nuño, and Passadore (2023) study issuance liquidity costs, from which we abstract, and Gomez Cram et al. (2025) show that fiscal redistribution risk can make Treasuries risky and generate nominal term premia.

Lastly, our paper relates to central bank balance-sheet risk and the fiscal backing needed to preserve monetary control. Hall and Reis (2015), Del Negro and Sims (2015), and Reis (2015) analyze monetary control absent sufficient fiscal support. Christensen, Lopez, and Rudebusch (2015) stress test Federal Reserve assets, income, capital shortfalls, and remittances along term-structure Monte Carlo paths; related projections appear in Greenlaw et al. (2013) and Carpenter et al. (2015). We share this simulation approach but assume fiscal backing, map remittance shortfalls into distortionary taxes, value the resulting flows *ex ante* under the no-arbitrage pricing kernel, and construct estimates program by program. Following the government-cost-of-capital logic of Jiang et al. (2026), we value cash flows with the government transaction-pricing kernel; marginal debt-management operations are zero net present value, so fiscal consequences come from second moments rather than expected cash flows or selected scenarios.

Relative to the existing literature, this paper makes three contributions. First, we develop an implementable fiscal-cost estimand grounded in the tax-smoothing tradition—the ex ante expected change in the present value of tax deadweight losses induced by a QE program under a conservative terminal financing rule—and show how it can be computed from the joint distribution of QE portfolio returns and taxes. Second, using granular SOMA holdings data, we quantify this fiscal-efficiency cost for each U.S. QE program, characterize the full distribution of outcomes including its tail, and compare the estimates to program-level output effects from the existing empirical literature. Third, we provide a tractable model that rationalizes the estimand: QE is a maturity-shortening policy for the consolidated government, and an interior optimality condition at the ZLB equates the marginal output benefit of risk extraction to the marginal fiscal-efficiency cost of refinancing risk.

I. QE, Debt Maturity, and the Cost of Tax-Rate Risk

In this section, we present a tractable three-period model that highlights the key trade-offs determining the optimal maturity structure of government debt when faced with a zero lower bound (ZLB) and distortionary taxes. This model also delivers the fiscal-efficiency object quantified in Section II; the measurement itself, however, does not depend on the model's specific microfoundations. We begin with a baseline setting in which the government selects its debt maturity to achieve perfect tax smoothing. We then introduce a binding ZLB constraint, prompting the government to optimally shorten its debt maturity relative to the full tax-smoothing benchmark, thereby exposing itself to refinancing risk.[6] We derive the optimal QE scale at the ZLB. Formal proofs are relegated to Appendix IA.1.

A. Environment

Consider an economy with three periods, $t \in \{0, 1, 2\}$, heterogeneous agents, one consumption good, and a single factor of production. Let $s \in \mathcal{S}$ represent the state of the economy determining period-1 productivity of capital, $a_1(s)$, which occurs with probability $\pi(s)$. This uncertainty resolves in period 1 and constitutes the sole source of risk in the economy. Productivity in periods 0 and 2, denoted a_0 and a_2, respectively, is assumed to be constant. Potential output per unit of capital equals capital's productivity a_t, but actual output per unit of capital y_t can fall below potential due to insufficient aggregate demand. The model comprises bondholders, households, and a government financing its initial spending through debt issuance and taxation. Bondholders include domestic and foreign agents.

Demography We normalize the total population in the economy to 1, of which a fraction ω consists of domestic bondholders and the remaining $1 - \omega$ of households. In addition, foreign bondholders hold a fixed fraction ϕ of government bonds. Each domestic bondholder and household owns one unit of capital, which is used to produce intermediate goods by the intermediate firms they own.

[6] Although our primary interpretation and empirical focus regard this maturity shortening as a central bank quantitative easing (QE) program, the model maintains a consolidated government perspective and remains neutral on specific implementation details. For instance, such a policy could equally be implemented by the Treasury swapping long-term Treasury bonds for short-term Treasury bills.

Notation We denote variables associated with households, domestic bondholders, and foreign bondholders by superscripts h, b, and f, respectively. When referring to the value of a variable in a specific period, we index it by the state s of the economy whenever it depends on s. For instance, productivity in periods 0, 1, and 2 is denoted by a_0, $a_1(s)$, and a_2, respectively. For notational convenience, we write a_t when referring to productivity in a generic period t.

Preferences Domestic bondholders have constant relative risk aversion (CRRA) utility over consumption c_t^b with risk aversion γ and time discount rate β:

$$V_0^b = \mathbb{E}_0 \left[\sum_{t=0}^{2} \beta^t \frac{(c_t^b)^{1-\gamma}}{1-\gamma} \right]. \tag{1}$$

Government The government spends G_0 in period 0 and does not spend in periods 1 and 2. To fund the initial spending in period 0, it can raise taxes τ_t and issue short-term bonds B_t^S and long-term bonds B_t^L. For expositional simplicity, we assume that taxes are raised only from households, whereas bonds are only held by bondholders.[7] Thus, taxes raised from each household are given by $\tau_t^h = \tau_t/(1-\omega)$.

Tax Distortions Taxes incur deadweight welfare losses of $\frac{\alpha}{2}(\tau_t^h)^2$ per household, where α controls the magnitude of these losses. Equivalently, aggregate deadweight losses in period t are given by $\frac{\alpha}{2}\frac{\tau_t^2}{1-\omega}$.

Government's Objective The government maximizes the total sum of aggregate output net of tax deadweight losses discounted by the stochastic discount factor of domestic bondholders.[8] In each period t, the government maximizes

$$V_t^g = E_t \left[\sum_{u=t}^{2} \frac{\Lambda_u}{\Lambda_t} \left(y_u - \frac{\alpha}{2} \frac{(\tau_u)^2}{1-\omega} \right) \right], \tag{2}$$

[7] In Appendix IA.3.1, we show that all our results remain valid as long as domestic bondholders hold a share of bonds larger than their tax incidence.

[8] As discussed in Section I.D below, this assumption is made to abstract from market-completion motives.

where Λ_t is the stochastic discount process of domestic bondholders and is taken as given by the government.

Production Homogeneous firms in the final good production sector take prices as given, buy intermediate goods $x_t(i)$ from intermediate firms i, and produce the final good y_t as demanded according to a constant elasticity of substitution (CES) technology:

$$y_t = \left(\int_i x_t(i)^{\frac{\varepsilon-1}{\varepsilon}} di \right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{3}$$

for some elasticity of substitution $\varepsilon > 1$. Thus, given (3), final good producers solve

$$\max_{x_t(i)} P_t y_t - \int_i P_t(i) x_t(i) di,$$

where P_t is the price of the aggregate good and $P_t(i)$ the price of the intermediate good i. To allow output to fall below potential while maintaining tractability, we follow Caballero and Simsek (2020) and assume that all intermediate firms take as given the nominal price $P_t(i) = P_t$. Hence, the problem of each individual intermediate firm $i \in [0, 1]$ is given by

$$\max_{0 \leq \eta_t(i) \leq 1} P_t(i) a_t \eta_t(i), \quad \text{s.t.} \quad a_t \eta_t(i) \leq x_t(i),$$

where $\eta_t(i) \in [0, 1]$ is the capital utilization rate it sets, and $x_t(i)$ is the demand for good i. The solution to intermediate firms' maximization problem characterizes the aggregate capital utilization rate:

$$\eta_t = \min \left\{ \frac{y_t}{a_t}, 1 \right\}. \tag{4}$$

Following Keynesian logic, a demand-driven recession is possible when demand for the final good y_t in the economy is lower than output capacity a_t per unit of capital.

B. Agents' Maximization Problem

Domestic Bondholders At time 0, domestic bondholders choose consumption and bond holdings $\{c_0^b, c_1^b(s), c_2^b(s), B_0^{S,b}, B_0^{L,b}, B_1^{S,b}(s)\}$ to maximize their lifetime expected value V_0^b, subject to the budget constraint in three periods:

$$c_0^b = y_0 - B_0^{S,b} p_0^S - B_0^{L,b} p_0^L, \tag{5}$$

$$c_1^b(s) = y_1(s) - B_1^{S,b}(s) p_1^S(s) + B_0^{S,b}, \tag{6}$$

$$c_2^b(s) = y_2(s) + B_1^{S,b}(s) + B_0^{L,b}. \tag{7}$$

Bondholders have access to the saving technology provided by the government: short-term bonds $B_t^{S,b}$ with price p_t^S in period $t \in \{0,1\}$ and long-term bonds $B_t^{L,b}$ with price p_t^L in period $t = 0$.

Foreign Bondholders Foreign bondholders purchase bonds, with consumption goods produced abroad, in proportion ϕ of the total supply. Their consumption $\{c_0^f, c_1^f(s), c_2^f(s)\}$ in each period equals their share ϕ of the government's net bond cash flows and is given explicitly in Appendix IA.1.

Households Hand-to-mouth households pay taxes $\tau_t/(1-\omega)$ and consume their income net of taxes and tax deadweight losses:

$$c_t^h = y_t - \frac{\tau_t}{1-\omega} - \frac{\alpha}{2}\left(\frac{\tau_t}{1-\omega}\right)^2 \qquad \text{for } t = 0, 1, 2. \tag{8}$$

Tax and Debt Policies Given its initial spending G_0, the government chooses taxes $\{\tau_0, \tau_1(s), \tau_2(s)\}$ and bonds $\{B_0^S, B_0^L, B_1^S(s)\}$ to maximize its objective (2) while satisfying its budget constraint in each period. Importantly, we assume that the government takes bond prices as given, which allows us to abstract from the government's cost reduction motive.[9]

[9] Since the government issues bonds to fund its spending, it has incentives to manipulate bond prices to increase its revenue and reduce the level of taxes. When a larger short-term bond share reduces the term premium, and the consolidated government's asset duration is larger than its liability duration, the government has the additional incentive to issue more short-term bonds.

Also, the government cannot commit to a prespecified policy rule.[10] Thus, in period 0, the government chooses $\{\tau_0, B_0^S, B_0^L\}$ subject to the budget constraint:

$$G_0 = \tau_0 + B_0^S p_0^S + B_0^L p_0^L. \tag{9}$$

In period 1, it chooses $\{\tau_1(s), B_1^S(s)\}$ subject to the budget constraint $0 = \tau_1(s) + B_1^S(s) p_1^S(s) - B_0^S$. Finally, in period 2, taxes $\tau_2(s)$ are raised to close the budget: $0 = \tau_2(s) - B_1^S(s) - B_0^L$.

Conventional Monetary Policy The central bank sets short-term interest rates $\{p_0^S, p_1^S(s)\}$ in order to maximize output.[11] Thus, the central bank's problem is given by $\max_{p_t^S} y_t$ for $t = 0, 1$. This assumption follows Caballero and Simsek (2020), where the short-term interest rate r_t^S is set such that output in the current period is maximized (e.g., set at full capacity) whenever possible. This target is not always admissible when the zero lower bound is binding.

C. Equilibrium

Equilibrium Definition Given government spending and productivity processes $\{G_t, a_t : t \in \{0, 1, 2\}\}$, the sequential competitive equilibrium is a set of (i) long-term bond price p_0^L; (ii) decisions for domestic bondholders $\{c_0^b, c_1^b(s), c_2^b(s), B_0^{S,b}, B_0^{L,b}, B_1^{S,b}(s)\}$; (iii) consumption by households $\{c_0^h, c_1^h(s), c_2^h(s)\}$; (iv) consumption by foreign bondholders $\{c_0^f, c_1^f(s), c_2^f(s)\}$; (v) tax policies $\{\tau_0, \tau_1(s), \tau_2(s)\}$; (vi) debt policies $\{B_0^S, B_0^L, B_1^S(s)\}$; and (vii) conventional monetary policies $\{p_0^S, p_1^S(s)\}$ such that (1) domestic bondholders' decisions and the government's policies are solutions to their respective problems given long-term bond price (i); (2) the short-term bond price in period 0, p_0^S, is bounded above by 1; and (3) markets for consumption goods, short-term bonds, and long-term bonds clear: $\omega c_t^b + (1 - \omega) c_t^h + c_t^f =$

[10] The absence of commitment, which follows from Greenwood, Hanson, and Stein (2015b) and Bhattarai, Eggertsson, and Gafarov (2022), corresponds to a more realistic assumption and yields a simpler solution.

[11] Note that allowing the government to set the interest rate instead of delegating this task to a central bank would yield a different solution since the optimal interest rate does not necessarily maximize output even if the ZLB is not binding. This discrepancy arises because the government is incentivized to reduce the debt burden by increasing the interest rate. In our analysis, we purposely abstract from the government's incentive to exploit monetary policy for pure fiscal cost reduction purposes because those have additional normative implications that are beyond the focus of this study.

$y_t - G_t - \alpha\tau_t^2/2(1-\omega)$ for $t = 0, 1, 2$; $\omega B_t^{S,b} = (1-\phi)B_t^S$ for $t = 0, 1$; and $\omega B_0^{L,b} = (1-\phi)B_0^L$.

Domestic Bondholders' First-Order Conditions The first-order conditions for domestic bondholders yield the Euler equations for bondholders:

$$p_0^S = \mathbb{E}_0\left[\frac{\Lambda_1(s)}{\Lambda_0}\right], \quad p_0^L = \mathbb{E}_0\left[\frac{\Lambda_2(s)}{\Lambda_0}\right], \quad p_1^S(s) = \frac{\Lambda_2(s)}{\Lambda_1(s)}, \tag{10}$$

where the stochastic discount factor is defined as $\Lambda_t \equiv \beta^t(c_t^b)^{-\gamma}$.

Since households are hand-to-mouth, their consumption in each period is equal to their endowment net of taxes, as given by (8). We then obtain domestic bondholders' consumption through their budget constraints and market-clearing conditions:

$$c_t^b = y_t - \frac{1-\phi}{\omega}G_t + \frac{1-\phi}{\omega}\tau_t \qquad \text{for } t = 0, 1, 2. \tag{11}$$

Because the ZLB can be binding only in period 0, the central bank is always able to follow the Wicksellian prescription of setting the period-1 interest rate equal to the natural rate, so that output is maximized: $y_1(s) = a_1(s)$. We also set $y_2 = a_2$ to pin down the equilibrium. Combining equations (10) and (11) and defining the short-term rate in period 0 as $r_0 \equiv 1/p_0^S - 1$, we get

$$r_0 = \max\left\{\left(\beta\mathbb{E}_0\left[\left(\frac{\omega a_1(s) + (1-\phi)\tau_1(s)}{\omega a_0 - (1-\phi)(G_0 - \tau_0)}\right)^{-\gamma}\right]\right)^{-1} - 1, 0\right\}, \tag{12}$$

where the max operator reflects that the ZLB is potentially binding. The corresponding expression for the period-1 rate $r_1(s)$, which involves no such constraint, is given in Appendix IA.1.

D. Discussion of Assumptions

This subsection clarifies five assumptions that isolate the mechanism studied below.

Segmentation and limited participation In frictionless complete-market benchmarks, open-market operations are neutral (Wallace, 1981). We therefore assume limited participation: hand-to-mouth households consume current income and do not trade bonds. This prevents duration risk removed from bondholders from being fully offset by private portfolio rebalancing. The hand-to-mouth structure is a tractable way to impose this property, not the only one.

Government objective In our segmented-market economy, a fully utilitarian planner would like to use state-contingent taxes to provide risk-sharing between households and bondholders. To abstract from that market-completion motive, we evaluate allocations using bondholders' stochastic discount factor, which prices government bonds. Appendix IA.3.2 allows for general government welfare weights and nests the baseline model.

Price-taking debt policy In the baseline model, we assume that the government takes bond prices as given when choosing maturity. This removes strategic price-manipulation motives aimed at lowering the tax burden (Greenwood, Hanson, and Stein, 2015b), while still allowing equilibrium prices to respond endogenously to maturity changes. Appendix IA.3.3 allows the government to partially internalize issuance-induced price effects and nests the baseline case.

Tax incidence In the baseline model, taxes fall on households rather than bondholders. QE then shifts interest-rate risk from bondholders to households through the government budget constraint, which strengthens the demand channel. This misalignment between the incidence of taxation and the ownership of duration risk is the same friction that makes QE non-neutral in the first place: if taxpayers held bonds in proportion to their tax exposure, private portfolio adjustments would undo the maturity swap (Wallace, 1981), eliminating QE's output effects and its fiscal risks alike. Appendix IA.3.1 allows for general tax incidence. The same result carries through as long as domestic bondholders' share of bondholdings exceeds their tax incidence.

Interest-bearing versus zero-interest liabilities The baseline model treats short public liabilities as a single remunerated instrument and does not distinguish between Treasury bills and reserves. This is appropriate in floor-system environments, where remunerated reserves and short Treasuries are close substitutes. Appendix IA.3.4 considers non-interest-bearing liabilities.

E. The Government's Problem

Period-1 Problem We solve the problem using backward induction and first characterize the solution in period 1, where the ZLB is never binding by assumption, output is always maximized ($y_1(s) = a_1(s)$), and the government minimizes the present value of tax deadweight losses. Given the optimality condition in (10) and the government's budget constraint, the solution to the government's period-1 problem is characterized by tax smoothing (Appendix IA.1 details the derivation):

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s)B_0^L}{1 + p_1^S(s)} \tag{13}$$

and the corresponding short-term bond issuance:

$$B_1^S(s) = \frac{B_0^S - B_0^L}{1 + p_1^S(s)}. \tag{14}$$

To minimize deadweight losses from taxation, the government rolls over short-term bonds to smooth taxes across time. If the amounts of short- and long-term bonds issued in period 0 are not equal, short-term bond issuance in period 1 is not equal to 0 ($B_1^S(s) \neq 0$), the government is exposed to refinancing risk due to its new position in short-term bonds, and taxes become risky across states. Because of the convexity of the tax cost function, this tax dispersion results in higher expected deadweight losses in period 0, which the government aims to minimize.

Period-0 Problem In Proposition 1, we characterize the optimal government policy when the ZLB is not binding.

Proposition 1 (Optimal Government Policy without the ZLB). *In period 0, if the ZLB is not binding, the government achieves first-best: Output is maximized, and tax deadweight losses are minimized. The optimal policy matches the duration of bonds with the* *duration of taxes as follows:*

(a) Tax plan: $\tau_0 = \tau_1(s) = \tau_2(s) = G_0/(1 + p_0^S + p_0^L)$*;*

(b) Bond issuance: $B_0^S = B_0^L = G_0/(1 + p_0^S + p_0^L)$*.*

When the ZLB is not binding, as in period 1, the only consideration determining the optimal bond issuance scheme is minimizing quadratic tax deadweight losses. Thus, the government smooths taxes across periods and states by matching bond and tax duration to hedge against interest-rate risk so that it never issues short-term bonds in period 1—that is, $B_1^S(s) = 0$. This result is a generalization of the benchmark solution by Greenwood, Hanson, and Stein (2015b) to a setting with risk-averse agents.

Period-0 Reformulation Since QE alters the maturity structure of debt while leaving its total level unchanged, we use the solution of the period-1 problem to restate the government's period-0 problem explicitly in terms of the total debt outstanding at period 0, $D = p_0^S B_0^S + p_0^L B_0^L$, and the short-term bond share, $S = p_0^S B_0^S / D$, which captures the interest rate risk inherent in the government's debt position. We further define $S^\star = p_0^S/(p_0^S + p_0^L)$ as the share of short-term debt that implements the complete tax-smoothing solution derived in Proposition 1, and $R_1^c(s) = p_1^S(s)/p_0^L - 1/p_0^S$ as the return on a carry position (borrowing short-term and investing long-term). The following lemma restates the government's period-0 problem according to these variable redefinitions:

Lemma 1. *The government's period-0 problem can be rewritten as:*

$$\max_{S,D} \left\{ y_0 - \frac{\alpha}{2(1-\omega)} \left((G_0 - D)^2 + D^2 \left(m \left(S - S^\star\right)^2 + \frac{1}{p_0^S + p_0^L} \right) \right) \right\}, \tag{15}$$

where $m \equiv \mathbb{E}_0 \left[\frac{\Lambda_1(s)}{\Lambda_0(1+p_1^S(s))} (R_1^c(s))^2 \right]$.

In Lemma 1 above, m is a weighted second moment of the bond market carry trade return and is closely related to the return variance, which affects the magnitude of refinancing risk assumed by the government. One can easily verify that the full tax-smoothing maturity share $S^\star$ indeed corresponds to zero refinancing risk: combining (14) with the definitions of (S, D) shows that period-1 issuance $B_1^S(s)$ is proportional to the maturity deviation $S - S^\star$

(see Appendix IA.1). Thus, setting $S = S^\star$ implies $B_1^S(s) = 0$ for all s, so the government does not issue any new short-term debt in period 1 and takes no refinancing risk, exactly as in Proposition 1. Outside the ZLB, where y_0 does not depend on S, Lemma 1 therefore implies $S = S^\star$ at the optimum.

F. Demand Recession at the ZLB

When the ZLB binds in period 0, the nominal short rate cannot fall below zero: the price of the one-period bond is at its upper bound, $p_0^S = 1$ (equivalently, $r_0 = 0$), and with sticky prices, output is demand-determined up to capacity, so equilibrium output satisfies $y_0 \leq a_0$. Because households are hand-to-mouth and do not trade assets, domestic bondholders are the marginal intertemporal decision makers, so the short-rate pricing relation (12) can be read as their Euler equation. When the ZLB binds and p_0^S is fixed, the Euler equation no longer determines the equilibrium short rate; instead, it pins down bondholders' desired current consumption and, through goods-market clearing, equilibrium output. A change in the short-term share S (holding the total market value of debt D fixed) then changes the state-contingent fiscal adjustment in period 1 and therefore bondholders' period-1 resources, to which equilibrium output must adjust so that their Euler equation continues to hold.

To make this dependence explicit, we combine bond-market clearing with bondholders' budget constraints, the period-0 government budget constraint (9), and the period-1 tax-smoothing policy (13) to express bondholders' consumption directly as a function of the maturity choice (D, S):

$$c_0^b = y_0 - \frac{1-\phi}{\omega} D, \tag{16}$$

$$c_1^b(s) = a_1(s) + \frac{1-\phi}{\omega} D \left[\frac{R_1^c(s)}{1 + p_1^S(s)} (S^\star - S) + \frac{1}{p_0^S + p_0^L} \right]. \tag{17}$$

For a given debt value D, bondholders' period-0 consumption (16) co-moves one-for-one with output and is decreasing in the amount of newly issued debt they absorb (their domestic share $(1-\phi)/\omega$), while their period-1 consumption (17) equals period-1 output plus the domestic share of debt-service transfers financed by households' taxes. The dependence on maturity is governed by the interaction of the maturity deviation $(S^\star - S)$ with the state-dependent carry return $R_1^c(s)/(1+p_1^S(s))$; under full tax smoothing ($S = S^\star$), this state-contingent component disappears. Imposing a binding ZLB ($p_0^S = 1$), bondholders' Euler equation determines c_0^b given the distribution of $c_1^b(s)$, and therefore determines y_0 through (16).

Lemma 2 (Demand Recession at the ZLB). *If the ZLB is binding, equilibrium output satisfies*

$$y_0 = \frac{1-\phi}{\omega} D + \left(\beta \mathbb{E}_0\left[\left(c_1^b(s)\right)^{-\gamma}\right]\right)^{-\frac{1}{\gamma}} \leq a_0. \tag{18}$$

Lemma 2 shows that a demand recession is possible at the ZLB. Even with $r_0 = 0$, bondholders may desire to postpone consumption (for example, because period-1 resources are low in expectation or risky), and with the short rate unable to fall further, this increase in desired saving is accommodated through a contraction in current activity: output falls below capacity ($y_0 < a_0$) rather than being offset by a lower interest rate.

We next characterize how the maturity policy affects output in a ZLB recession by differentiating the equilibrium condition in Lemma 2 with respect to the short-term share S:

Lemma 3 (Output effect of maturity at the ZLB). *If the ZLB is binding, the impact of changing debt maturity S on output y_0 is*

$$\frac{\partial y_0}{\partial S} = \frac{1-\phi}{\omega} D\, \mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(1+\gamma)} \frac{-R_1^c(s)}{1+p_1^S(s)}\right]. \tag{19}$$

Equation (19) from Lemma 3 summarizes the marginal demand effect of a maturity swap at the ZLB. The term $-R_1^c(s)/(1+p_1^S(s))$ captures how a marginal increase in S shifts bondholders' period-1 resources across states (through the refinancing component in (17)), while $\beta(c_1^b/c_0^b)^{-(1+\gamma)}$ is the marginal-utility weight that determines how these state-by-state shifts map into current consumption demand when p_0^S is fixed. The expectation in (19) decomposes into the product of the two factors' expectations, an intertemporal-substitution channel, and their covariance, a precautionary-saving channel (Appendix IA.1 details the decomposition).

The first channel captures intertemporal substitution. When long bonds carry a positive term premium, the expected discounted carry $\mathbb{E}_0\left[R_1^c(s)/(1+p_1^S(s))\right]$ is positive: on average, borrowing short and lending long is profitable. By shortening maturity, the government takes the opposite side of this trade, reducing bondholders' expected period-1 resources and making them want to save more today rather than consume, so the intertemporal-substitution component is negative and depresses consumption demand.

The second channel captures how the maturity swap reshapes bondholders' precautionary savings. Increasing S reduces the exposure of bondholders' period-1 consumption to fluctuations in the carry return and shifts the associated refinancing risk onto hand-to-mouth households through the tax adjustment. With concave utility, this reduction in bondholders' consumption risk lowers precautionary saving and raises current consumption demand; since households do not smooth intertemporally, the risk transfer is not offset by an increase in their savings.

In general, the two channels can work in opposite directions, so the sign of $\partial y_0/\partial S$ is ambiguous. Empirically, the survey evidence in Fabo et al. (2021) finds that QE raises output in ZLB environments, motivating a focus on parameterizations with $\partial y_0/\partial S > 0$. Appendix IA.2 provides a sufficient condition under which the precautionary-saving channel dominates—a restriction on the maturity deviation relative to terminal resources ensuring that $c_1^b(s)$ remains increasing in the state, so that the precautionary-saving channel is positive and sufficiently large to imply $\partial y_0/\partial S > 0$.[12] In the next section, we assume this condition is satisfied.

G. Optimal Size of a QE Program at the ZLB

We now characterize how a binding ZLB modifies the government's optimal maturity choice. In the reformulated period-0 problem (Lemma 1), the benchmark maturity share $S^\star$ implements full tax smoothing and, conditional on the level of debt D, minimizes the dispersion of future taxes across states and therefore expected deadweight losses from tax distortions; deviating from it exposes the fiscal authority to refinancing risk. At the ZLB, however, a

[12] The three-period structure is chosen to deliver a transparent trade-off between demand stabilization at the ZLB and fiscal exposure to refinancing risk. In richer environments—such as HANK models (Kaplan, Moll, and Violante, 2018) or settings with endogenous investment horizons (Hubert de Fraisse, 2024)—QE may affect output through additional general-equilibrium channels. Our quantitative analysis does not hinge on the specific microfoundation of the output effect in this section, as it disciplines the magnitude of QE's output response using program-level estimates from the empirical literature.

maturity deviation also affects equilibrium output through the aggregate-demand channel characterized in Lemma 3, so the government's optimal maturity balances a fiscal-efficiency cost against a demand-stabilization benefit. Taking the first-order condition with respect to S yields

$$\underbrace{\frac{\alpha m}{1-\omega}\,(S-S^{\star})\,D}_{\text{fiscal-efficiency cost}} = \underbrace{\frac{1}{D}\frac{\partial y_0}{\partial S}}_{\text{output benefit}}\ . \tag{20}$$

The left-hand side is the fiscal cost of QE at the margin. Starting from the tax-smoothing benchmark $S^{\star}$, a tilt toward shorter maturity makes government financing needs, and therefore future taxes, more sensitive to interest-rate realizations; because the deadweight loss from taxation is convex, this additional tax dispersion raises expected tax-distortion costs ex ante. The marginal cost is scaled by α, which governs the curvature of the deadweight-loss function; by the outstanding debt level D, which determines the fiscal exposure to refinancing shocks; and by the factor m, which summarizes how uncertainty along the yield curve translates into dispersion of the government's financing needs when maturity is shortened. The denominator $(1-\omega)$ reflects that taxes are levied on households, so that a smaller tax base generates larger per-capita tax movements and hence larger quadratic distortionary costs. Section II studies this fiscal-efficiency cost in the data by quantifying the expected tax-distortion cost associated with the maturity tilt generated by observed QE programs under a maintained holding and financing rule.

The right-hand side is the marginal demand benefit at the ZLB. Increasing S by $\mathrm{d}S$ increases the outstanding maturity swap by $\mathrm{d}Q = D\,\mathrm{d}S$, where $Q \equiv (S-S^{\star})D$ is the nominal size of the swap relative to full tax smoothing, so $\frac{1}{D}\frac{\partial y_0}{\partial S}$ can be interpreted as the output effect per additional dollar of maturity transformation, $\partial y_0/\partial Q$. When the ZLB is slack, conventional monetary policy can implement $y_0 = a_0$, the marginal output benefit is zero, and (20) collapses to $S = S^{\star}$, recovering Proposition 1. When the ZLB binds and $\partial y_0/\partial S > 0$, the government optimally chooses $S > S^{\star}$: it accepts some refinancing risk in order to raise current output.

Proposition 2 (Optimal Size of QE at the ZLB). *If the ZLB is binding and $\partial y_0/\partial S > 0$ (which holds under Condition 1), the optimal deviation from full tax smoothing (or, equiva-*

lently, the optimal size of a QE program) is given by

$$Q^\star \equiv (S - S^\star)D = \min\left\{\frac{1-\omega}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S}, \overline{Q}\right\}, \tag{21}$$

where $\overline{Q}$ is the smallest maturity swap that makes the ZLB constraint slack (i.e., the smallest $\overline{Q}$ such that the economy can attain $y_0 = a_0$).

Proposition 2 summarizes the optimal maturity intervention relative to the full tax-smoothing benchmark $S^\star$: the object $Q^\star$ is the nominal quantity of short-term liabilities that is optimally swapped—holding the total debt value D fixed—for the purpose of demand stabilization at the ZLB. From a consolidated public sector perspective, this corresponds to a QE-style maturity swap in which long-duration government claims held by the public are replaced by short-term floating-rate liabilities (reserves), and we accordingly refer to $Q^\star$ as the optimal QE program size.

The minimum operator captures that the demand-stabilization motive for maturity shortening is operative only while the ZLB binds: when the maturity swap is sufficiently large that the ZLB becomes slack, monetary policy can restore $y_0 = a_0$, so additional maturity shortening yields no marginal output benefit while continuing to raise expected tax-distortion costs. Therefore, the government never chooses a maturity swap larger than $\overline{Q}$.

In the next section, we propose a method to quantitatively evaluate the isolated fiscal-efficiency cost of QE as introduced in this theoretical section, and apply this method to the five U.S. QE programs.

II. Quantification

In this section, we quantify the fiscal-efficiency cost of quantitative easing from the consolidated government perspective formalized in the previous section. The goal is to bring to the data the part of the QE trade-off that operates through the remittances channel: QE reallocates duration risk onto the consolidated public sector, and—under fiscal backing of the central bank—any realized net income shortfall must ultimately be absorbed by the Treasury through future fiscal adjustment. Our quantitative exercise isolates the resulting change in the expected present value of tax deadweight losses, holding fixed (i) the primary fiscal stance and baseline Treasury debt-management policy and (ii) the general-equilibrium macroeconomic effects of QE.

Thus, this approach is distinct from a full general-equilibrium welfare comparison between economies with and without QE, which would require a fully specified macroeconomic model to generate a joint no-QE counterfactual for the relevant aggregates and an explicit policy rule for monetary, fiscal, and debt-management policies as well as any other general equilibrium feedback. Our focus is narrower but also more directly implementable and disciplined: it isolates a component that any broader welfare evaluation must confront once QE reallocates interest-rate risk onto the public sector. A general-equilibrium model would face the same underlying measurement challenges—constructing the QE portfolio's security-level cash-flow profile, forecasting the joint distribution of long-horizon bond returns, inflation, and GDP, calibrating the curvature of the tax-distortion function, and selecting an appropriate stochastic discount factor—but these choices would be folded into the model's parametric structure alongside auxiliary assumptions about preferences, nominal rigidities, and policy rules, making it harder to identify which inputs drive the estimated fiscal cost. Our framework keeps each of these objects explicit and open to separate scrutiny. Appendix IA.4 derives the (model-agnostic) consolidated-government accounting that maps QE portfolio returns into Treasury resources via remittances and clarifies which channels are included and which are held fixed in our quantification.

Within this framework, we report two complementary measures: an exact pricing-kernel estimate of the expected present value of the tax deadweight losses, and a conservative upper bound constructed under the most adverse comovement between QE payoffs and the fiscal environment. The next subsection derives both objects formally; we then compare the resulting estimates to program-level output effects from the empirical literature.

A. Cost Estimands

We now derive the two cost measures under the conservative terminal financing rule and provide an interpretation of the terms that will guide the empirical implementation. All formal proofs are relegated to Appendix IA.1.

Government objective and deadweight losses To accommodate the long horizon of QE programs, we extend the notation of Section I to an infinite-horizon setting with inflation. Define Y_t as the aggregate real output, P_t as the price level, τ_t as nominal taxes, and $\Lambda_{0,t}$ as the real stochastic discount factor from 0 to t entering the government's welfare objective. Further define $\theta_t = \tau_t/(P_t Y_t)$ as the tax rate. The period-0 objective of the consolidated government is

$$\max \ \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \Lambda_{0,t} Y_t \left(1 - \xi\left(\theta_t\right)\right) \right]. \tag{22}$$

We take $\xi(\theta_t)$ to represent the tax-distortion function, i.e., the real efficiency cost of taxation as a fraction of the real output. This objective is the counterpart of the government problem in Equation (2).

Notation For any variable x_t, let x_t^{QE} denote its value with the QE program in place and x_t^{nQE} its value in the no-QE counterfactual. Define $\Delta^{\mathrm{QE}} x_t \equiv x_t^{\mathrm{QE}} - x_t^{\mathrm{nQE}}$. The change in expected tax deadweight losses induced by QE is defined as

$$\Delta^{\mathrm{QE}} L_0 \equiv \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \Lambda_{0,t} Y_t \xi\left(\theta_t^{\mathrm{QE}}\right) \right] - \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \Lambda_{0,t} Y_t \xi\left(\theta_t^{\mathrm{nQE}}\right) \right]. \tag{23}$$

Interpretation of the estimand Equation (23) makes precise the ceteris-paribus nature of the exercise described above: it is the change in the expected present value of tax deadweight losses with respect to the incremental QE portfolio, holding fixed the stochastic discount factor $\Lambda_{0,t}$, the output path Y_t, the term-structure pricing environment (invariant to QE under no price impact), and the counterfactual tax process θ_t^{nQE}—a finite change evaluated at the observed incremental QE portfolio. The estimand therefore excludes by construction the general-equilibrium responses of asset prices and output to the purchases themselves; the output stimulus that motivates QE enters the cost–benefit comparison separately, through the empirical output-effect benchmarks introduced below.

QE portfolio Consider a QE program initiated on date 0 (defined empirically as the end of the purchase phase). The consolidated government purchases zero-coupon bonds $B_0(n)$

that pay \$1 in $n \leq N$ periods in quantity $\Delta^{\text{QE}}B_0(n)$ and finances the position by issuing one-period zero-coupon bonds $B_0(1)$ in quantity

$$-\Delta^{\text{QE}}B_0(1) = \sum_{n=2}^{N} \Delta^{\text{QE}}B_0(n) \times \frac{p_0(n)}{p_0(1)}, \tag{24}$$

where $p_t(n)$ denotes the date-t price of an n-period nominal zero-coupon bond.[13] During every subsequent period, we assume the government follows a passive strategy that keeps the asset-side maturity structure of the incremental portfolio constant.[14] The one-period position $\Delta^{\text{QE}}B_t(1)$ (short-term funding) is then chosen each period to finance these trades and roll over outstanding short-term funding.

Terminal financing rule At date T, the position is unwound in a single-date sale after rebalancing the portfolio. The resulting cash flow from liquidation R_T^{QE} is given by the net nominal value of the QE portfolio:

$$R_T^{\text{QE}} = \sum_{n=1}^{N} p_T(n)\, \Delta^{\text{QE}}B_T(n). \tag{25}$$

And under the terminal financing rule, QE-related taxes adjust only at the unwind date T:

$$\Delta^{\text{QE}}\tau_T = -R_T^{\text{QE}}, \qquad \Delta^{\text{QE}}\tau_t = 0 \text{ for all } t \neq T. \tag{26}$$

Cost estimate derivation We consider a second-order approximation of the tax-distortion function with a Taylor expansion around a benchmark θ_0, $\xi(\theta) \approx \xi(\theta_0) + \kappa(\theta - \theta_0) + \frac{\alpha}{2}(\theta - \theta_0)^2$, where $\kappa = \xi'(\theta_0)$ and $\alpha = \xi''(\theta_0)$. The fiscal-efficiency cost admits the following decomposition, as shown in the proof of Proposition 3 below:

$$\Delta^{\text{QE}}L_0 = \frac{\alpha}{2}\,\mathbb{E}_0\left[\Lambda_{0,T}Y_T\left(r_T^{\text{QE}}\right)^2\right] - \alpha\,\mathbb{E}_0\left[\Lambda_{0,T}Y_T\,\theta_T^{\text{nQE}}\,r_T^{\text{QE}}\right], \tag{27}$$

[13] We work with zero-coupon claims for expositional convenience. This representation is without loss of generality, as any portfolio of coupon-paying Treasury securities can be represented as a portfolio of zero-coupon bonds that replicates the same sequence of dated nominal cash flows.

[14] That is, the government purchases $\Delta^{\text{QE}}B_0(n) - \Delta^{\text{QE}}B_{t-1}(n+1)$ additional n-year zero-coupon bonds $B_t(n)$ for all $n > 1$, such that $\Delta^{\text{QE}}B_t(n) = \Delta^{\text{QE}}B_0(n)$ for every period $t \leq T$.

where $r_T^{\text{QE}} = R_T^{\text{QE}}/P_T Y_T$.

For the quantitative implementation, we rewrite the cost under the risk-neutral measure $\mathbb{Q}$. The main reason is numerical. In affine term structure models, direct evaluation of equation (27) requires simulating long-horizon moments weighted by the physical-measure stochastic discount factor, and such weighted long-horizon calculations are known to be numerically unstable: a small number of simulated paths can receive extremely large weights and dominate Monte Carlo estimates. The $\mathbb{Q}$-measure representation avoids this instability and is therefore much more reliable to implement in an arbitrage-free term structure model.

Proposition 3 (Pricing-kernel cost estimate). *Let $1/\lambda_{0,T}$ denote the date-T real value of a strategy that rolls over one-period nominal risk-free bonds up to date T (equivalently, $\lambda_{0,T}$ is the stochastic discount factor from 0 to T for real payoffs under $\mathbb{Q}$).*[15] *Then*

$$\Delta^{\text{QE}} L_0 = \frac{\alpha}{2}\, \mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T \left(r_T^{\text{QE}}\right)^2\right] - \alpha\, \mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T\, \theta_T^{\text{nQE}}\, r_T^{\text{QE}}\right]. \tag{28}$$

Equation (28) provides a convenient pricing-kernel representation of the fiscal-efficiency cost of QE, $\Delta^{\text{QE}} L_0$, implied by the terminal financing rule in equation (26). The key determinant of this cost is the state-contingent tax adjustment induced by QE at the unwind date T. Because QE-related taxes adjust only at T, the QE tax rate satisfies $\theta_T^{\text{QE}} = \theta_T^{\text{nQE}} - r_T^{\text{QE}}$, so positive portfolio returns ($r_T^{\text{QE}} > 0$) reduce the required tax rate at date T, whereas negative returns increase it. In equation (28), the factor $\lambda_{0,T} Y_T$ is the state-dependent present-value weight applied to date-T tax distortions: $\lambda_{0,T}$ is the (real) discount factor from 0 to T under the risk-neutral measure $\mathbb{Q}$, and Y_T scales deadweight losses.

In our setting, the incremental QE portfolio is constructed to have zero net worth at inception: long-bond purchases are exactly funded by short-term liabilities at prevailing prices. Under no arbitrage, the $\mathbb{Q}$-discounted expected payoff of the self-financing QE strategy is zero.[16] Accordingly, any change in expected deadweight losses arises from state-contingent second-moment properties—that is, from how QE payoffs covary with baseline fiscal conditions and the present-value weights—rather than from mechanical "carry" or arbitrage

[15] Here $\mathbb{Q}$ denotes the (real) risk-neutral measure associated with the money-market numeraire (bank account), under which discounted real asset prices are martingales. Thus, $\mathbb{E}_0[\Lambda_{0,T} X_T] = \mathbb{E}_0^{\mathbb{Q}}[\lambda_{0,T} X_T]$ for any random variable X_T. See, e.g., Duffie (2001) or Björk (2019).

[16] That is, $\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} r_T^{\text{QE}} Y_T\right] = \mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} R_T^{\text{QE}}/P_T\right] = 0$.

profits.

A common observation is that QE earns a profit in expectation: under the physical measure, the same self-financing position earns a positive expected discounted excess return whenever the term premium on the strategy is positive—the expected carry. This carry is quantitatively large; Table II reports it for each program as a memo item alongside the cost estimates. The carry nonetheless does not enter equation (28) and is not netted against the fiscal-efficiency cost. As shown above, the discounted expected payoff of the self-financing position is zero under the stochastic discount factor that prices Treasury securities—the same bondholders' kernel with which we evaluate allocations in the model (Section I). Its positive expected payoff under the physical measure is therefore compensation for priced risk, not a free fiscal resource: it is offset by the covariance of the payoff with that kernel. Booking the expected carry as a fiscal resource—valuing the government's own market transactions at a kernel other than the one at which they clear—would let the government manufacture resources through pure secondary-market trading, an argument that would apply equally to levering the public balance sheet into equities or any other priced risk factor (Jiang et al., 2026). Taxpayers who ultimately absorb the associated gains and losses may value the same state-contingent payoff differently from Treasury investors because their marginal utility across states may differ. That difference is a separate welfare effect of reallocating risk between bondholders and taxpayers, which our fiscal-cost estimand does not attempt to measure; it is not an additional fiscal cash flow generated by the QE position.[17]

The first term in (28) is a *return-dispersion* component driven by the convexity of the tax distortions. It is always weakly positive because it captures a mechanical asymmetry: when deadweight losses are convex in the tax rate, tax volatility is intrinsically costly. As the marginal excess burden of taxation increases with the tax rate, a tax rise of a given size in a "bad" state—when taxes are already high—raises distortions more than an equally sized tax cut in a "good" state—when taxes are already low—reduces them. Put differently, fluctuations push the economy into the steep part of the deadweight-loss schedule in high-tax states, while the relief obtained in low-tax states is comparatively small. Under the terminal

[17] A distinct source of genuinely cheap government financing is the convenience premium on money-like public liabilities (Krishnamurthy and Vissing-Jorgensen, 2012; Greenwood, Hanson, and Stein, 2015b). Convenience premia are level effects measurable from spreads and are conceptually separate from the compensation for duration risk; our estimand isolates the latter.

financing rule, the state-contingent unwind return on the QE position translates one-for-one into a state-contingent tax adjustment at date T. This creates a mean-preserving spread in the terminal tax rate around its counterfactual level, which necessarily raises the present value of tax distortions.

The second term of (28) is a *tax-covariance* component. Because marginal deadweight losses are increasing in the tax rate, the same unit of fiscal tightening is more costly in states in which the counterfactual tax rate θ_T^{nQE} is already high. This term therefore adjusts the QE-induced tax change r_T^{QE} against the baseline marginal distortion. If QE tends to pay off in high-tax states (θ_T^{nQE} is high when r_T^{QE} is high), then the product $\theta_T^{\mathrm{nQE}} r_T^{\mathrm{QE}}$ is large precisely when the marginal deadweight loss is high, and the negative sign implies that QE reduces the expected present value of distortions by delivering fiscal resources in the states where they are most valuable. Conversely, if QE losses are concentrated in high-tax states (r_T^{QE} is low when θ_T^{nQE} is high), then QE forces additional taxation precisely when marginal distortions are largest, raising expected deadweight losses. Hence, the largest possible contribution of the state-contingent interaction term is obtained when tax rates and real discounted QE portfolio returns are perfectly negatively correlated, which motivates the conservative upper bound provided below.

Overall, equation (28) shows that the fiscal cost of QE is pinned down by two moment objects of the terminal tax adjustment: its magnitude (captured by the weighted second moment of r_T^{QE}) and its covariance properties (captured by the weighted comovement between r_T^{QE} and the counterfactual tax rate θ_T^{nQE}). The first object is always costly under convex taxation. The second either offsets that cost when QE provides fiscal insurance, or amplifies it when QE performs poorly in states with high marginal tax distortions.

Conservative upper bound In addition to the pricing-kernel estimate (28), we derive a conservative upper bound using Cauchy–Schwarz under the assumption that tax rates and real discounted QE portfolio returns are perfectly negatively correlated and imposing a tail restriction on the path of short-term rates. This bound mitigates concerns that our affine term-structure model may understate adverse long-horizon comovement between the counterfactual tax rate and QE returns by replacing the covariance term in equation (28) with its worst-case value, so that QE delivers its most negative payoffs precisely in states where the counterfactual tax rate is high. In that sense, the bound can be interpreted as evaluating QE under an extreme scenario in which the maturity-transformation payoff provides no insurance against fiscal stress and instead amplifies it to the maximum extent.

Proposition 4 (Conservative bound). *Assume there exists $\bar{\lambda}$ such that $\lambda_{0,T} \leq \bar{\lambda}$ state-by-state. Then $\Delta^{\mathrm{QE}} L_0 \leq \overline{\Delta^{\mathrm{QE}} L_0}$, where the conservative bound is*

$$\overline{\Delta^{\mathrm{QE}} L_0} \equiv \frac{\alpha\bar{\lambda}}{2}\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[r_T^{\mathrm{QE}}\right]\mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]} + \alpha\bar{\lambda}\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[\theta_T^{\mathrm{nQE}}\right]\mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]}. \tag{29}$$

The condition $\lambda_{0,T} \leq \bar{\lambda}$ is a tail-regularity restriction on long-horizon real state prices under the $\mathbb{Q}$-measure. Recall that $1/\lambda_{0,T}$ is the date-T real value of rolling over one-period nominal risk-free bonds up to T. The restriction, therefore, rules out states in which the real discount factor becomes implausibly large, which would otherwise allow a small-probability tail event to dominate the valuation-weighted moments in equation (28). For our baseline $T = 10$-year horizon, we set $\bar{\lambda} = 2$. In the Monte Carlo simulations used for our program-level forecasts, 99.5% of paths satisfy this bound (see Appendix Figure IA.6). This choice is also historically conservative. Using yield data (Liu and Wu, 2021) and CPI data (U.S. Bureau of Labor Statistics), annual real T-bill returns always exceeded −3% (1971–2023), except in 2021 (−6.7%), 2022 (−5.9%), and 1974 (−4.4%). A worst-case 10-year scenario with these three extremes and seven years at −3% yields $\lambda_{0,10} = 1.47$, below our cap $\bar{\lambda} = 2$.

B. Data

SOMA holdings We measure each program's incremental asset purchases using the Federal Reserve's System Open Market Account (SOMA) holdings data, which reports weekly security-level positions in U.S. Treasury securities, agency debt, and agency mortgage-backed securities (MBS).[18] In Appendix IA.3.6, we provide a brief institutional description of these programs and their timing, which defines the QE episodes used in the empirical analysis.

[18]SOMA holdings are available from the Federal Reserve Bank of New York at https://www.newyorkfed.org/markets/soma-holdings.

Yield curve data We retrieve nominal zero-coupon Treasury yield curves from Liu and Wu (2021), who construct and regularly update the U.S. Treasury yield curve at monthly frequency from 1961 to the present. The maturity coverage of the underlying Treasury universe expands over time with issuance at the long end: 10-year notes begin in September 1971, 15-year bonds in December 1971, 20-year bonds in July 1981, and 30-year bonds in November 1985.

Fiscal aggregates For historical tax rates, we use annual Federal Receipts as a percentage of GDP from FRED (series FYFRGDA188S). We also retrieve nominal annual GDP from FRED (series GDPA) and CPI from U.S. Bureau of Labor Statistics (All Urban Consumers, Current Series).

C. Measurement

To compute the pricing-kernel cost estimate (28) and the upper bound (29), we construct, for each QE program, the joint distribution of the terminal QE payoff and the fiscal environment at the unwind date, conditional on information available at date 0, where date 0 denotes the end of the QE program's net purchase phase. Concretely, for a holding horizon T, we require the convexity parameter α and the joint distribution of $\left(r_T^{\mathrm{QE}}, \lambda_{0,T}, P_T, Y_T, \theta_T^{\mathrm{nQE}}\right)$.

We proceed in three steps. First, we construct the incremental portfolio acquired by the Federal Reserve during each QE program's purchase phase. Second, we employ an estimated affine term structure model to (i) forecast holding-period returns and macroeconomic variables under the physical measure and the $\mathbb{Q}$-measure and (ii) evaluate conditional expectations under the $\mathbb{Q}$-measure. Third, we calibrate the tax-distortion function using standard parameter values from the public finance literature. In doing so, we proxy the real stochastic discount factor in the loss evaluation with the term-structure-model-implied real stochastic discount factor of the marginal Treasury investor. This market-based proxy is disciplined by observed asset prices.

Construction of QE portfolios For each QE program, the initial QE portfolio is a zero-net-worth long–short position: the central bank takes long position in the long-maturity securities acquired during the program's purchase phase, $\Delta^{\mathrm{QE}}B_0(n)$, and shorts the one-year bonds issued to finance these acquisitions, $-\Delta^{\mathrm{QE}}B_0(1)$.[19] We measure net purchases in Treasury securities, agency debt, and agency MBS as the cumulative change in the Federal Reserve's SOMA holdings over each program's purchase phase.[20] For each episode, we define date 0 as the end of net purchases and treat the resulting incremental portfolio as passively held thereafter under the rollover rule described above.

To capture the interest-rate risk exposure of the incremental QE portfolios, we convert each security into its zero-coupon cash-flow representation by decomposing it into coupon payments and principal at maturity. For Treasury and agency-debt holdings, maturity dates and coupons in the SOMA releases allow us to construct dated nominal cash flows directly. For agency MBS, the SOMA releases do not report enough information on principal repayment schedules and prepayment behavior to recover dated cash flows directly. We approximate the maturity structure of MBS cash flows using the aggregate maturity profile of Treasury and agency-debt holdings. Two features of agency MBS cut in opposite directions for our purposes: agency MBS generally exhibit lower effective duration than comparable Treasury portfolios, including on an option-adjusted basis, so that our approximation overstates their average interest-rate exposure, while their negative convexity extends duration as rates rise (Hanson, 2014) and may therefore concentrate losses in the high-rate states where tax distortions are steepest. Appendix IA.5.1 documents the duration ordering in the data, including during episodes of sharply rising rates.

Table I reports program size in face value and in 10-year duration equivalents. QE1, QE3, and QE4 combine Treasury, agency, and MBS net purchases, whereas QE2's purchases were Treasuries only, with the small negative MBS entries reflecting runoff of existing holdings over its purchase phase; the MEP is close to zero in face-value terms but sizable in duration-equivalent units, reflecting the maturity swap.

[19]In practice, purchases are financed with reserves. We model funding at the one-year maturity because term-structure models typically forecast short-term rates very poorly. Using one-year rates mitigates measurement error at the short end of the curve.

[20]Although our model in Section I abstracts from agency debt and MBS, mapping these assets into the model is innocuous if (i) their interest-rate exposures are comparable to those of Treasuries at similar maturities and (ii) they are held by the same bondholder sector, so that purchases reallocate duration risk from bondholders to households in the same way as Treasury purchases. Appendix IA.3.5 formalizes this mapping and shows that, absent default, agency MBS are one-for-one fungible with long-term Treasuries in our model.

	QE1	QE2	MEP	QE3	QE4
Panel A: Face Value (% of GDP)					
Treasury and Agency	3.09	4.86	-0.22	4.50	13.13
MBS	7.19	-0.93	-0.26	5.08	5.37
All	10.27	3.93	-0.48	9.59	18.50
Panel B: 10-Year Dur. Equiv. (% of GDP)					
Treasury and Agency	2.37	3.83	4.85	6.94	12.40
MBS	5.53	-0.73	5.69	7.83	5.07
All	7.90	3.10	10.54	14.77	17.47

Table I: Size of the Federal Reserve's QE programs. The table reports net asset purchases during each program's purchase phase. Net purchases are measured as the cumulative change in SOMA holdings over the purchase phase and are expressed as a percentage of nominal GDP measured at the year-end following the end of net purchases (P_0Y_0). "Face value" reports par amounts. "10-year duration equivalents" convert the stream of promised total payments into duration-weighted units, $\sum_{t=1}^{30}(t/10)\,Pay_t$, where Pay_t denotes the principal and coupon payments due in t years.

Liquidity premia and convenience yields Treasury prices embed liquidity and convenience premia that vary across maturities and over time (Fontaine and Garcia, 2012; Krishnamurthy and Vissing-Jorgensen, 2012); the concern is that these premia, which the term-structure model does not capture separately, could contaminate the cost estimates. What matters for our exercise is not the level of these premia but their effect on the value of a zero-net-worth long–short position within the Treasury market. Three design features limit this exposure. First, both legs of the incremental portfolio are priced off the same fitted Treasury curve at inception and at unwind, so no relative cross-sectional mispricing is introduced between the securities purchased and those issued to fund the position. Because the position is duration-mismatched, a premium common to all Treasury claims does not cancel in the position value; what matters instead is the time variation of these premia over the holding period. Second, the Liu and Wu (2021) curve is constructed from a broad cross-section of coupon securities, which limits the influence of on-the-run specialness on fitted zero-coupon yields. Third, the position is funded at the one-year maturity rather than at the very short end of the curve, where the convenience premia specific to money-like instruments are concentrated (Greenwood, Hanson, and Stein, 2015b; Nagel, 2016). To the extent that very short public liabilities command larger convenience premia than one-year Treasuries, the modeled funding leg omits a source of cheap short-term funding and therefore tends, all else equal, to overstate average funding costs; the comparison is not signed path by path, since locked one-year funding can be cheaper than rolled short-term funding when rates rise. For our cost measures, however, the relevant omission is the time variation in this maturity gradient and its covariance with the QE payoff; the average spread is a level effect, conceptually separate from the compensation for duration risk that our estimand isolates.

Term-structure model We estimate a three-factor arbitrage-free affine term-structure model (see Piazzesi, 2010, for a survey) on monthly U.S. yield curve data up to and including the month in which the program's purchase phase ends. Following Hamilton and Wu (2012b), we assume the 1-month, 1-year, and 5-year yields are perfectly spanned by the three latent factors, and estimate parameters using data for 1-month, 1-year, 5-year, and 15-year yields. We follow their estimation strategy to promote global convergence. Details of the term-structure model are relegated to Appendix IA.5.2.[21] Starting from the end-of-purchases state, we simulate 1,000,000 paths of future factor realizations and construct the implied yield-curve paths needed to evaluate holding-horizon outcomes up to 15 years. The estimated affine model delivers both the physical transition of the factors and the risk-neutral dynamics under the money-market numeraire. Appendix IA.6 illustrates the physical-measure distribution of 10-year-ahead forecasts for the 1-year and 10-year yields from the end of the purchase phase of each QE program, together with realized yields. The model implies slow-moving median paths with substantial tail dispersion. For each macroeconomic variable—the tax rate, real GDP growth, and inflation—we estimate a separate forecasting equation that projects the variable on the latent term-structure factors, with an AR(1) error term to allow for serial correlation.[22] We then generate 1,000,000 paths for these variables using the corresponding

[21] We do not impose a zero lower bound (ZLB) in the term structure model. The literature offers no consensus on incorporating the ZLB in affine models, and evidence suggests it is less relevant for long maturities. Omitting the ZLB is conservative, as allowing rates to move freely increases simulated yield dispersion and thus the volatility-driven components in equations (28) and (29).

[22] Appendix IA.5.3 provides the full time-series specifications. In Appendix IA.7.1, we also show that the results are robust to using a VAR specification for the macro variables that includes the latent factors as exogenous regressors.

factor paths. These simulations deliver the horizon-T nominal GDP level $P_T Y_T$ used to compute $R_T^{\mathrm{QE}}/(P_T Y_T)$ and the no-QE tax rate θ_T^{nQE}.

Model requirements and identification The pricing-kernel estimate (28) and the conservative bound (29) are valuation objects computed under $\mathbb{Q}$, built from second-order moments of the terminal tax adjustment: the return-dispersion term is a valuation-weighted $\mathbb{Q}$-second moment of the scaled QE payoff r_T^{QE}, and, because the discounted payoff of the zero-net-worth position has zero $\mathbb{Q}$-expectation, the tax-covariance term is a pure $\mathbb{Q}$-covariance between the counterfactual tax rate θ_T^{nQE} and the discounted QE payoff $\lambda_{0,T} Y_T r_T^{\mathrm{QE}}$. Both terms are disciplined by market prices: the $\mathbb{Q}$-drift used in valuation is disciplined by the observed date-0 yield curve through no arbitrage, and the required term-structure inputs—the factor loadings of yields and the risk-neutral transition matrix $\rho^{\mathbb{Q}}$ (see Appendix IA.5.2)—are pinned down by the cross-maturity no-arbitrage restrictions on the yield panel. The counterfactual tax rate and output enter separately through macro projection equations estimated under $\mathbb{P}$ (see Appendix IA.5.3). The cost measures therefore do not use the physical-measure trend or long-run mean of the factors, nor the decomposition of yields into expectations and term premia—the objects at issue in the well-documented limitation of stationary Gaussian models, which, estimated on post-1971 samples, imply long-horizon physical forecasts that revert to sample averages while interest rates are better described by slow-moving trend components (Kozicki and Tinsley, 2001; Bauer and Rudebusch, 2020). The estimated risk-neutral factor dynamics are persistent at monthly frequency—the largest eigenvalue of $\rho^{\mathbb{Q}}$ lies between 0.9975 and 0.9986 across program estimations—so conditional yield dispersion under $\mathbb{Q}$ builds up over long horizons and remains substantial at the 10-to-15-year horizons relevant for the cost calculation. The physical measure enters the expected carry, the yield forecasts in Appendix IA.6, and the scenario analysis in Section II.G, where this mean-reverting behavior should be kept in mind.

Unspanned macro variation Our forecasting system does not restrict macro outcomes to lie in the span of the yield curve: each macroeconomic variable loads on the latent factors but retains an orthogonal, serially correlated disturbance, consistent with the evidence that yield-curve factors span only part of the variation in macroeconomic variables and their forecastable dynamics (Duffee, 2011; Joslin, Priebsch, and Singleton, 2014; Bauer and Rudebusch, 2017). These orthogonal innovations are assumed not to carry a separate price of risk: the Radon–Nikodym derivative for the risk-neutral measure depends only on the latent factors, so the $\mathbb{P} \to \mathbb{Q}$ change of measure tilts only the factor shocks and—because the residual innovations are independent of the factor shocks—leaves the conditional distribution of the unspanned residual component unchanged (Appendix IA.5.2.2 gives the formal derivation). Simulated macro paths therefore retain unspanned variation under $\mathbb{Q}$, even though their factor-driven component follows the $\mathbb{Q}$ factor dynamics.[23] The conservative bound (29) neutralizes the component most exposed to misspecified yield–macro comovement—the tax-covariance term—by replacing the estimated covariance between the counterfactual tax rate and the QE payoff with its worst-case value; the bound does not rely on the estimated sign or magnitude of that comovement, though it still uses the model-implied $\mathbb{Q}$ marginal moments of taxes and of the scaled QE return.

Maturity grid In the baseline specification, the longest maturity of the zero-coupon yield curve that is consistently available over our estimation sample (starting in 1971) is 15 years.[24] We therefore estimate the term structure model on maturities up to 15 years and implement QE portfolio cash flows on the same annual maturity grid. Each dated cash flow is mapped to the next integer maturity (in years), so that it can be priced using the corresponding model-implied zero-coupon price $p_t(n)$. This discretization slightly shifts payments outward and, therefore, if anything, increases interest-rate exposure. Cash flows with maturities beyond 15 years are aggregated at the 15-year point while preserving maturity-weighted payments. Concretely, a \$1 zero-coupon payment due in 30 years is replaced by a \$2 zero-coupon payment due in 15 years. In Appendix IA.7.1, we relax the 15-year maturity restriction by re-estimating the term structure model on a post-November-1985 sample that includes yields out to 30 years. This extension leaves all conclusions unchanged.

QE portfolio returns We compute the terminal QE payoff R_T^{QE} by applying equation (25) to each simulated yield-curve path. We set the baseline unwinding horizon to $T = 10$ years,

[23] Allowing unspanned macro risks to carry risk premia, as in Joslin, Priebsch, and Singleton (2014), would tilt the $\mathbb{Q}$-distribution of the macro variables and can in principle affect both cost terms, since the scaled QE payoff is deflated by nominal GDP $P_T Y_T$.

[24] From December 1971, yield curve data for all maturities less than or equal to 15 years are available; maturities up to 30 years become available only from November 1985 onward.

which we interpret as a conservative exposure window. This horizon is broadly consistent with both the realized pace of subsequent balance-sheet runoff and contemporaneous survey expectations about reinvestment and runoff.[25] For completeness, we also report results for alternative holding horizons in Section II.E. Figure 2 illustrates the resulting distribution of R_T^{QE} scaled by nominal GDP at the end of each program's purchase phase, $P_0 Y_0$, for the five programs at the baseline horizon $T = 10$.

Tax distortions Recall that $\xi(\theta_t)$ denotes the efficiency cost of taxation as a fraction of output, where $\theta_t \equiv \tau_t/(P_t Y_t)$ is the effective average tax rate. In the baseline framework, we used a second-order local approximation around a reference rate θ_0,

$$\xi(\theta) \approx \xi(\theta_0) + \xi'(\theta_0)(\theta - \theta_0) + \frac{1}{2}\xi''(\theta_0)(\theta - \theta_0)^2,$$

and showed that our cost estimates depend only on the curvature of the tax-distortion function $\alpha \equiv \xi''(\theta_0)$.[26] Because α enters the pricing-kernel estimate (28) and the conservative bound (29) multiplicatively, it acts as a conversion factor between a market-priced object—the $\mathbb{Q}$-measure valuation of the tax dispersion induced by QE—and its efficiency cost, and all cost estimates below can be read per unit of α.

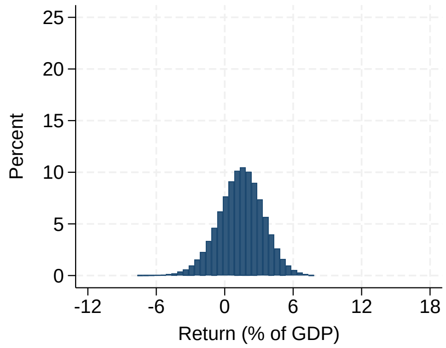
To discipline α, we use the sufficient-statistics benchmark in Saez, Slemrod, and Giertz (2012) for the marginal excess burden (MEB) of raising revenue through income taxation in the top bracket:

$$\mathrm{MEB}(\theta_m) = \frac{e\,a\,\theta_m}{1 - \theta_m - e\,a\,\theta_m},$$

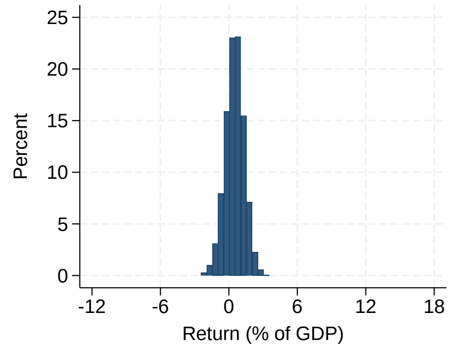
where e is the elasticity of taxable income, a is the Pareto parameter governing the thickness of the top tail of the taxable-income distribution, and θ_m is the top-bracket marginal tax rate. Under their baseline U.S. calibration ($a = 1.5$, $e = 0.25$, $\theta_m = 0.425$), this implies $\mathrm{MEB}(0.425) = 0.38$, meaning that raising an additional dollar of revenue at that margin generates an additional efficiency loss of roughly 38 cents.

[25] Appendix IA.5 provides supporting evidence based on effective holding periods and expected ones from real-time Survey of Primary Dealers vintages.

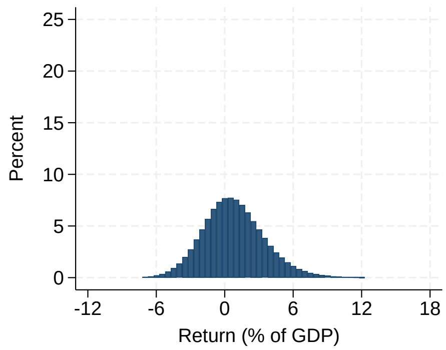
[26] In the three-period model of Section I, aggregate deadweight losses per unit of output are $\frac{\alpha}{2(1-\omega)}\theta_t^2$, where ω is the bondholder population share. The reduced-form curvature $\xi''(\theta_0)$ used here therefore corresponds to $\alpha/(1-\omega)$ in the model notation, absorbing the population-share normalization.



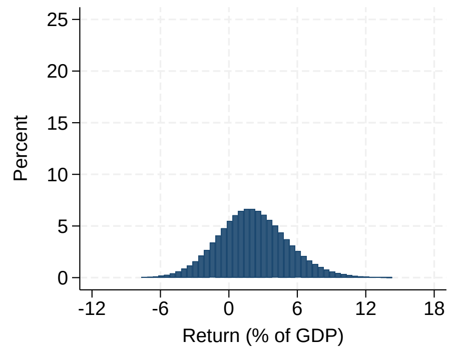
(a) QE1



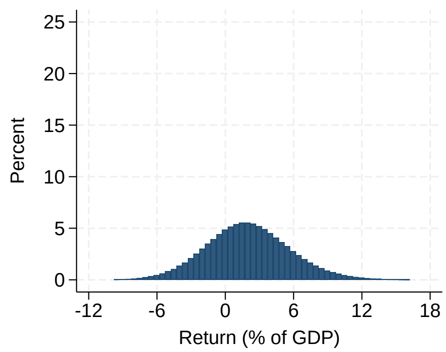
(b) QE2



(c) MEP



(d) QE3



(e) QE4

Figure 2: Distribution of QE portfolio returns. This figure plots the distribution of incremental nominal QE portfolio returns under the physical measure at a 10-year holding horizon ($T = 10$) as a percentage of nominal GDP measured at the year-end following the end of net purchases ($P_0 Y_0$) for each program. For each episode, the unwind payoff R_T^{QE} is computed from equation (25) along each of 1,000,000 term-structure simulations under physical measure. Returns are trimmed at the 0.1% level in both tails for presentation.

We calibrate α using the MEB function. Formally, we set

$$\xi'(\theta) = \text{MEB}(\theta_m(\theta)).$$

This relation implies that the local curvature of the tax-distortion function $\alpha = \xi''(\theta_0)$ is approximated by the local slope of the MEB function:

$$\xi''(\theta_0) = \text{MEB}'(\theta_m(\theta_0))\theta_m'(\theta_0),$$

where $\theta_m'(\theta_0)$ is the pass-through from the average tax rate to the top-bracket marginal tax rate. In our baseline calibration, we fix $\alpha = \text{MEB}'(0.425) = 2.17$, which corresponds to a pass-through of one, and set the benchmark rate θ_0 equal to the average federal receipts-to-GDP ratio over 2000–2009, approximately 17%, to match the U.S. tax environment underlying the Saez, Slemrod, and Giertz (2012) calibration. Two checks support this calibration. First, an alternative route assumes a globally quadratic distortion function, $\xi(\theta) = \frac{\alpha}{2}\theta^2$, as in Greenwood, Hanson, and Stein (2015b), and calibrates its slope at the benchmark rate to the same marginal excess burden, $\alpha\theta_0 = \text{MEB}(0.425)$; this yields $\alpha = 2.24$, essentially the same value under a different functional-form assumption. Second, the local approximation is applied only in the region where it is calibrated: at the 10-year horizon, the forecast distribution of the tax rate has a mean of about 17%—equal to θ_0—with an interquartile range from 15.3% to 18.4%. In Appendix IA.7.3, we demonstrate that our conclusions are robust to relaxing these assumptions, considering values of α across the range implied by estimated elasticities of taxable income, as well as considering a local third-order approximation of the tax-distortion function.

D. Output effects

We measure QE program-level output effects using the 18 studies of U.S. QE programs surveyed by Fabo et al. (2021). We express the reported estimates in a common unit: the change in output, as a percentage of GDP, per 1% of GDP in 10-year duration-equivalent purchases. When studies report a joint effect for multiple QE episodes, we allocate the total effect across programs in proportion to program size measured in 10-year duration equivalents. Because the MEP and QE4 are not directly covered in the surveyed set, we impute their output effects using the pooled size-normalized effect and scale by program size. Appendix IA.5.5 reports the detailed conversions, the trimming procedure, and the resulting list of paper–program observations.[27]

Fabo et al. (2021) document systematic differences across study characteristics: papers with central-bank-affiliated authors tend to report larger QE effects than academic-only papers, and VAR-based studies tend to imply larger effects than DSGE-based studies. Overall, aggregating across studies and programs, purchases equal to 1% of GDP in 10-year duration equivalents increase output by 0.06% of GDP. The implied effect is smaller in academic-only studies (0.02%) and in DSGE studies (0.03%), and larger in VAR studies (0.08%). We therefore report results across these study classifications. More recent papers published after the Fabo et al. (2021) survey deliver effects of comparable, and in some cases larger, magnitude once expressed in the same units.[28] Because the surveyed empirical literature reports cumulative output effects rather than a full-time profile of gains, it is difficult to assign the gains to a unique horizon. We therefore discount all output gains using the smallest real risk-free price over horizons up to T, which also makes the comparison conservative relative to discounted fiscal-efficiency cost estimates.[29] More generally, the two sides of the comparison differ in timing and in risk: the surveyed output effects are typically interpreted as short- to medium-run gains, whereas the fiscal-efficiency cost is the priced value of risky tax adjustments realized at the distant unwind date. Discounting the gains at the smallest risk-free price treats them as if they accrued at the most heavily discounted horizon, which lowers measured benefits and thus tilts the comparison conservatively. The cost side, by contrast, is a date-0 valuation that already embeds both discounting and the market price of its duration risk; this pricing is part of the estimand rather than a separate, independently established source of conservatism.

[27]The studies surveyed by Fabo et al. (2021) largely predate 2018. We therefore treat these estimates as disciplined benchmarks and examine sensitivity across study classifications.

[28]Appendix IA.5.5 reports back-of-the-envelope conversions for Fieldhouse, Mertens, and Ravn (2018), Di Maggio, Kermani, and Palmer (2020), and Ray, Droste, and Gorodnichenko (2024).

[29]That is, $\underline{p}_0^r(T) \equiv \min_{t\in\{1,\dots,T\}} p_0^r(t)$, where $p_0^r(t)$ is the date-0 price of a real risk-free claim paying one unit of consumption at date t.

E. Fiscal-Efficiency Costs

Table II reports fiscal-efficiency costs for each QE episode. For $T = 10$, the pricing-kernel estimate ranges from 0.01% to 0.12% of GDP across programs, with a cumulative cost of 0.35% of GDP. The conservative upper bound is larger by construction and ranges from 0.08% to 0.41% of GDP across programs. Dividing the cumulative estimates by cumulative purchases of 53.8 percentage points of GDP in 10-year duration equivalents (the sum across programs in Table I, each measured relative to GDP at its program-specific reference date) yields an average cost per percentage point of purchases of less than 0.01% of GDP for the pricing-kernel estimate and about 0.03% of GDP for the conservative upper bound.[30]

Cost heterogeneity across programs is economically meaningful. It mainly reflects differences in purchase size and duration composition (Table I), as well as differences in the distribution of interest rate forecasts (see Appendix Figures IA.3–IA.5).

The fiscal-efficiency costs reported in Table II are modest when benchmarked against estimated output effects from the literature. Aggregating across all programs, the cumulative output gains implied by Fabo et al. (2021) range from 1.17% of GDP (academic-only studies) to 4.22% (VAR-based studies), with an overall average of 3.28%. Consequently, at our baseline horizon, the conservative upper bound lies well below the overall mean discounted output-effect estimate and slightly above the mean for academic-only studies.

Break-even convexity Because the cost estimates are linear in α (see Section II.C), this comparison admits a transparent break-even representation. In aggregate, the pricing-kernel cost is 0.16% of GDP per unit of α and the conservative upper bound is 0.62% of GDP per unit of α, so any preferred calibration can be evaluated by rescaling. For the pricing-kernel estimate to reach the mean output-effect benchmark of 3.28% of GDP, the convexity would have to reach roughly $\alpha = 20$; to reach the academic-only benchmark of 1.17%, it would have to reach roughly $\alpha = 7$. Expressed in terms of the underlying elasticity of taxable income, these thresholds correspond to $e = 0.58$ and $e = 0.43$, both above the entire range of estimates that Saez, Slemrod, and Giertz (2012) consider most reliable ($e \in [0.12, 0.40]$).

[30] Because the return-dispersion component is quadratic in the size of the position while the tax-covariance component is linear (Table III), this pooled average describes costs at the observed program scales rather than the cost of a stand-alone one-percentage-point program; the latter would require specifying the program's portfolio composition and pricing environment.

	QE1	QE2	MEP	QE3	QE4	All
Panel A: Fiscal-efficiency cost (% of GDP)						
Pricing-kernel estimate	0.04	0.01	0.08	0.09	0.12	0.35
Conservative upper bound	0.22	0.08	0.30	0.33	0.41	1.35
Memo: expected carry (physical measure)	0.32	0.08	0.28	0.62	0.26	1.57
Panel B: Benefit (survey, % of GDP)						
Output effect (all)	0.31	0.27	0.71	0.81	1.17	3.28
Output effect (academia)	0.16	0.05	0.21	0.42	0.34	1.17
Output effect (DSGE)	0.16	0.07	0.26	0.49	0.43	1.40
Output effect (VAR)	0.38	0.38	0.94	0.97	1.55	4.22

Table II: Fiscal-efficiency costs and output effects of QE programs. This table reports the pricing-kernel cost estimate (28), the conservative upper bound (29), and output-effect benchmarks from Fabo et al. (2021), all expressed as a percentage of real GDP at the end of each program's purchase phase (Y_0). Results are shown program-by-program and aggregated across programs. The baseline holding horizon is $T = 10$ years. We discount output effects using the conservative approach detailed in Section II.C. $\mathbb{Q}$-measure distributions of QE returns are computed from 1,000,000 term-structure simulations. The conservative upper bound is computed under the assumption that $\lambda_{0,T} \leq 2$. The memo row reports the expected discounted payoff of the incremental QE portfolio under the physical measure, $\mathbb{E}_0^{\mathbb{P}}[\lambda_{0,T} R_T^{\mathrm{QE}} / P_T]/Y_0$—the expected carry, a positive entry being an expected payoff to the government rather than a cost—whose risk-neutral counterpart is zero under no arbitrage.

The conservative upper bound reaches the mean benchmark at $\alpha = 5.3$ ($e = 0.38$), near the top of that range: only the joint worst case—perfectly adverse comovement between QE payoffs and the fiscal state combined with a top-of-range elasticity—brings the estimated costs to the level of mean output-effect estimates.

We emphasize, however, that this comparison should be viewed as a disciplined benchmark of magnitudes rather than a comprehensive welfare evaluation. Specifically, our fiscal-efficiency measure captures exclusively the distortionary-financing consequences of QE under the terminal financing rule, while abstracting from additional channels and broader general-equilibrium feedback mechanisms that could influence welfare (see Appendix IA.4).

	QE1	QE2	MEP	QE3	QE4	All
Panel A: Pricing-kernel estimate (% of GDP)						
Pricing-kernel estimate	0.04	0.01	0.08	0.09	0.12	0.35
Return-dispersion component	0.02	0.00	0.05	0.05	0.09	0.22
Tax-covariance component	0.02	0.01	0.03	0.03	0.04	0.13
Panel B: Conservative upper bound (% of GDP)						
Conservative upper bound	0.22	0.08	0.30	0.33	0.41	1.35
Return-dispersion component	0.05	0.01	0.08	0.11	0.16	0.41
Tax-covariance component	0.17	0.07	0.22	0.22	0.26	0.94

Table III: Decomposition of fiscal-efficiency costs. This table decomposes the pricing-kernel cost estimate (28) and the conservative upper bound (29) into a *return-dispersion* component and a *tax-covariance* component, expressed as a percentage of real GDP at the end of each program's purchase phase (Y_0). The holding horizon is $T = 10$ years. Moments are computed from 1,000,000 term-structure simulations under the $\mathbb{Q}$-measure. The conservative bound is computed under the envelope $\lambda_{0,T} \leq 2$.

Efficiency versus incidence Our estimand is an efficiency object: it measures the deadweight burden of the state-contingent taxes required to finance QE gains and losses, not the distribution of that burden across households. The two are conceptually distinct. In the model, QE reallocates interest-rate risk from bondholders to taxpayers; in the data, this reallocation has a distributional dimension because holdings of fixed-income claims are concentrated at the top of the wealth distribution, while the incidence of the financing adjustment depends on which tax instruments adjust.[31] The simulations underlying our cost estimates do trace the state-contingent transfer between taxpayers and bondholders—the realized QE portfolio payoff itself, whose distribution is reported in Figure 2—and Section II.G separates the portion transferred to foreign holders from the domestic tax-deadweight loss. A distributional evaluation would additionally require weighting these transfers across households and specifying the incidence of the marginal tax adjustment; beyond the dispersed-tax-incidence extension in Appendix IA.3.1, a full household-level treatment is left for future work.

[31] In the Survey of Consumer Finances, bond holdings—both direct and through investment funds and retirement accounts—are held disproportionately by households in the top decile of the wealth distribution.

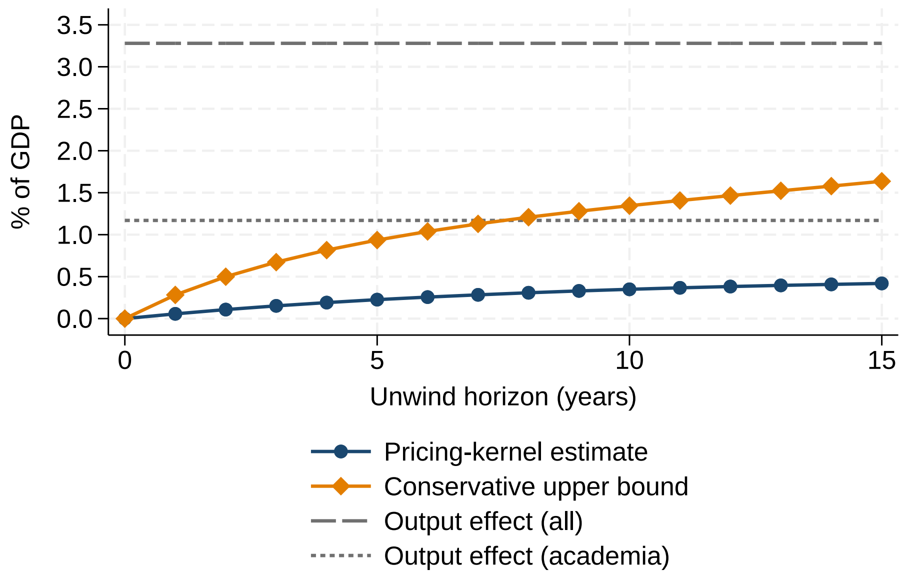


Figure 3: Aggregate fiscal-efficiency costs versus unwind horizon. The figure plots the aggregate pricing-kernel estimate (28) and the conservative upper bound (29) as a function of the unwind horizon T, expressed as a percentage of real GDP at the end of each program's purchase phase (Y_0). The dashed horizontal lines report the output-effect benchmarks from Fabo et al. (2021) described in Section II.C. Costs are computed from 1,000,000 term-structure simulations under the $\mathbb{Q}$-measure. The conservative bound is computed under the envelope $\lambda_{0,T} \leq 2$.

Decomposition As discussed above, the pricing-kernel cost in (28) can be decomposed into a *return-dispersion* component and a *tax-covariance* component. Table III reports this decomposition for both the pricing-kernel estimate and the conservative bound. At $T = 10$, both components contribute positively to the aggregate pricing-kernel estimate (0.22% of GDP from return dispersion and 0.13% from tax covariance). The positive tax-covariance contribution means that the covariance term in (28) is unfavorable in sign: because it enters as $-\alpha \, \mathrm{Cov}^{\mathbb{Q}}(\theta_T^{\mathrm{nQE}}, \lambda_{0,T} Y_T r_T^{\mathrm{QE}})$ with $\alpha > 0$, a positive contribution corresponds to a *negative* underlying covariance. Economically, QE exhibits adverse fiscal exposure in this benchmark: the discounted, output-scaled QE payoff $\lambda_{0,T} Y_T r_T^{\mathrm{QE}}$ is lower in states in which the counterfactual tax rate is higher.

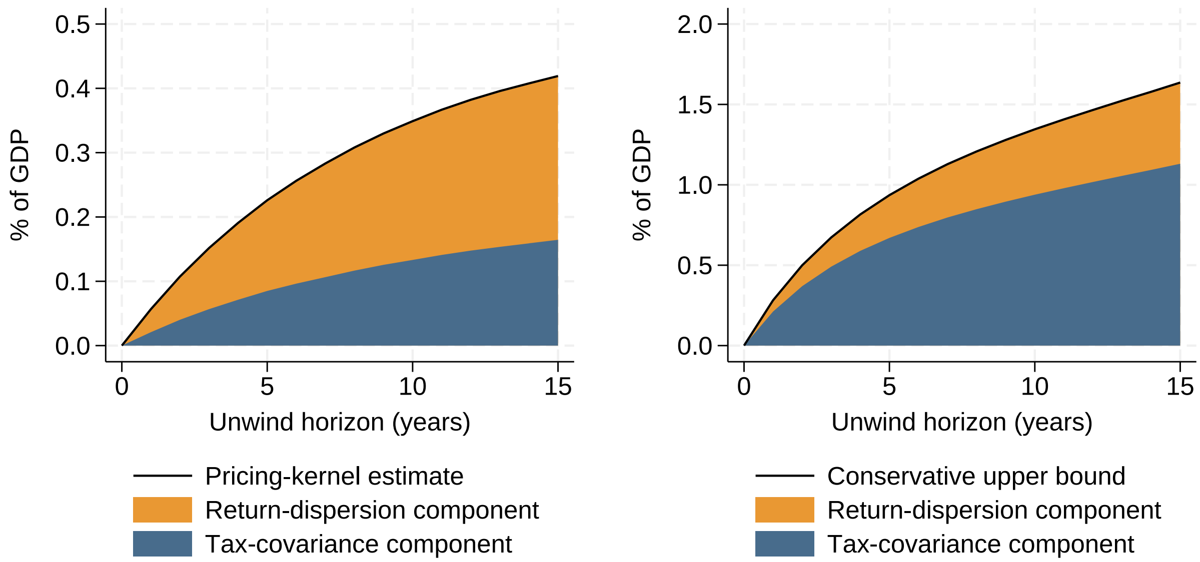
F. Sensitivity Analyses

We conduct a range of sensitivity analyses around the baseline cost–benefit comparison. Specifically, we examine alternative unwind horizons, term-structure model parameterizations, financing rules, and calibrations of the tax-distortion function. Across these exercises, the qualitative conclusions remain unchanged.

Holding horizons Figure 3 plots aggregate fiscal-efficiency costs as a function of the unwind horizon T. Holding the initial QE portfolio fixed, a longer unwind horizon keeps the consolidated government exposed to the long–short duration position for longer. Because the short leg is refinanced over time, uncertainty about future short rates cumulates as T increases, widening the distribution of the terminal QE payoff and raising expected deadweight losses under convex tax distortions. The conservative upper bound rises more steeply, reflecting its worst-case treatment of pricing weights and payoff–tax comovement. For comparison, the long-dashed and short-dashed lines report the discounted output-effect benchmark from Fabo et al. (2021), based respectively on all papers and on papers written by academics. These benchmarks are constant in T because the underlying studies report cumulative output effects at a stated horizon, which we discount back to date 0. Across horizons, the pricing-kernel estimate remains below both output-effect benchmarks. The conservative upper bound remains below the output-effect benchmark based on all papers at all horizons and below the benchmark based on academic papers up to horizons of approximately 7 years.

Figure 4 shows that both the return-dispersion and tax-covariance components rise with the unwind horizon, while their relative contributions remain broadly stable.

Term-structure parameters In Appendix IA.7.1, we recompute the aggregate pricing-kernel cost and conservative upper bound using alternative term-structure parameterizations: the model parameterized as in Hamilton and Wu (2012b), a model we estimate on post-1985 data allowing for a maturity grid up to 30 years (see discussion in Section II.C), and a model we estimate on pre-QE data (1971–2007). As shown in Appendix IA.7.1, the resulting cost profiles are close to the baseline: the estimates are slightly larger under the Hamilton–Wu and pre-2007 parameterizations, which exclude the low-interest-rate period from estimation, and slightly smaller under the post-1985 specification, which excludes the volatile interest-rate



(a) Pricing-kernel estimate

(b) Conservative upper bound

Figure 4: Decomposition by horizon. The figure decomposes aggregate fiscal-efficiency costs into *return dispersion* and *tax covariance* components as a function of the unwind horizon T. (a) reports the decomposition of the pricing-kernel estimate (28). (b) reports the decomposition of the conservative upper bound (29). Both panels are based on 1,000,000 term-structure simulations under the $\mathbb{Q}$-measure. The conservative bound is computed under the envelope $\lambda_{0,T} \leq 2$.

environment of the 1970s.

In Appendix IA.7.1, we also show that the resulting cost profiles are quantitatively similar when we replace the baseline forecasting equations for macroeconomic variables with a VAR specification that includes the latent term-structure factors as exogenous regressors.

Financing rule In the baseline analysis, we assume a terminal financing rule under which each QE portfolio's cumulative net return is offset by a one-time tax adjustment at the unwind date T. Appendix IA.7.2 relaxes this assumption by considering staggered tax adjustments instead. As shown in Appendix Figure IA.9, the cost declines when tax adjustments are spread over more future periods. This supports our interpretation that the baseline terminal financing rule is conservative.

This pattern is driven mainly by the reduction of the return-dispersion term due to convexity of the tax-distortion function, reflecting the incentive to smooth QE-induced tax adjustments over time so as to equalize marginal losses. By contrast, the tax-covariance term remains nearly unchanged because it captures the covariance between the cyclical component of counterfactual taxes and the overall tax adjustments induced by QE.

Tax distortion calibrations To relax the baseline calibration assumptions, we compute the fiscal-efficiency costs under alternative calibrations of the tax-loss convexity parameter α. Because α enters multiplicatively in (28) and (29), the cost estimates scale one-for-one with α. Appendix Figure IA.10 reports cost profiles across the interval $\alpha \in [0.72, 5.86]$, obtained by mapping the range of taxable-income elasticities that Saez, Slemrod, and Giertz (2012) consider most reliable (see Section II.E) through the marginal-excess-burden formula used in the baseline calibration, so that the interval is anchored in estimated elasticities rather than chosen values. Across this entire interval, the pricing-kernel estimate remains below the most conservative output-effect benchmark at the baseline horizon.

In Appendix IA.7.3, we also report a robustness exercise that replaces our second-order approximation of the tax-distortion function by a third-order approximation, which does not alter the qualitative conclusions.

Taken together, these results imply that the estimated cost of historical QE programs is modest relative to output-effect estimates.

G. Scenario Analysis

This section complements the expected-cost analysis with a stress-test perspective. Rather than averaging ex ante over the full distribution of outcomes, we quantify losses under adverse yield-curve paths. The exercise is analogous to stress tests applied to financial institutions: given a path for the term structure (and the associated macro path implied by our forecasting system), we compute the terminal payoff of the incremental QE position and translate it into total losses under the maintained terminal-financing rule. In this respect, we follow the probability-based stress test of Christensen, Lopez, and Rudebusch (2015), applied here to the incremental QE portfolios rather than the Federal Reserve's full balance sheet.

From portfolio payoffs to total losses We map a given payoff realization into two components: (i) the net real transfer to foreign bondholders, $\Delta^{\mathrm{QE}}F_T$, and (ii) the efficiency loss due to distortionary taxation, $\Delta^{\mathrm{QE}}L_T$. We report both because, in adverse states, part of the QE portfolio loss corresponds to a leakage of resources from domestic residents to foreign holders, while the remainder reflects the efficiency cost of financing QE gains and losses through taxes.[32] This transfer component does not enter our expected deadweight-loss cost measures: because the incremental QE portfolio has zero initial value, its pricing-kernel discounted expected payoff is zero under no arbitrage. It can, however, be quantitatively important in realized adverse scenarios.

Transfers to foreigners To capture the transfer component, we set ϕ to the share of foreign holdings of U.S. Treasury securities around each episode's purchase phase (see Table IV). The implied terminal net real transfer to foreigners is

$$\Delta^{\mathrm{QE}}F_T = -\phi \; R_T^{\mathrm{QE}}/P_T. \tag{30}$$

The share ϕ is the average foreign ownership share of outstanding Treasury securities, not the marginal foreign share in QE-related purchases and sales, which is not observable ex ante for a future unwind. Since the transfer in equation (30) is linear in ϕ, readers can rescale the transfer component of Table V one-for-one under any preferred marginal share; the tax-deadweight-loss component is unaffected by this choice.

Tax deadweight losses As in our main analysis, QE-induced gains or losses are financed by distortionary taxes at the terminal date, generating real efficiency losses. Using a second-order approximation of the tax-distortion function around the benchmark tax rate θ_0, terminal deadweight losses are

$$\Delta^{\mathrm{QE}}L_T \;=\; Y_T\left(-(\kappa-\alpha\theta_0)r_T^{\mathrm{QE}} - \alpha\theta_T^{\mathrm{nQE}}r_T^{\mathrm{QE}} + \frac{1}{2}\alpha\left(r_T^{\mathrm{QE}}\right)^2\right), \tag{31}$$

[32]From a global-welfare perspective, the transfer component nets out, so $\Delta^{\mathrm{QE}}L_T$ is the relevant welfare cost. We nevertheless report $\Delta^{\mathrm{QE}}F_T$ because the consolidated-government perspective in our framework is naturally domestic.

QE Episode	QE1	QE2	MEP	QE3	QE4
Foreign Holdings (%)	57.3	60.9	61.7	63.4	39.0

Table IV: Proportion of foreign Treasury holdings. The table reports the share of U.S. Treasury securities held by the rest of the world, computed from the Federal Reserve Board Financial Accounts of the United States (Z.1), table L.210 (Federal Government Debt Securities, Liabilities). For each QE episode, the share is measured at the end of the calendar year corresponding to the end of net purchases. The foreign-holdings share is constructed as foreign Treasury holdings divided by Treasury securities outstanding held outside the Federal Reserve and federal retirement funds.

where, following the calibration in Section II.C, we set $\theta_0 = 0.17$, $\kappa = 0.38$, $\alpha = 2.17$.[33]

Total (domestic) losses are the sum of the two components:

$$\Delta^{\mathrm{QE}} TL_T \;=\; \Delta^{\mathrm{QE}} F_T + \Delta^{\mathrm{QE}} L_T. \tag{32}$$

Stress-test results For each program, we evaluate (30)–(32) at horizon $T = 10$ using the simulated joint distribution of $(r_T^{\mathrm{QE}}, P_T, Y_T, \theta_T^{\mathrm{nQE}})$ under the physical measure generated in Section II.C. Panel A (Panel B) of Table V reports, for each program, the 95th (75th) percentiles of the marginal distribution of each loss component across simulated paths at the 10-year horizon. Throughout the table, positive entries denote losses (costs) and negative entries denote gains (benefits). Panel C reports realized outcomes along the historical yield-curve path. For QE3 and QE4, realized values are computed through 2023, since a full 10-year post-purchase window is not yet observed.

Panels D and E benchmark these loss magnitudes against the program-level output-effect estimates from Fabo et al. (2021) and report the break-even percentile at which $\Delta^{\mathrm{QE}} TL_T$ equals the output-effect benchmark. As discussed in Section II.C, we do not apply additional discounting in this tail-outcome comparison, which is plausibly conservative because fiscal impacts are realized at T while output gains typically accrue earlier.

[33] These values follow the calibration in Section II.C: $\theta_0 = 0.17$ is the benchmark tax rate, $\kappa = \mathrm{MEB}(0.425) = 0.38$ is the marginal excess burden from Saez, Slemrod, and Giertz (2012), and $\alpha = 2.17$ is the local curvature. The derivation of equation (31) takes essentially the same steps as the derivation of equation (27) in Appendix IA.1.6.

The table shows that sufficiently adverse tail realizations can imply economically meaningful losses: at the 95th percentile of the portfolio-loss distribution, total losses can be large enough to exceed conservative output-effect benchmarks for some programs. At more moderate stress scenarios (e.g., the 75th percentile), implied losses are typically modest relative to benchmark output gains. This stress-test evidence complements the expected-cost results by making transparent the magnitude of fiscal exposure implied by the duration risk embedded in the incremental QE portfolios.

III. Conclusion

This paper develops a framework to evaluate QE through the trade-off between demand stabilization and fiscal refinancing risk. In a tractable model with nominal rigidities, market segmentation, distortionary taxation, and a zero lower bound, QE is a maturity-shortening policy for the consolidated government: it extracts duration risk from bondholders, stimulating aggregate demand, but increases the dispersion of government financing needs across states. Because deadweight losses are convex in the tax rate, this dispersion raises expected tax-distortion costs. The optimal scale of QE at the ZLB equates the marginal output benefit from risk extraction to the marginal increase in expected deadweight losses from refinancing risk.

Quantifying this trade-off for the five major U.S. maturity-transformation episodes since 2008 using SOMA holdings data and an affine term structure model, we obtain a cumulative fiscal-efficiency cost of 0.35% of GDP under risk-neutral valuation and 1.35% of GDP as a conservative upper bound. When benchmarked against program-level output effects from Fabo et al. (2021)—ranging from 1.17% to 4.22% of GDP—the expected cost of 0.35% suggests that U.S. QE programs had positive expected net value at origination, and even the conservative upper bound remains below the mean output-effect estimate. The qualitative conclusions are unchanged across term-structure parameterizations, holding horizons, financing rules, and calibrations of the tax-distortion function. The conclusion is also robust to how one values the term premium QE earns in expectation: under the market kernel that prices the government's transactions, the carry is risk compensation and the zero-net-worth position is zero net present value; crediting the positive physical expected carry as a taxpayer

	QE1	QE2	MEP	QE3	QE4	All
Panel A: 95th percentile						
QE portfolio losses	1.28	0.60	2.26	1.92	2.54	8.60
Tax-deadweight losses	0.28	0.13	0.52	0.44	0.59	1.96
Transfer to Foreign Investors	0.73	0.36	1.39	1.22	0.99	4.70
Total domestic losses	1.02	0.49	1.91	1.65	1.58	6.65
Panel B: 75th percentile						
QE portfolio losses	-0.11	0.03	0.74	-0.03	0.33	0.96
Tax-deadweight losses	-0.02	0.01	0.16	-0.01	0.07	0.21
Transfer to Foreign Investors	-0.06	0.02	0.46	-0.02	0.13	0.52
Total domestic losses	-0.09	0.02	0.62	-0.02	0.20	0.73
Panel C: Realized (ex-post)						
QE portfolio losses	-3.00	-0.82	-0.41	0.58	0.43	-3.23
Tax-deadweight losses	-0.55	-0.17	-0.08	0.13	0.09	-0.58
Transfer to Foreign Investors	-1.72	-0.50	-0.25	0.37	0.17	-1.94
Total domestic losses	-2.27	-0.67	-0.34	0.50	0.26	-2.52
Panel D: Output effect (survey)						
Output effect (all)	0.36	0.28	0.71	0.85	1.18	3.38
Output effect (academia)	0.18	0.05	0.21	0.44	0.34	1.21
Panel E: Breakeven percentile						
Breakeven percentile (all)	85.59	88.52	77.10	87.94	91.31	86.09
Breakeven percentile (academia)	81.72	76.61	65.22	82.41	77.99	76.79

Table V: Cost-benefit analysis under adverse scenarios. This table reports outcomes at a 10-year horizon as a percentage of real GDP at the end of each program's purchase phase (Y_0). The first line of Panel A (Panel B) reports QE portfolio losses in stress scenarios defined by the 95th (75th) percentile of the real *portfolio-loss* distribution across the 1,000,000 term-structure simulations under physical measure (i.e., percentiles of $-Y_{10}r_{10}^{\mathrm{QE}}$). The remaining entries in Panels A and B report the 95th (75th) percentile of tax deadweight losses, transfers to foreigners, and total losses across corresponding simulations. Tax deadweight losses, transfers to foreigners, and total losses are calculated using (31), (30), and (32), respectively. Panel C reports realized outcomes given the historical yield-curve path. For QE3 and QE4, the realized horizon is truncated through 2023. Panel D reports (non-discounted) output-effect benchmarks from Fabo et al. (2021). Panel E reports the break-even percentile of QE programs, that is, the quantile of the distribution of total economic losses ($\Delta^{\mathrm{QE}}TL_{10}$) at which the corresponding output-effect benchmark is reached.

gain would only lower the measured fiscal cost.

The distribution of outcomes, however, matters. Summing program-level 95th-percentile losses gives a cumulative 6.65% of GDP, and total domestic losses exceed non-discounted estimates of output benefits from the full survey for individual programs on roughly one in ten to one in four simulated paths—modest expected costs can coexist with large realized losses in adverse states. Our estimates isolate the distortionary-financing consequences of QE-induced duration exposure and provide a disciplining input for general-equilibrium analyses incorporating distortionary taxation. Quantifying these state-contingent fiscal exposures transparently is necessary for the design of institutional arrangements that preserve the credibility and independence of monetary policy.

These conclusions rest on the assumption of fiscal backing of the central bank, whereby fiscal policy adjusts taxes to offset central-bank losses. We leave to future work the study of how inflation dynamics, sovereign spreads, and default probabilities may endogenously respond to QE-induced fiscal exposures.

References

Abadi, Joseph, 2023, Monetary policy with inelastic asset markets, Working paper.

Adrian, Tobias, Christopher J. Erceg, Marcin Kolasa, Jesper Lindé, and Pawel Zabczyk, 2025, Macroeconomic and fiscal consequences of quantitative easing, IMF Working Paper 2025/158, International Monetary Fund.

Ang, Andrew, and Monika Piazzesi, 2003, A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables, *Journal of Monetary Economics* 50, 745–787.

Angeletos, George-Marios, 2002, Fiscal policy with noncontingent debt and the optimal maturity structure, *The Quarterly Journal of Economics* 117, 1105–1131.

Balatti, Mirco, Chris Brooks, Michael P. Clements, and Konstantina Kappou, 2017, Did quantitative easing only inflate stock prices? Macroeconomic evidence from the US and UK, Working paper.

Barro, Robert J., 1979, On the determination of the public debt, *Journal of Political Economy* 87, 940–971.

Bauer, Michael D., and Glenn D. Rudebusch, 2017, Resolving the spanning puzzle in macro-finance term structure models, *Review of Finance* 21, 511–553.

Bauer, Michael D., and Glenn D. Rudebusch, 2020, Interest rates under falling stars, *American Economic Review* 110, 1316–1354.

Baumeister, Christiane, and Luca Benati, 2013, Unconventional monetary policy and the great recession: Estimating the macroeconomic effects of a spread compression at the zero lower bound, *International Journal of Central Banking* 9, 165–212.

Begenau, Juliane, Monika Piazzesi, and Martin Schneider, 2025, Banks' risk exposures, Working paper.

Bhandari, Anmol, David Evans, Mikhail Golosov, and Thomas J. Sargent, 2017, Fiscal policy and debt management with incomplete markets, *The Quarterly Journal of Economics* 132, 617–663.

Bhattarai, Saroj, Gauti B. Eggertsson, and Bulat Gafarov, 2022, Time consistency and duration of government debt: A model of quantitative easing, *The Review of Economic Studies* 90, 1759–1799.

Bigio, Saki, Galo Nuño, and Juan Passadore, 2023, Debt-maturity management with liquidity costs, *Journal of Political Economy: Macroeconomics* 1, 119–190.

Björk, Tomas, 2019, *Arbitrage Theory in Continuous Time*, fourth edition (Oxford University Press).

Bohn, Henning, 1990, Tax smoothing with financial instruments, *The American Economic Review* 80, 1217–1230.

Boucheron, Stéphane, Gábor Lugosi, and Pascal Massart, 2013, *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford University Press).

Buera, Francisco, and Juan Pablo Nicolini, 2004, Optimal maturity of government debt without state contingent bonds, *Journal of Monetary Economics* 51, 531–554.

Caballero, Ricardo J., and Alp Simsek, 2020, A risk-centric model of demand recessions and speculation, *The Quarterly Journal of Economics* 135, 1493–1566.

Caballero, Ricardo J., and Alp Simsek, 2021, A model of endogenous risk intolerance and LSAPs: Asset prices and aggregate demand in a "COVID-19" shock, *The Review of Financial Studies* 34, 5522–5580.

Carpenter, Seth B., Jane E. Ihrig, Elizabeth C. Klee, Daniel W. Quinn, and Alexander H. Boote, 2015, The federal reserve's balance sheet and earnings: A primer and projections, *International Journal of Central Banking* 11, 237–283.

Cecchetti, Stephen G., and Jens Hilscher, 2024, Fiscal consequences of central bank losses, NBER Working Paper 32478, National Bureau of Economic Research.

Chen, Han, Vasco Cúrdia, and Andrea Ferrero, 2012, The macroeconomic effects of large-scale asset purchase programmes, *The Economic Journal* 122, F289–F315.

Christensen, Jens H.E., Jose A. Lopez, and Glenn D. Rudebusch, 2015, A probability-based stress test of Federal Reserve assets and income, *Journal of Monetary Economics* 73, 26–43.

Chung, Hess, Jean-Philippe Laforte, David Reifschneider, and John C. Williams, 2012, Have we underestimated the likelihood and severity of zero lower bound events? *Journal of Money, Credit and Banking* 44, 47–82.

Corhay, Alexandre, Thilo Kind, Howard Kung, and Gonzalo Morales, 2023, Discount rates, debt maturity, and the fiscal theory, *Journal of Finance* 78, 3561–3620.

Culbertson, John M., 1957, The term structure of interest rates, *The Quarterly Journal of Economics* 71, 485–517.

Dahlhaus, Tatjana, Kristina Hess, and Abeer Reza, 2018, International transmission channels of U.S. quantitative easing: Evidence from Canada, *Journal of Money, Credit and Banking* 50, 545–563.

Dai, Qiang, and Kenneth J. Singleton, 2000, Specification analysis of affine term structure models, *The Journal of Finance* 55, 1943–1978.

D'Amico, Stefania, and Thomas B. King, 2013, Flow and stock effects of large-scale treasury purchases: Evidence on the importance of local supply, *Journal of Financial Economics* 108, 425–448.

De Grauwe, Paul, and Yuemei Ji, 2023, Monetary policies without giveaways to banks, CEPR Discussion Paper 18103, CEPR.

Del Negro, Marco, Gauti Eggertsson, Andrea Ferrero, and Nobuhiro Kiyotaki, 2017, The great escape? A quantitative evaluation of the Fed's liquidity facilities, *American Economic Review* 107, 824–57.

Del Negro, Marco, and Christopher A. Sims, 2015, When does a central bank's balance sheet require fiscal support? *Journal of Monetary Economics* 73, 1–19.

Di Maggio, Marco, Amir Kermani, and Christopher J. Palmer, 2020, How quantitative easing works: Evidence on the refinancing channel, *Review of Economic Studies* 87, 1498–1528.

Duffee, Gregory R., 2011, Information in (and not in) the term structure, *The Review of Financial Studies* 24, 2895–2934.

Duffie, Darrell, 2001, *Dynamic Asset Pricing Theory*, third edition (Princeton University Press, Princeton, NJ).

Duffie, Darrell, and Rui Kan, 1996, A yield-factor model of interest rates, *Mathematical Finance* 6, 379–406.

Engen, Eric M., Thomas Laubach, and David L. Reifschneider, 2015, The macroeconomic effects of the Federal Reserve's unconventional monetary policies, Finance and Economics Discussion Series 2015-5, Board of Governors of the Federal Reserve System.

Eren, Egemen, Timothy Jackson, and Giovanni Lombardo, 2024, The macroprudential role of central bank balance sheets, BIS Working Papers 1173, Bank for International Settlements.

Fabo, Brian, Martina Jančoková, Elisabeth Kempf, and Luboš Pástor, 2021, Fifty shades of QE: Comparing findings of central bankers and academics, *Journal of Monetary Economics* 120, 1–20.

Falagiarda, Matteo, 2014, Evaluating quantitative easing: A DSGE approach, *International Journal of Monetary Economics and Finance* 7, 302–327.

Faria-e-Castro, Miguel, and Samuel Jordan-Wood, 2023, The Fed's remittances to the Treasury: Explaining the "deferred asset", On the economy, Federal Reserve Bank of St. Louis.

Fieldhouse, Andrew J., Karel Mertens, and Morten O. Ravn, 2018, The macroeconomic effects of government asset purchases: Evidence from Postwar U.S. Housing Credit Policy, *Quarterly Journal of Economics* 133, 1503–1560.

Fontaine, Jean-Sébastien, and René Garcia, 2012, Bond liquidity premia, *The Review of Financial Studies* 25, 1207–1254.

Fuhrer, Jeffrey C., and Giovanni P. Olivei, 2011, The estimated macroeconomic effects of the Federal Reserve's large-scale Treasury purchase program, Public policy briefs, Federal Reserve Bank of Boston.

Gagnon, Joseph, Matthew Raskin, Julie Remache, and Brian Sack, 2011, The financial market effects of the Federal Reserve's large-scale asset purchases, *International Journal of Central Banking* 7, 3–43.

Gambacorta, Leonardo, Boris Hofmann, and Gert Peersman, 2014, The effectiveness of unconventional monetary policy at the zero lower bound: A cross-country analysis, *Journal of Money, Credit and Banking* 46, 615–642.

Garbade, Kenneth, 2015, Treasury debt management under the rubric of regular and predictable issuance: 1983-2012, Technical report, Federal Reserve Bank of New York.

Gertler, Mark, and Peter Karadi, 2013, QE 1 vs. 2 vs. 3. . . : A framework for analyzing large-scale asset purchases as a monetary policy tool, *International Journal of Central Banking* 9, 5–53.

Gomez Cram, Roberto, Howard Kung, Hanno Lustig, and David Zeke, 2025, Fiscal redistribution risk in Treasury markets, NBER Working Paper 33769, National Bureau of Economic Research.

Greenlaw, David, James D. Hamilton, Peter Hooper, and Frederic S. Mishkin, 2013, Crunch time: Fiscal crises and the role of monetary policy, Report, U.S. Monetary Policy Forum, Also NBER Working Paper No. 19297.

Greenwood, Robin, Samuel Hanson, Joshua S. Rudolph, and Lawrence Summers, 2015a, The optimal maturity of government debt, in *The $13 Trillion Question: How America Manages Its Debt*, 1–41 (Brookings Institution Press).

Greenwood, Robin, Samuel G. Hanson, and Jeremy C. Stein, 2015b, A comparative-advantage approach to government debt maturity, *The Journal of Finance* 70, 1683–1722.

Greenwood, Robin, and Dimitri Vayanos, 2014, Bond supply and excess bond returns, *The Review of Financial Studies* 27, 663–713.

Haldane, Andrew, Matt Roberts-Sklar, Chris Young, and Tomasz Wieladek, 2016, QE: The story so far, Working paper, Bank of England.

Hall, Robert E., and Ricardo Reis, 2015, Maintaining central-bank financial stability under new-style central banking, Working Paper 21173, National Bureau of Economic Research.

Hamilton, James D., and Jing Cynthia Wu, 2012a, The effectiveness of alternative monetary policy tools in a zero lower bound environment, *Journal of Money, Credit and Banking* 44, 3–46.

Hamilton, James D., and Jing Cynthia Wu, 2012b, Identification and estimation of gaussian affine term structure models, *Journal of Econometrics* 168, 315–331.

Hanson, Samuel G., 2014, Mortgage convexity, *Journal of Financial Economics* 113, 270–299.

Hesse, Henning, Boris Hofmann, and James Michael Weber, 2018, The macroeconomic effects of asset purchases revisited, *Journal of Macroeconomics* 58, 115–138.

Hubert de Fraisse, Antoine, 2024, Crowding out long-term corporate investment: The role of long-term government debt supply, Working paper.

Iovino, Luigi, and Dmitriy Sergeyev, 2023, Central bank balance sheet policies without rational expectations, *The Review of Economic Studies* 90, 3119–3152.

Jiang, Zhengyang, Hanno Lustig, Stijn Van Nieuwerburgh, and Mindy Z. Xiaolan, 2026, Manufacturing risk-free government debt, *Journal of Financial Economics* 176, 104203.

Joslin, Scott, Marcel Priebsch, and Kenneth J. Singleton, 2014, Risk premiums in dynamic term structure models with unspanned macro risks, *The Journal of Finance* 69, 1197–1233.

Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante, 2018, Monetary policy according to HANK, *American Economic Review* 108, 697–743.

Kekre, Rohan, Moritz Lenel, and Federico Mainardi, 2024, Monetary policy, segmentation, and the term structure, Working Paper 32324, National Bureau of Economic Research, Revision date: March 2025.

Kim, Kyuhoon, Thomas Laubach, and Min Wei, 2020, Macroeconomic effects of large-scale asset purchases, Finance and Economics Discussion Series 2020-047r1, Board of Governors of the Federal Reserve System, Revised August 2023.

Kozicki, Sharon, and Peter A. Tinsley, 2001, Shifting endpoints in the term structure of interest rates, *Journal of Monetary Economics* 47, 613–652.

Krishnamurthy, Arvind, and Annette Vissing-Jorgensen, 2011, The effects of quantitative easing on interest rates: Channels and implications for policy, *Brookings Papers on Economic Activity* 42, 215–287.

Krishnamurthy, Arvind, and Annette Vissing-Jorgensen, 2012, The aggregate demand for treasury debt, *Journal of Political Economy* 120, 233–267.

Liu, Yan, and Jing Cynthia Wu, 2021, Reconstructing the yield curve, *Journal of Financial Economics* 142, 1395–1425.

Lucas Jr., Robert E., and Nancy L. Stokey, 1983, Optimal fiscal and monetary policy in an economy without capital, *Journal of Monetary Economics* 12, 55–93.

Lustig, Hanno, Christopher Sleet, and Şevin Yeltekin, 2008, Fiscal hedging with nominal assets, *Journal of Monetary Economics* 55, 710–727.

Modigliani, Franco, and Richard Sutch, 1966, Innovations in interest rate policy, *The American Economic Review* 56, 178–197.

Nagel, Stefan, 2016, The liquidity premium of near-money assets, *The Quarterly Journal of Economics* 131, 1927–1971.

Piazzesi, Monika, 2010, Affine term structure models, in *Handbook of Financial Economics*, chapter 12, 691–766 (Elsevier).

Popescu, Adina, 2015, Did large-scale asset purchases work? Working paper.

Ray, Walker, 2019, Monetary policy and the limits to arbitrage: Insights from a new Keynesian preferred habitat model, 2019 Meeting Papers 692, Society for Economic Dynamics.

Ray, Walker, Michael Droste, and Yuriy Gorodnichenko, 2024, Unbundling quantitative easing: Taking a cue from Treasury auctions, *Journal of Political Economy* 132, 3115–3172.

Reis, Ricardo, 2015, Different types of central bank insolvency and the central role of seignorage, Working Paper 21226, National Bureau of Economic Research.

Roubini, Nouriel, and Stephen Miran, 2024, The US Treasury's backdoor stimulus is hampering the Fed, Project Syndicate, Accessed: 2024-10-19.

Saez, Emmanuel, Joel Slemrod, and Seth H. Giertz, 2012, The elasticity of taxable income with respect to marginal tax rates: A critical review, *Journal of Economic Literature* 50, 3–50.

Silva, Dejanir H., 2016, The risk channel of unconventional monetary policy, Working paper.

Swanson, Eric T., 2011, Let's twist again: A high-frequency event-study analysis of Operation Twist and its implications for QE2, *Brookings Papers on Economic Activity* 42, 151–207.

Tobin, James, 1969, A general equilibrium approach to monetary theory, *Journal of Money, Credit and Banking* 1, 15–29.

Vayanos, Dimitri, and Jean-Luc Vila, 2021, A preferred-habitat model of the term structure of interest rates, *Econometrica* 89, 77–112.

Vissing-Jorgensen, Annette, 2023, Balance sheet policy above the ELB, in *ECB Forum on Central Banking*, Sintra.

Wallace, Neil, 1981, A Modigliani-Miller theorem for open-market operations, *The American Economic Review* 71, 267–274.

Weale, Martin, and Tomasz Wieladek, 2016, What are the macroeconomic effects of asset purchases? *Journal of Monetary Economics* 79, 81–93.

Wu, Jing Cynthia, and Fan Dora Xia, 2016, Measuring the macroeconomic impact of monetary policy at the zero lower bound, *Journal of Money, Credit and Banking* 48, 253–291.

Internet Appendix

IA.1. Proofs

Foreign bondholders hold a fraction ϕ of each bond issue, so their consumption in each period equals their share ϕ of the government's net bond cash flows:

$$c_0^f = -\phi(p_0^S B_0^S + p_0^L B_0^L), \tag{IA.1}$$

$$c_1^f(s) = \phi(B_0^S - p_1^S(s) B_1^S(s)), \tag{IA.2}$$

$$c_2^f(s) = \phi(B_1^S(s) + B_0^L). \tag{IA.3}$$

Because the zero lower bound can bind only in period 0, the period-1 short rate $r_1(s) \equiv 1/p_1^S(s) - 1$, obtained from bondholders' Euler equation (10) and their equilibrium consumption (11), involves no lower-bound constraint:

$$r_1(s) = \left(\beta \left(\frac{\omega a_2 + (1-\phi)\tau_2(s)}{\omega a_1(s) + (1-\phi)\tau_1(s)}\right)^{-\gamma}\right)^{-1} - 1. \tag{IA.4}$$

IA.1.1. Proof of Proposition 1

The government's problem in period 0 is given by

$$V_0^g = \mathbb{E}_0 \left[\sum_{t=0}^{2} \frac{\Lambda_t}{\Lambda_0} \left(y_t - \frac{\alpha}{2} \frac{\tau_t^2}{1-\omega}\right)\right], \tag{IA.5}$$

subject to the budget constraint

$$G_0 = \tau_0 + B_0^S p_0^S + B_0^L p_0^L. \tag{IA.6}$$

From the solution of the period-1 problem in equation (13), we know

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s) B_0^L}{1 + p_1^S(s)} \quad \text{and} \quad B_1^S(s) = \frac{B_0^S - B_0^L}{1 + p_1^S(s)}. \tag{IA.7}$$

Since the ZLB is not binding, $y_t = a_t$. Substituting in all constraints and taking the first-order derivatives with respect to B_0^S and B_0^L yield

$$\mathbb{E}_0\left[-\tau_0 p_0^S + \frac{\Lambda_1(s)}{\Lambda_0}\tau_1(s)\frac{1}{1+p_1^S(s)} + \frac{\Lambda_2(s)}{\Lambda_0}\tau_2(s)\frac{1}{1+p_1^S(s)}\right] = 0, \tag{IA.8}$$

$$\mathbb{E}_0\left[-\tau_0 p_0^L + \frac{\Lambda_1(s)}{\Lambda_0}\tau_1(s)\frac{p_1^S(s)}{1+p_1^S(s)} + \frac{\Lambda_2(s)}{\Lambda_0}\tau_2(s)\frac{p_1^S(s)}{1+p_1^S(s)}\right] = 0. \tag{IA.9}$$

Combining the government's first-order conditions, bondholders' first-order conditions and market-clearing conditions gives

$$\tau_0 = \tau_1(s) = \tau_2(s) = \frac{1}{1+p_0^S+p_0^L}G_0, \tag{IA.10}$$

$$B_0^S = B_0^L = \frac{1}{1+p_0^S+p_0^L}G_0. \tag{IA.11}$$

IA.1.2. Proof of Lemma 1

We plug the solution to the period-1 problem into the period-0 objective, and omit y_1, y_2 terms because they are exogeneous. The objective is then formulated as follows

$$\mathbb{E}_0\left[y_0 - \frac{\alpha}{2(1-\omega)}\left((\tau_0)^2 + \frac{\Lambda_1(s)}{\Lambda_0}(\tau_1(s))^2 + \frac{\Lambda_2(s)}{\Lambda_0}(\tau_2(s))^2\right)\right] \tag{IA.12}$$

$$=\mathbb{E}_0\left[y_0 - \frac{\alpha}{2(1-\omega)}\left((G_0 - p_0^S B_0^S - p_0^L B_0^L)^2 + \left(\frac{\Lambda_1(s)}{\Lambda_0} + \frac{\Lambda_2(s)}{\Lambda_0}\right)\left(\frac{B_0^S + p_1^S(s)B_0^L}{1+p_1^S(s)}\right)^2\right)\right] \tag{IA.13}$$

$$=\mathbb{E}_0\left[y_0 - \frac{\alpha}{2(1-\omega)}\left((G_0 - p_0^S B_0^S - p_0^L B_0^L)^2 + \frac{\Lambda_1(s)}{\Lambda_0}\frac{(B_0^S + p_1^S(s)B_0^L)^2}{1+p_1^S(s)}\right)\right], \tag{IA.14}$$

where we have used the Euler equation $p_1^S(s) = \frac{\Lambda_2(s)}{\Lambda_1(s)}$ to simplify the expression.

Define $D = p_0^S B_0^S + p_0^L B_0^L, S = p_0^S B_0^S / D$. Then

$$\mathbb{E}_0 \left[y_0 - \frac{\alpha}{2(1-\omega)} \left((G_0 - p_0^S B_0^S - p_0^L B_0^L)^2 + \frac{\Lambda_1(s)}{\Lambda_0} \frac{(B_0^S + p_1^S(s) B_0^L)^2}{1 + p_1^S(s)} \right) \right] \tag{IA.15}$$

$$= \mathbb{E}_0 \left[y_0 - \frac{\alpha}{2(1-\omega)} \left((G_0 - D)^2 + \frac{\Lambda_1(s)}{\Lambda_0} D^2 \frac{\left(\frac{S}{p_0^S} + p_1^S(s) \frac{1-S}{p_0^L} \right)^2}{1 + p_1^S(s)} \right) \right] \tag{IA.16}$$

$$= y_0 - \frac{\alpha}{2(1-\omega)} \left[(G_0 - D)^2 + D^2 \left[m \left(S - \frac{p_0^S}{p_0^S + p_0^L} \right)^2 + \frac{1}{p_0^S + p_0^L} \right] \right], \tag{IA.17}$$

where

$$m = E_0 \left[\frac{\Lambda_1(s)}{\Lambda_0 (1 + p_1^S(s))} \left(\frac{1}{p_0^S} - \frac{p_1^S(s)}{p_0^L} \right)^2 \right]. \tag{IA.18}$$

Finally, combining the period-1 issuance (14) with the definitions $D = p_0^S B_0^S + p_0^L B_0^L$ and $S = p_0^S B_0^S / D$ shows that period-1 issuance is proportional to the maturity deviation from full tax smoothing:

$$B_1^S(s) = \frac{B_0^S - B_0^L}{1 + p_1^S(s)} = \frac{D}{1 + p_1^S(s)} \left(\frac{1}{p_0^S} + \frac{1}{p_0^L} \right) (S - S^\star), \tag{IA.19}$$

so that $S = S^\star$ implies $B_1^S(s) = 0$ for all s.

IA.1.3. Proof of Lemma 2

From the first-order condition for B_0^S, we get

$$p_0^S = \beta \mathbb{E}_0 \left[\left(\frac{c_1^b(s)}{c_0^b} \right)^{-\gamma} \right]. \tag{IA.20}$$

Given the market-clearing conditions for consumption goods, we have

$$c_0^b = y_0 - \frac{1-\phi}{\omega} G_0 + \frac{1-\phi}{\omega}\tau_0, \tag{IA.21}$$

$$c_1^b(s) = y_1(s) + \frac{1-\phi}{\omega}\tau_1(s). \tag{IA.22}$$

Then,

$$p_0^S = \beta \mathbb{E}_0 \left[\left(\frac{\omega y_1(s) + (1-\phi)\tau_1(s)}{\omega y_0 - (1-\phi)G_0 + (1-\phi)\tau_0} \right)^{-\gamma} \right]. \tag{IA.23}$$

When the ZLB is binding, $p_0^S = 1$. Since period-1 output $y_1(s)$ is always maximized, $y_1(s) = a_1(s)$. From the solution to the period-1 problem,

$$\tau_1(s) = \frac{B_0^S + p_1^S(s) B_0^L}{1 + p_1^S(s)} = D \left(\frac{-R_1^c(s)}{1 + p_1^S(s)} (S - S^\star) + \frac{1}{p_0^S + p_0^L} \right). \tag{IA.24}$$

Then, we can rewrite equation (IA.23) as

$$y_0 = \frac{1-\phi}{\omega} D + \left(\beta \mathbb{E}_0 \left[\left(a_1(s) + \frac{1-\phi}{\omega} D \left[\frac{-R_1^c(s)}{1 + p_1^S(s)} (S - S^\star) + \frac{1}{p_0^S + p_0^L} \right] \right)^{-\gamma} \right] \right)^{-\frac{1}{\gamma}}. \tag{IA.25}$$

Since the aggregate output can never exceed productivity, $y_0 \leq a_0$.

IA.1.4. Proof of Lemma 3

If the ZLB is binding,

$$y_0 = \frac{1-\phi}{\omega} D + \left(\beta \mathbb{E}_0 \left[\left(a_1(s) + \frac{1-\phi}{\omega} D \left[\frac{-R_1^c(s)}{1 + p_1^S(s)} (S - S^\star) + \frac{1}{p_0^S + p_0^L} \right] \right)^{-\gamma} \right] \right)^{-\frac{1}{\gamma}}. \tag{IA.26}$$

Taking the partial derivative of aggregate output with respect to D and S yields

$$\frac{\partial y_0}{\partial D} = \frac{1-\phi}{\omega} + \frac{1-\phi}{\omega}\beta\mathbb{E}_0\left[\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)} \frac{\frac{S}{p_0^S} + \frac{1-S}{p_0^L}p_1^S(s)}{1+p_1^S(s)}\right] \tag{IA.27}$$

and

$$\frac{\partial y_0}{\partial S} = \frac{1-\phi}{\omega}\beta\mathbb{E}_0\left[\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)} \frac{-R_1^c(s)}{1+p_1^S(s)}\right] D, \tag{IA.28}$$

where we have simplified the expression with $c_0^b = y_0 - \frac{1-\phi}{\omega}(G_0 - \tau_0)$, $c_1^b(s) = a_1(s) + \frac{1-\phi}{\omega}\tau_1(s)$, $\tau_0 = G_0 - D$, and $\tau_1(s) = D\left[\frac{-R_1^c(s)}{1+p_1^S(s)}(S - S^\star) + \frac{1}{p_0^S + p_0^L}\right]$.

Writing the expectation in (19) as the product of the two factors' expectations plus their covariance separates the marginal output effect into an intertemporal-substitution channel and a precautionary-saving channel:

$$\begin{aligned}\frac{\partial y_0}{\partial S} = \frac{1-\phi}{\omega} D \Bigg(& \underbrace{\mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(1+\gamma)}\right]\mathbb{E}_0\left[\frac{-R_1^c(s)}{1+p_1^S(s)}\right]}_{\text{intertemporal substitution}} \\ & + \underbrace{\mathrm{Cov}_0\left(\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(1+\gamma)}, \frac{-R_1^c(s)}{1+p_1^S(s)}\right)}_{\text{precautionary saving}}\Bigg).\end{aligned} \tag{IA.29}$$

IA.1.5. Proof of Proposition 2

We take the first-order derivative with respect to S in the reformulated problem in Lemma 1 and get

$$\frac{\alpha}{1-\omega} D^2 m \left(S - \frac{p_0^S}{p_0^S + p_0^L}\right) = \frac{\partial y_0}{\partial S}. \tag{IA.30}$$

Given an inner solution, the optimal short-term bond share is thereby

$$S = \frac{p_0^S}{p_0^S + p_0^L} + \frac{1-\omega}{\alpha m D}\frac{1}{D}\frac{\partial y_0}{\partial S}, \tag{IA.31}$$

and the optimal size of QE is

$$Q = D(S - S^\star) = \frac{1-\omega}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S}. \tag{IA.32}$$

If $y_0 \geq a_0$ in the equilibrium characterized by the first-order condition above, the inner solution is not admissible. To also account for the corner solutions, we consider the Karush–Kuhn–Tucker conditions:

$$\frac{\alpha}{1-\omega}D^2 m(S - S^\star) - (1-\mu)\frac{\partial y_0}{\partial S} = 0, \tag{IA.33}$$

where $\mu \geq 0$ is the Lagrange multiplier associated with the constraint $a_0 - y_0 \geq 0$, with complementary slackness $\mu\,(a_0 - y_0) = 0$. Therefore, the optimal size of QE is

$$Q = (1-\mu)\frac{1-\omega}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S}. \tag{IA.34}$$

We define $\overline{Q}$ as the minimum size of QE pushing the economy out of the ZLB, that is, the solution to the following implicit function:

$$\left(a_0 - \frac{1-\phi}{\omega}D\right)^{-\gamma} - \beta\mathbb{E}_0\left[\left(a_1(s) + \frac{1-\phi}{\omega}\frac{-R_1^c(s)}{1+p_1^S(s)}Q + \frac{1-\phi}{\omega}\frac{D}{p_0^S + p_0^L}\right)^{-\gamma}\right] = 0. \tag{IA.35}$$

Since $\mu(a_0 - y_0) = 0$, $Q = \overline{Q}$ when $\mu > 0$. Remember $\frac{1}{D}\frac{\partial y_0}{\partial S} > 0$. Then, $Q = \overline{Q} = (1 - \mu)\frac{1-\omega}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S} < \frac{1-\omega}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S}$ when $\mu > 0$, and $Q = \frac{1-\omega}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S}$ when $\mu = 0$. We have $y_0 \leq a_0$ when $\mu = 0$, and thus $Q = \frac{1-\omega}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S} \leq \overline{Q}$ since $\frac{1}{D}\frac{\partial y_0}{\partial S} > 0$. In summary, $Q = \min\{\frac{1-\omega}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S}, \overline{Q}\}$.

IA.1.6. Proof of Proposition 3

In our modified model, the expression for time-0 expected total deadweight losses is

$$L_0 = \mathbb{E}_0\left[\sum_{t=0}^{\infty}\Lambda_{0,t}Y_t\xi\left(\theta_t\right)\right]. \tag{IA.36}$$

We denote time t tax rate θ_t^{QE}, and the counterfactual tax rate in the absence of QE programs θ_t^{nQE}. The contribution of QE to expected total deadweight losses can be expressed as:

$$\Delta^{\text{QE}} L_0 = \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \Lambda_{0,t} Y_t \left(\xi \left(\theta_t^{\text{QE}} \right) - \xi \left(\theta_t^{\text{nQE}} \right) \right) \right]. \tag{IA.37}$$

We define $\Delta^{\text{QE}} \theta_t = \theta_t^{\text{QE}} - \theta_t^{\text{nQE}}$ where $\Delta^{\text{QE}} \theta_t$ is the contribution of QE on the tax rate. Assume the loss function $\xi(\theta)$ is at least twice continuously differentiable. For all θ close to the benchmark θ_0, we approximate $\xi(\theta)$ by its second-order Taylor expansion around θ_0, $\xi(\theta_0) + \xi'(\theta_0)(\theta - \theta_0) + \frac{\xi''(\theta_0)}{2}(\theta - \theta_0)^2$. Then the difference of loss functions is

$$\xi \left(\theta_t^{\text{QE}} \right) - \xi \left(\theta_t^{\text{nQE}} \right) \tag{IA.38}$$

$$= \xi'(\theta_0)(\theta_t^{\text{QE}} - \theta_t^{\text{nQE}}) + \frac{\xi''(\theta_0)}{2}(\theta_t^{\text{QE}} - \theta_t^{\text{nQE}})(\theta_t^{\text{QE}} + \theta_t^{\text{nQE}} - 2\theta_0) \tag{IA.39}$$

$$= (\xi'(\theta_0) - \xi''(\theta_0)\theta_0)\Delta^{\text{QE}}\theta_t + \xi''(\theta_0)\theta_t^{\text{nQE}}\Delta^{\text{QE}}\theta_t + \frac{\xi''(\theta_0)}{2}(\Delta^{\text{QE}}\theta_t)^2, \tag{IA.40}$$

and $\Delta^{\text{QE}} L_0$ is approximated by:

$$\Delta^{\text{QE}} L_0 = \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \Lambda_{0,t} Y_t \left((\kappa - \alpha\theta_0)\Delta^{\text{QE}}\theta_t + \alpha\theta_t^{\text{nQE}}\Delta^{\text{QE}}\theta_t + \frac{\alpha}{2}(\Delta^{\text{QE}}\theta_t)^2 \right) \right], \tag{IA.41}$$

where we denote $\kappa = \xi'(\theta_0)$, $\alpha = \xi''(\theta_0)$. We assume $\Delta^{\text{QE}}\tau_T = -R_T^{\text{QE}}$ and $\Delta^{\text{QE}}\tau_t = 0$ for any $t \neq T$. Also note that $\mathbb{E}_0 \left[\Lambda_{0,T} Y_T \Delta^{\text{QE}} \theta_T \right] = -\mathbb{E}_0 \left[\Lambda_{0,T} R_T^{\text{QE}} / P_T \right] = 0$ according to the Euler equation. Therefore, the expression for $\Delta^{\text{QE}} L_0$ can be rewritten as

$$\Delta^{\text{QE}} L_0 = \frac{\alpha}{2} \mathbb{E}_0 \left[\Lambda_{0,T} Y_T \left(r_T^{\text{QE}} \right)^2 \right] - \alpha \mathbb{E}_0 \left[\Lambda_{0,T} Y_T \left(\theta_T^{\text{nQE}} r_T^{\text{QE}} \right) \right], \tag{IA.42}$$

where $r_T^{\text{QE}} = R_T^{\text{QE}} / P_T Y_T$. We can write an equivalent expression for equation (IA.42) using the risk-neutral $\mathbb{Q}$-measure, as specified in Appendix IA.5.2.2:

$$\Delta^{\text{QE}} L_0 = \frac{\alpha}{2} \mathbb{E}_0^{\mathbb{Q}} \left[\lambda_{0,T} Y_T \left(r_T^{\text{QE}} \right)^2 \right] - \alpha \, \mathbb{E}_0^{\mathbb{Q}} \left[\lambda_{0,T} Y_T \left(\theta_T^{\text{nQE}} r_T^{\text{QE}} \right) \right], \tag{IA.43}$$

where $1/\lambda_{0,T}$ denotes the date-T real value of a strategy that rolls over one-period nominal risk-free bonds up to date T.

IA.1.7. Proof of Proposition 4

To derive the upper bound, first note that

$$\mathbb{E}_0^{\mathbb{Q}}\Big[\lambda_{0,T} Y_T \left(\theta_T^{\mathrm{nQE}}\, r_T^{\mathrm{QE}}\right)\Big] = \mathrm{Cov}_0^{\mathbb{Q}}\left[\theta_T^{\mathrm{nQE}}, \lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right] \leq \sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[\theta_T^{\mathrm{nQE}}\right] \mathrm{Var}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right]} \tag{IA.44}$$

because $\mathbb{E}_0^{\mathbb{Q}}[\lambda_{0,T} \cdot Y_T r_T^{\mathrm{QE}}] = \mathbb{E}_0^{\mathbb{Q}}[\lambda_{0,T} \cdot R_T^{\mathrm{QE}}/P_T] = 0$ according to the real Euler equation and the absolute value of a correlation is not greater than 1. Similarly,

$$\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T \left(r_T^{\mathrm{QE}}\right)^2\right] \leq \sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[r_T^{\mathrm{QE}}\right] \mathrm{Var}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right]}. \tag{IA.45}$$

Assume that $\lambda_{0,T} \leq \bar{\lambda}$. Under the assumption, we have

$$\mathrm{Var}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right] = \mathbb{E}_0^{\mathbb{Q}}\left[\left(\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right)^2\right] - \left(\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right]\right)^2 \leq (\bar{\lambda})^2 \mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]. \tag{IA.46}$$

Therefore,

$$\begin{aligned}
\Delta^{\mathrm{QE}} L_0 = {} & \frac{\alpha}{2}\, \mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T \left(r_T^{\mathrm{QE}}\right)^2\right] - \alpha\, \mathbb{E}_0^{\mathbb{Q}}\Big[\lambda_{0,T} Y_T \left(\theta_T^{\mathrm{nQE}}\, r_T^{\mathrm{QE}}\right)\Big] \\
= {} & \frac{\alpha}{2} \mathrm{Corr}_0^{\mathbb{Q}}\left[r_T^{\mathrm{QE}}, \lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right] \sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[r_T^{\mathrm{QE}}\right] \mathrm{Var}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right]} \\
& - \alpha \mathrm{Corr}_0^{\mathbb{Q}}\left[\theta_T^{\mathrm{nQE}}, \lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right] \sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[\theta_T^{\mathrm{nQE}}\right] \mathrm{Var}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right]} \\
\leq {} & \frac{\alpha\bar{\lambda}}{2} \sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[r_T^{\mathrm{QE}}\right] \mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]} + \alpha\bar{\lambda} \sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[\theta_T^{\mathrm{nQE}}\right] \mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]}.
\end{aligned} \tag{IA.47}$$

IA.2. Sign of the Output Effect

This section provides a sufficient condition under which a maturity-shortening policy (an increase in S holding D fixed) is expansionary at the ZLB.

The fundamental shock is the period-1 productivity $a_1(s)$, indexed by an exogenous scalar state s. Assume that the conditional distribution of s at date 0 is non-degenerate, and that $a_1(s)$ is continuously differentiable and strictly increasing in s, i.e. $\partial a_1(s)/\partial s > 0$.[34] The period-2 productivity a_2 is constant. We restrict attention to equilibria with $D > 0$ and $D(S - S^\star) \geq 0$, where $S^\star \equiv \frac{p_0^S}{p_0^S + p_0^L}$.

Tax smoothing in periods 1 and 2 implies

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s)B_0^L}{1 + p_1^S(s)} = D\left[-\frac{R_1^c(s)}{1 + p_1^S(s)}(S - S^\star) + \frac{1}{p_0^S + p_0^L}\right], \tag{IA.48}$$

where $R_1^c(s) \equiv \frac{p_1^S(s)}{p_0^L} - \frac{1}{p_0^S}$. Bondholders' consumption is therefore

$$c_0^b = y_0 - \frac{1-\phi}{\omega}D, \tag{IA.49}$$

$$c_1^b(s) = a_1(s) + \frac{1-\phi}{\omega}\tau_1(s), \tag{IA.50}$$

$$c_2^b(s) = a_2 + \frac{1-\phi}{\omega}\tau_2(s). \tag{IA.51}$$

Since $D > 0$ and bond prices are positive, tax smoothing implies $\tau_1(s) = \tau_2(s) > 0$, hence $c_2^b(s) > a_2$ for all s.

When the ZLB binds, p_0^S is fixed by policy (in the main text, $p_0^S = 1$). Differentiating the period-0 Euler equation as in Lemma 3 yields

$$\frac{\partial y_0}{\partial S} = \frac{1-\phi}{\omega} D\, \mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)} \frac{-R_1^c(s)}{1 + p_1^S(s)}\right]. \tag{IA.52}$$

We now give a simple sufficient condition for the expectation in (IA.52) to be strictly positive.

[34] The scalar state assumption is only used for the monotone-covariance argument below.

Condition 1 (Small maturity deviation). *Assume*

$$\frac{D(S - S^\star)}{a_2} \leq \frac{4\omega}{\gamma(1-\phi)\left(\frac{1}{p_0^S} + \frac{1}{p_0^L}\right)}. \tag{IA.53}$$

Lemma IA.1 (Monotone covariance). *Let X be a scalar random variable and let f and g be measurable functions that are both (weakly) increasing, or both (weakly) decreasing. Then $Cov(f(X), g(X)) \geq 0$, with strict inequality whenever X is non-degenerate and both $f(X)$ and $g(X)$ are non-constant.*[35]

Proposition IA.1. *Under Appendix 1, $p_1^S(s)$ and $c_1^b(s)$ are strictly increasing in s, and the ZLB output effect is strictly positive:*

$$\frac{\partial y_0}{\partial S} > 0.$$

Proof. Let $p(s) \equiv p_1^S(s)$ and $\tau(s) \equiv \tau_1(s) = \tau_2(s)$. From tax smoothing,

$$\tau(s) = \frac{B_0^S + p(s)B_0^L}{1 + p(s)}.$$

Differentiating with respect to p and using $B_0^S = \frac{SD}{p_0^S}$, $B_0^L = \frac{(1-S)D}{p_0^L}$, and $S^\star = \frac{p_0^S}{p_0^S + p_0^L}$ gives

$$\frac{\partial \tau}{\partial p}(s) = \frac{B_0^L - B_0^S}{(1 + p(s))^2} = -\frac{D(S - S^\star)}{(1 + p(s))^2}\left(\frac{1}{p_0^S} + \frac{1}{p_0^L}\right). \tag{IA.54}$$

Define

$$H(s) \equiv \frac{1-\phi}{\omega}\frac{D(S - S^\star)}{(1 + p(s))^2}\left(\frac{1}{p_0^S} + \frac{1}{p_0^L}\right) \geq 0. \tag{IA.55}$$

Then (IA.54)–(IA.51) imply

$$\frac{\partial c_1^b}{\partial s}(s) = a_1'(s) - H(s)\frac{\partial p}{\partial s}(s), \qquad \frac{\partial c_2^b}{\partial s}(s) = -H(s)\frac{\partial p}{\partial s}(s). \tag{IA.56}$$

[35] See, for example, Theorem 2.14 in Boucheron, Lugosi, and Massart (2013).

Next, the date-1 Euler equation implies

$$p(s) = \beta \left(\frac{c_2^b(s)}{c_1^b(s)} \right)^{-\gamma}. \tag{IA.57}$$

Taking logs in (IA.57), differentiating with respect to s, substituting (IA.56), and solving for $p'(s) \equiv \partial p / \partial s(s)$ yields

$$\frac{\partial p}{\partial s}(s) = \frac{\gamma\, p(s)}{c_1^b(s)} \left[1 + \gamma H(s) p(s) \frac{c_2^b(s) - c_1^b(s)}{c_1^b(s) c_2^b(s)} \right]^{-1} a_1'(s). \tag{IA.58}$$

Combining (IA.56) and (IA.58) gives

$$\frac{\partial c_1^b}{\partial s}(s) = \left(1 - \gamma H(s) \frac{p(s)}{c_2^b(s)} \right) \left[1 + \gamma H(s) p(s) \frac{c_2^b(s) - c_1^b(s)}{c_1^b(s) c_2^b(s)} \right]^{-1} a_1'(s). \tag{IA.59}$$

We claim that $1 - \gamma H(s) \frac{p(s)}{c_2^b(s)} > 0$ for all s. Since $c_2^b(s) > a_2$ and $\frac{p}{(1+p)^2} \leq \frac{1}{4}$ for all $p > 0$ (equivalently, $(p-1)^2 \geq 0$), (IA.55) implies

$$\gamma H(s) \frac{p(s)}{c_2^b(s)} < \gamma\, \frac{1 - \phi}{\omega} \frac{D(S - S^\star)}{a_2} \left(\frac{1}{p_0^S} + \frac{1}{p_0^L} \right) \frac{1}{4}. \tag{IA.60}$$

Under condition 1, the right-hand side of (IA.60) is weakly smaller than one, hence

$$1 - \gamma H(s) \frac{p(s)}{c_2^b(s)} > 0 \qquad \forall s. \tag{IA.61}$$

Moreover, the denominator in (IA.58) and (IA.59) satisfies

$$1 + \gamma H(s) p(s) \frac{c_2^b(s) - c_1^b(s)}{c_1^b(s) c_2^b(s)} = \left(1 - \gamma H(s) \frac{p(s)}{c_2^b(s)} \right) + \gamma H(s) \frac{p(s)}{c_1^b(s)} > 0,$$

where the strict inequality uses (IA.61) and $c_1^b(s) > 0$. Since $a_1'(s) > 0$, (IA.58) and (IA.59) imply that $p_1^S(s)$ and $c_1^b(s)$ are strictly increasing in s.

We now study the sign of the expectation in (IA.52). Define

$$X(s) \equiv \beta \left(\frac{c_1^b(s)}{c_0^b}\right)^{-\gamma} \big(-R_1^c(s)\big), \qquad Y(s) \equiv \left(\frac{c_1^b(s)}{c_0^b}\right)^{-1} \frac{1}{1+p_1^S(s)}.$$

Then $X(s)Y(s)$ equals the integrand in (IA.52). We claim that $\mathbb{E}_0[X(s)] = 0$. Using $\Lambda_1(s)/\Lambda_0 = \beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-\gamma}$ and $p_1^S(s) = \Lambda_2(s)/\Lambda_1(s)$,

$$\begin{aligned}\mathbb{E}_0[X(s)] &= -\mathbb{E}_0\left[\frac{\Lambda_1(s)}{\Lambda_0}\left(\frac{p_1^S(s)}{p_0^L} - \frac{1}{p_0^S}\right)\right] = -\frac{1}{p_0^L}\mathbb{E}_0\left[\frac{\Lambda_2(s)}{\Lambda_0}\right] + \frac{1}{p_0^S}\mathbb{E}_0\left[\frac{\Lambda_1(s)}{\Lambda_0}\right] \\ &= -\frac{p_0^L}{p_0^L} + \frac{p_0^S}{p_0^S} = 0,\end{aligned}$$

where the last equality uses the Euler equations $p_0^S = \mathbb{E}_0[\Lambda_1(s)/\Lambda_0]$ and $p_0^L = \mathbb{E}_0[\Lambda_2(s)/\Lambda_0]$. Therefore,

$$\mathbb{E}_0[X(s)Y(s)] = \mathrm{Cov}_0(X(s), Y(s))\,. \tag{IA.62}$$

Finally, $Y(s)$ is strictly decreasing in s because both $c_1^b(s)$ and $p_1^S(s)$ are strictly increasing. Moreover, writing $-R_1^c(s) = \frac{1}{p_0^S} - \frac{p_1^S(s)}{p_0^L}$, we have $\partial(-R_1^c(s))/\partial s = -(1/p_0^L)\,\partial p_1^S(s)/\partial s < 0$. A direct differentiation of $X(s)$ yields

$$\frac{\partial X(s)}{\partial s} = -\gamma\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)} \frac{a_1'(s)}{c_0^b}\left[1 + \gamma H(s) p_1^S(s)\frac{c_2^b(s) - c_1^b(s)}{c_1^b(s)c_2^b(s)}\right]^{-1}\left[\frac{1}{p_0^S} + \gamma H(s)\frac{p_1^S(s)}{c_2^b(s)}R_1^c(s)\right]. \tag{IA.63}$$

Since $p_1^S(s) > 0$, we have $R_1^c(s) = \frac{p_1^S(s)}{p_0^L} - \frac{1}{p_0^S} \geq -\frac{1}{p_0^S}$, and therefore, using (IA.61),

$$\frac{1}{p_0^S} + \gamma H(s)\frac{p_1^S(s)}{c_2^b(s)}R_1^c(s) \geq \frac{1}{p_0^S}\left(1 - \gamma H(s)\frac{p_1^S(s)}{c_2^b(s)}\right) > 0.$$

All remaining factors in (IA.63) are strictly positive except for the leading minus sign, hence $\partial X(s)/\partial s < 0$. Thus both $X(s)$ and $Y(s)$ are strictly decreasing in s. By Appendix IA.1, $\mathrm{Cov}_0(X(s), Y(s)) > 0$. Combining (IA.52) and (IA.62) yields $\partial y_0/\partial S > 0$. ■

IA.3. Model Extensions

IA.3.1. Dispersed Tax Incidence

In this section, we relax our assumption that households bear all tax liabilities. We assume instead the tax incidence is as follows: a proportion of ψ is collected from domestic bondholders, and $1-\psi$ is collected from households. Then $\omega\tau_t^b + (1-\omega)\tau_t^h = \tau_t$, and $\omega\tau_t^b = \psi\tau_t$.

We start from the budget constraints of all market participants. For the government, we have

$$G_0 = \tau_0 + p_0^S B_0^S + p_0^L B_0^L \tag{IA.64}$$
$$0 = \tau_1(s) - B_0^S + p_1^S(s)B_1^S(s) \tag{IA.65}$$
$$0 = \tau_2(s) - B_1^S(s) - B_0^L. \tag{IA.66}$$

For households, we have

$$(1-\omega)c_0^h = (1-\omega)y_0 - (1-\omega)\tau_0^h - (1-\omega)\xi(\tau_0^h) \tag{IA.67}$$
$$(1-\omega)c_1^h(s) = (1-\omega)y_1(s) - (1-\omega)\tau_1^h(s) - (1-\omega)\xi(\tau_1^h(s)) \tag{IA.68}$$
$$(1-\omega)c_2^h(s) = (1-\omega)y_2 - (1-\omega)\tau_2^h(s) - (1-\omega)\xi(\tau_2^h(s)). \tag{IA.69}$$

For domestic bondholders,

$$\omega c_0^b = \omega y_0 - (1-\phi)(p_0^S B_0^S + p_0^L B_0^L) - \omega\tau_0^b - \omega\xi(\tau_0^b) \tag{IA.70}$$
$$\omega c_1^b(s) = \omega y_1(s) - (1-\phi)(-B_0^S + p_1^S(s)B_1^S(s)) - \omega\tau_1^b(s) - \omega\xi(\tau_1^b(s)) \tag{IA.71}$$
$$\omega c_2^b(s) = \omega y_2 - (1-\phi)(-B_1^S(s) - B_0^L) - \omega\tau_2^b(s) - \omega\xi(\tau_2^b(s)), \tag{IA.72}$$

and for foreign bondholders

$$c_0^f = -\phi(p_0^S B_0^S + p_0^L B_0^L) \tag{IA.73}$$
$$c_1^f(s) = -\phi(-B_0^S + p_1^S(s)B_1^S(s)) \tag{IA.74}$$
$$c_2^f(s) = -\phi(-B_1^S(s) - B_0^L). \tag{IA.75}$$

We combine the budget constraints for government, households, domestic bondholders and express τ_t^b, τ_t^h as a function of τ_t:

$$c_0^b = y_0 - \frac{1-\phi}{\omega} G_0 + \frac{1-\phi-\psi}{\omega} \tau_0 - \xi \left(\frac{\psi}{\omega} \tau_0 \right) \tag{IA.76}$$

$$c_1^b(s) = y_1(s) + \frac{1-\phi-\psi}{\omega} \tau_1(s) - \xi \left(\frac{\psi}{\omega} \tau_1(s) \right) \tag{IA.77}$$

$$c_2^b(s) = y_2 + \frac{1-\phi-\psi}{\omega} \tau_2(s) - \xi \left(\frac{\psi}{\omega} \tau_2(s) \right). \tag{IA.78}$$

Under our new specification, the total deadweight loss in period t is expressed as

$$\omega \xi(\tau_t^b) + (1-\omega)\xi(\tau_t^h) = \omega \xi \left(\frac{\psi}{\omega} \tau_t \right) + (1-\omega) \xi \left(\frac{1-\psi}{1-\omega} \tau_t \right). \tag{IA.79}$$

We take $\xi(\tau) = \frac{\alpha}{2} \tau^2$ as before. Then we have

$$\omega \xi(\tau_t^b) + (1-\omega)\xi(\tau_t^h) = \frac{\alpha}{2} \left[\frac{\psi^2}{\omega} + \frac{(1-\psi)^2}{1-\omega} \right] (\tau_t)^2. \tag{IA.80}$$

Domestic bondholders' problem is the same as before, to which the solution is still characterized by Euler equations:

$$p_0^S = \beta \mathbb{E}_0 \left[\left(\frac{c_1^b(s)}{c_0^b} \right)^{-\gamma} \right] = \mathbb{E}_0 \left[\frac{\Lambda_1(s)}{\Lambda_0} \right] \tag{IA.81}$$

$$p_0^L = \beta^2 \mathbb{E}_0 \left[\left(\frac{c_2^b(s)}{c_0^b} \right)^{-\gamma} \right] = \mathbb{E}_0 \left[\frac{\Lambda_2(s)}{\Lambda_0} \right] \tag{IA.82}$$

$$p_1^S(s) = \beta \left(\frac{c_2^b(s)}{c_1^b(s)} \right)^{-\gamma} = \frac{\Lambda_2(s)}{\Lambda_1(s)}. \tag{IA.83}$$

Next, we solve the government's problem. The objective of the government is the same as before: it maximizes the sum of discounted aggregate output as a price taker. We assume no demand recession in periods 1 and 2 as before, that is, $y_2 = a_2$, and the period-1 interest rate is set exogenously such that $y_1(s) = a_1(s)$.

For the government, the period-1 problem is:

$$\max_{\{\tau_1(s),\tau_2(s),B_1^S(s)\}} \left\{ \left(a_1(s) + \frac{\Lambda_2(s)}{\Lambda_1(s)} a_2 \right) - \frac{\alpha}{2} \left[\frac{\psi^2}{\omega} + \frac{(1-\psi)^2}{1-\omega} \right] \left((\tau_1(s))^2 + \frac{\Lambda_2(s)}{\Lambda_1(s)} (\tau_2(s))^2 \right) \right\}, \tag{IA.84}$$

subject to the budget constraints:

$$0 = \tau_1(s) + B_1^S(s) p_1^S(s) - B_0^S \tag{IA.85}$$

$$0 = \tau_2(s) - B_1^S(s) - B_0^L. \tag{IA.86}$$

First-order condition is

$$\tau_1(s)(-p_1^S(s)) + \frac{\Lambda_2(s)}{\Lambda_1(s)} \tau_2(s) = 0. \tag{IA.87}$$

Then the period-1 problem solution is

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s) B_0^L}{1 + p_1^S(s)}. \tag{IA.88}$$

Define $D = p_0^S B_0^S + p_0^L B_0^L$, $S = p_0^S B_0^S / D$,

$$\tau_1(s) = \tau_2(s) = D \frac{S/p_0^S + p_1^S(s)(1-S)/p_0^L}{1 + p_1^S(s)} \tag{IA.89}$$

$$\tau_0 = G_0 - (p_0^S B_0^S + p_0^L B_0^L) = G_0 - D. \tag{IA.90}$$

The government's period-0 objective is

$$\mathbb{E}_0 \left[y_0 + \frac{\Lambda_1(s)}{\Lambda_0} a_1(s) + \frac{\Lambda_2(s)}{\Lambda_0} a_2 \right] \tag{IA.91}$$

$$-\frac{\alpha}{2}\left(\frac{\psi^2}{\omega} + \frac{(1-\psi)^2}{1-\omega}\right) \left\{ \mathbb{E}_0 \left[(\tau_0)^2 + \frac{\Lambda_1(s)}{\Lambda_0} (\tau_1(s))^2 + \frac{\Lambda_2(s)}{\Lambda_0} (\tau_2(s))^2 \right] \right\}. \tag{IA.92}$$

We ignore $a_1(s), a_2$ in the objective because they are deterministic, and substitute the expression for $\tau_0, \tau_1(s), \tau_2(s)$ as well as the Euler equation into the objective. The objective eventually takes the following form:

$$y_0 - \frac{\alpha}{2}\left(\frac{\psi^2}{\omega} + \frac{(1-\psi)^2}{1-\omega}\right)\left\{(G_0 - D)^2 + D^2\left[m\left(S - \frac{p_0^S}{p_0^S + p_0^L}\right)^2 + \frac{1}{p_0^S + p_0^L}\right]\right\}, \quad \text{(IA.93)}$$

where

$$m = \mathbb{E}_0\left[\left(\frac{R_1^c(s)}{1 + p_1^S(s)}\right)^2\left(\frac{\Lambda_1(s)}{\Lambda_0} + \frac{\Lambda_2(s)}{\Lambda_0}\right)\right]. \quad \text{(IA.94)}$$

We take the first-order condition with respect to S:

$$\frac{\partial y_0}{\partial S} - \alpha\left(\frac{\psi^2}{\omega} + \frac{(1-\psi)^2}{1-\omega}\right)D^2 m(S - S^\star) = 0. \quad \text{(IA.95)}$$

The new multiplier in marginal cost $\frac{\psi^2}{\omega} + \frac{(1-\psi)^2}{1-\omega}$ reflects the smoothness in tax distribution across agents. We next calculate the output effect $\partial y_0/\partial S$. From the domestic bondholders' Euler equation,

$$p_0^S = \beta\mathbb{E}_0\left[\left(\frac{a_1(s) + \frac{1-\phi-\psi}{\omega}\tau_1(s) - \frac{\alpha}{2}(\frac{\psi}{\omega}\tau_1(s))^2}{y_0 - \frac{1-\phi-\psi}{\omega}(G_0 - \tau_0) - \frac{\alpha}{2}(\frac{\psi}{\omega}\tau_0)^2}\right)^{-\gamma}\right], \quad \text{(IA.96)}$$

which pins down the domestic bondholders' (and therefore aggregate) demand y_0. When the ZLB binds, p_0^S is constrained at 1. We can find the output effect:

$$\begin{aligned}\frac{\partial y_0}{\partial S} &= \frac{\beta}{p_0^S}\mathbb{E}_0\left[\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)}\left[\frac{1-\phi-\psi}{\omega} - \alpha\left(\frac{\psi}{\omega}\tau_1(s)\right)\frac{\psi}{\omega}\right]\frac{\partial\tau_1(s)}{\partial S}\right] && \text{(IA.97)}\\ &= \frac{D}{p_0^S}\frac{1-\phi-\psi}{\omega}\mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)}\frac{-R_1^c(s)}{1+p_1^S(s)}\right] && \text{(IA.98)}\\ &- \frac{D}{p_0^S}\frac{\psi}{\omega}\mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)}\frac{-R_1^c(s)}{1+p_1^S(s)}\alpha\left(\frac{\psi}{\omega}\tau_1(s)\right)\right]. && \text{(IA.99)}\end{aligned}$$

Here $\frac{1-\phi-\psi}{\omega}$ is the marginal decrease in domestic bondholders' net exposure to the interest rate risk, our regular term; the second term comes from the change in deadweight loss bondholders need to pay, because part of taxes are collected from them. Since deadweight loss is monotonically increasing in taxes, it inherits the cyclicality of taxes. Therefore, when QE reduces the risk exposure of domestic bondholders' bond return by purchasing positive-beta long-term bonds from them, it also reduces the risk exposure of deadweight loss. As deadweight loss reduces consumption, the aggregate effect of QE on domestic bondholders' consumption will be smaller than that in the bond return, and the output effect will be smaller or even negative.

IA.3.2. Government Welfare Weights

In the baseline model, the government discounts future outcomes using bondholders' stochastic discount factor Λ_t. This subsection relaxes this assumption. We let the government evaluate allocations using an adapted, strictly positive process Λ_t^g, while bond prices continue to be determined by bondholders' kernel Λ_t through (10). We show how Proposition 1 and Proposition 2 are nested as the special case $\Lambda_t^g \equiv \Lambda_t$.

Period 1. Fix a realized state s at date 1 and take (B_0^S, B_0^L) as given. Since output in periods 1 and 2 is at capacity, the period-1 government problem reduces to minimizing discounted quadratic tax distortions:

$$\min_{\tau_1(s),\tau_2(s)} \tau_1(s)^2 + \frac{\Lambda_2^g(s)}{\Lambda_1^g(s)}\tau_2(s)^2 \quad \text{s.t.} \quad \tau_1(s) + p_1^S(s)\tau_2(s) = B_0^S + p_1^S(s)B_0^L. \tag{IA.100}$$

Define the wedge

$$\chi(s) \equiv \frac{\Lambda_2^g(s)/\Lambda_1^g(s)}{\Lambda_2(s)/\Lambda_1(s)} = \frac{\Lambda_2^g(s)/\Lambda_1^g(s)}{p_1^S(s)}. \tag{IA.101}$$

Solving (IA.100) yields

$$\tau_2(s) = \frac{B_0^S + p_1^S(s)B_0^L}{p_1^S(s) + \chi(s)}, \qquad \tau_1(s) = \chi(s)\,\tau_2(s). \tag{IA.102}$$

Substituting (IA.102) into the continuation tax-loss term gives the exact simplification

$$\frac{\Lambda_1^g(s)}{\Lambda_0^g}\tau_1(s)^2 + \frac{\Lambda_2^g(s)}{\Lambda_0^g}\tau_2(s)^2 = \frac{\Lambda_1^g(s)}{\Lambda_0^g} \cdot \frac{\chi(s)}{p_1^S(s) + \chi(s)} \left(B_0^S + p_1^S(s) B_0^L\right)^2. \tag{IA.103}$$

When $\Lambda_t^g \equiv \Lambda_t$, we have $\chi(s) \equiv 1$ and (IA.102) collapses to $\tau_1(s) = \tau_2(s) = (B_0^S + p_1^S(s)B_0^L)/(1 + p_1^S(s))$, as in the baseline.

Period-0 reformulation. Using the definitions from Lemma 1,

$$D = p_0^S B_0^S + p_0^L B_0^L, \qquad S = \frac{p_0^S B_0^S}{D}, \qquad S^\star = \frac{p_0^S}{p_0^S + p_0^L}, \qquad R_1^c(s) = \frac{p_1^S(s)}{p_0^L} - \frac{1}{p_0^S},$$

we can rewrite

$$B_0^S + p_1^S(s) B_0^L = D\left(\frac{1 + p_1^S(s)}{p_0^S + p_0^L} - (S - S^\star) R_1^c(s)\right). \tag{IA.104}$$

Plugging (IA.103)–(IA.104) into the period-0 objective (and using $\tau_0 = G_0 - D$) yields the following analogue of Lemma 1. Define

$$\begin{aligned}
A_0 &\equiv \mathbb{E}_0\left[\frac{\Lambda_1^g(s)}{\Lambda_0^g}\,\frac{\chi(s)}{p_1^S(s) + \chi(s)}\,\left(\frac{1 + p_1^S(s)}{p_0^S + p_0^L}\right)^2\right], \\
A_1 &\equiv \mathbb{E}_0\left[\frac{\Lambda_1^g(s)}{\Lambda_0^g}\,\frac{\chi(s)}{p_1^S(s) + \chi(s)}\,\left(\frac{1 + p_1^S(s)}{p_0^S + p_0^L}\right)\,R_1^c(s)\right], \\
A_2 &\equiv \mathbb{E}_0\left[\frac{\Lambda_1^g(s)}{\Lambda_0^g}\,\frac{\chi(s)}{p_1^S(s) + \chi(s)}\,(R_1^c(s))^2\right].
\end{aligned} \tag{IA.105}$$

Then the period-0 problem can be written as

$$\max_{S,D}\left\{y_0 - \frac{\alpha}{2(1-\omega)}\left((G_0 - D)^2 + D^2\left[A_2\,(S - S^g)^2 + \left(A_0 - \frac{A_1^2}{A_2}\right)\right]\right)\right\}, \tag{IA.106}$$

where

$$S^g \equiv S^\star + \frac{A_1}{A_2}. \tag{IA.107}$$

When $\Lambda_t^g \equiv \Lambda_t$, $\chi(s) \equiv 1$ and (IA.105) satisfies $A_1 = 0$, $A_2 = m$, and $A_0 = 1/(p_0^S + p_0^L)$, so

(IA.106) reduces to Lemma 1.

Implications outside the ZLB. If the ZLB is not binding in period 0, $y_0 = a_0$ is independent of (S, D), and (IA.106) implies $S = S^g$. The baseline full tax-smoothing maturity share $S^\star$ is recovered when $\Lambda_t^g \equiv \Lambda_t$, in which case $S^g = S^\star$.

Implications at the ZLB. When the ZLB is binding in period 0, the maturity choice affects output through $\tau_1(s)$. Using (IA.102) and (IA.104), taxes in period 1 satisfy

$$\tau_1(s) = \frac{\chi(s)}{p_1^S(s) + \chi(s)} D \left(R_1^c(s)(S^\star - S) + \frac{1 + p_1^S(s)}{p_0^S + p_0^L} \right). \tag{IA.108}$$

Differentiating the ZLB output expression as in Lemma 3 yields

$$\frac{\partial y_0}{\partial S} = \frac{1-\phi}{\omega} D\, \mathbb{E}_0 \left[\beta \left(\frac{c_1^b(s)}{c_0^b} \right)^{-(1+\gamma)} \frac{\chi(s)}{p_1^S(s) + \chi(s)} \left(-R_1^c(s)\right) \right], \tag{IA.109}$$

which coincides with (19) when $\chi(s) \equiv 1$. Taking the first-order condition with respect to S in (IA.106) gives

$$\frac{\alpha}{1-\omega} D^2 A_2 \left(S - S^g\right) = \frac{\partial y_0}{\partial S}. \tag{IA.110}$$

It is therefore natural to define the ZLB-specific maturity deviation relative to the no-ZLB optimum S^g as

$$Q_g \equiv D\left(S - S^g\right),$$

which reduces to $Q = D(S - S^\star)$ in the baseline case $\Lambda_t^g \equiv \Lambda_t$.

IA.3.3. Price-Strategic Government

In the baseline model, the government chooses its decisions (B_0^S, B_0^L) taking bond prices as given. This subsection relaxes this assumption in a nested way. We assume equilibrium prices are differentiable functions of issuance and let the government internalize a fraction $\zeta \in [0, 1]$ of the induced price changes at time-0; the government does not internalize the price changes at period-1. We recover Proposition 1 and Proposition 2 as the special case $\zeta = 0$.

Equilibrium price schedules. Assume that, in equilibrium, the date-0 bond prices satisfy

$$p_0^S = p_0^S(B_0^S, B_0^L), \qquad p_0^L = p_0^L(B_0^S, B_0^L), \tag{IA.111}$$

with all functions differentiable. The parameter ζ captures how much of the derivatives of these schedules the government internalizes when forming first-order conditions. Thus, $\zeta = 0$ corresponds to price taking, while $\zeta = 1$ corresponds to full internalization.

Period 1. Fix a realized state s at date 1 and take (B_0^S, B_0^L) as given. Since output in periods 1 and 2 is at capacity and we assume the government does not internalize the price changes at period-1, the period-1 government problem is unchanged and taxes are smoothed across periods 1 and 2:

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s) B_0^L}{1 + p_1^S(s)}. \tag{IA.112}$$

Define $w_S(p) \equiv 1/(1+p)$ and $w_L(p) \equiv p/(1+p)$.

Period 0 outside the ZLB. The date-0 budget constraint implies

$$\tau_0 = G_0 - p_0^S B_0^S - p_0^L B_0^L. \tag{IA.113}$$

Under (IA.111), issuance affects τ_0 through equilibrium proceeds. For $j \in \{S, L\}$, define the ζ-internalized marginal-proceeds price

$$\tilde{p}_{0,j} \equiv p_{0,j} + \zeta \Big(B_0^S \frac{\partial p_0^S}{\partial B_{0,j}} + B_0^L \frac{\partial p_0^L}{\partial B_{0,j}} \Big), \tag{IA.114}$$

so that the government perceives

$$\left(\frac{\partial \tau_0}{\partial B_{0,j}} \right)_\zeta = -\tilde{p}_{0,j}. \tag{IA.115}$$

In addition, if $p_1^S(s)$ depends on issuance, differentiating (IA.112) yields the exact identity

$$\frac{\partial \tau_1(s)}{\partial B_{0,j}} = w_j(p_1^S(s)) + \frac{B_0^L - B_0^S}{(1 + p_1^S(s))^2} \frac{\partial p_1^S(s)}{\partial B_{0,j}}, \tag{IA.116}$$

and the government perceives the second term in (IA.116) scaled by ζ.

Let $W(s) \equiv \Lambda_1(s)/\Lambda_0 + \Lambda_2(s)/\Lambda_0$. Using $\tau_2(s) = \tau_1(s)$, the date-0 first-order condition for $B_{0,j}$ can be written as

$$0 = \mathbb{E}_0\left[-\tau_0\,\tilde{p}_{0,j} + W(s)\,\tau_1(s)\,w_j(p_1^S(s)) + \zeta\,W(s)\,\tau_1(s)\,\frac{B_0^L - B_0^S}{(1+p_1^S(s))^2}\frac{\partial p_1^S(s)}{\partial B_{0,j}}\right], \qquad j \in \{S, L\}. \tag{IA.117}$$

When $\zeta = 0$, (IA.117) reduces to the price-taking conditions used in the proof of Proposition 1. For $\zeta > 0$, the optimal maturity reflects two additional incentives: (i) issuance is evaluated at marginal proceeds (IA.114), and (ii) the government internalizes how issuance shifts the continuation short price schedule $p_1^S(s)$ in (IA.116).

Implications at the ZLB. We use the period-0 reformulation from Lemma 1. When equilibrium prices depend on issuance, the objects m, $S^\star$, and $1/(p_0^S + p_0^L)$ may vary with S (holding D fixed). The exact first-order condition with respect to S becomes

$$\frac{\partial y_0}{\partial S} = \frac{\alpha}{2(1-\omega)}D^2\left[2m(S-S^\star) + \zeta\left\{(S-S^\star)^2\frac{\partial m}{\partial S} - 2m(S-S^\star)\frac{\partial S^\star}{\partial S} + \frac{\partial}{\partial S}\Big(\frac{1}{p_0^S + p_0^L}\Big)\right\}\right], \tag{IA.118}$$

where the derivatives are taken with respect to S holding D fixed. When $\zeta = 0$, the bracketed ζ-term drops out and (IA.118) collapses to the main-text optimality condition (31), yielding Proposition 2. For $\zeta > 0$, the optimal deviation from tax smoothing (and hence the optimal QE size) is generally characterized implicitly by (IA.118).

IA.3.4. Zero-interest Liabilities

In the main text, we have treated the consolidated government's short-maturity liabilities as a single instrument, without distinguishing whether they are issued as Treasury bills or as reserve balances. This abstraction is appropriate in operating frameworks in which reserve balances are fully remunerated at the policy rate: absent regulatory or institutional wedges, an interest-bearing reserve balance and a short Treasury security represent the same payoff risk and therefore command the same return in equilibrium.

Recent discussions of central-bank losses, however, have emphasized policies that rely on *non-interest-bearing* central bank liabilities. First, De Grauwe and Ji (2023) argue that ceasing to remunerate (infra-marginal) reserves could limit central bank operating losses when policy rates rise. Second, the existence of non-interest-bearing currency implies that the consolidated public sector earns seigniorage revenues whose present value increases when interest rates are high, providing a partial hedge against the interest-rate exposure created by QE. This logic is reflected in accounting mechanisms that smooth remittances when net income turns negative, such as the "deferred asset" treatment discussed by Faria-e-Castro and Jordan-Wood (2023).

In this section, we study how introducing zero-interest liabilities affects the design and consequences of QE. We consider two extensions. The first introduces an exogenous stock of non-interest-bearing currency outstanding. We show that currency changes the level of distortionary financing through seigniorage and alters the full tax-smoothing benchmark, but that the optimal QE rule at the ZLB retains the same functional form once expressed in terms of *effective* debt and the effective short-term share. The second extension considers a counterfactual implementation in which QE is financed by issuing reserves that pay zero interest and that bondholders are required to hold. In that case, the liability issued to fund purchases is itself long duration in our three-period setup; the maturity transformation that underlies the duration-risk-extraction channel is therefore absent, and the remaining effect operates through seigniorage redistribution.

Modified setup Assume that bondholders must hold a fixed quantity of non-interest-bearing money, M.[36] We assume that M is outstanding in periods 0 and 1 and is redeemed in period 2. The government budget constraints become

$$G_0 = \tau_0 + B_0^S p_0^S + B_0^L p_0^L + M, \tag{IA.119}$$

$$0 = \tau_1(s) + B_1^S(s) p_1^S(s) - B_0^S, \tag{IA.120}$$

$$0 = \tau_2(s) - B_1^S(s) - B_0^L - M. \tag{IA.121}$$

[36] This assumption can be interpreted either as a binding reserve requirement for banks (De Grauwe and Ji, 2023) or as a reduced-form way of capturing a transactions demand (a convenience yield) that we do not model explicitly.

Solving the government's period-1 problem delivers full tax smoothing between periods 1 and 2, taking into account that M is a long-maturity liability:

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s)\big(B_0^L + M\big)}{1 + p_1^S(s)}.$$

It is convenient to define the *effective* long-maturity liability $\widetilde{B}_0^L \equiv B_0^L + M$ and to rewrite the period-0 budget constraint as

$$\widetilde{G}_0 = \tau_0 + B_0^S p_0^S + \widetilde{B}_0^L p_0^L, \qquad \text{where} \qquad \widetilde{G}_0 \equiv G_0 - M(1 - p_0^L).$$

The term $M(1 - p_0^L)$ is the present value of seigniorage: issuing M units of a claim that is redeemed at par in period 2 provides M units of resources in period 0 and has a market value Mp_0^L in terms of the period-0 price of a claim paying in period 2. Thus, $\widetilde{G}_0$ is the amount of period-0 spending that remains to be financed after accounting for seigniorage.

QE in the presence of seigniorage We next characterize the optimal maturity tilt at the ZLB when currency is outstanding. Let $\widetilde{D}$ denote the total effective value of outstanding liabilities in period 0,

$$\widetilde{D} \equiv p_0^S B_0^S + p_0^L \widetilde{B}_0^L,$$

and let $\widetilde{S} \equiv p_0^S B_0^S / \widetilde{D}$ be the effective short-term share.

Proposition IA.2 (Optimal QE size at the ZLB with currency). *If the ZLB is binding and $\partial y_0/\partial \widetilde{S} > 0$, the optimal deviation from full tax smoothing (equivalently, the optimal size of a QE program) is*

$$\widetilde{Q}^{\star} = \min\left\{\frac{1-\omega}{\alpha m}\frac{1}{\widetilde{D}}\frac{\partial y_0}{\partial \widetilde{S}}, \overline{Q}\right\}, \tag{IA.122}$$

where $\overline{Q}$ is the minimum maturity swap that makes the ZLB constraint slack.

Proof. We plug the solution to the period-1 problem as well as the government's budget constraint into the period-0 objective, where y_1, y_2 terms are omitted because they are exogenous. The objective is then reformulated as

$$\mathbb{E}_0 \left\{ y_0 - \frac{\alpha}{2(1-\omega)} \left[(G_0 - p_0^S B_0^S - p_0^L B_0^L - M)^2 + \frac{\Lambda_1(s)}{\Lambda_0} \frac{(B_0^S + p_1^S(s)(B_0^L + M))^2}{1 + p_1^S(s)} \right] \right\}, \tag{IA.123}$$

where we have used the Euler equation $p_1^S(s) = \frac{\Lambda_2(s)}{\Lambda_1(s)}$ to simplify the expression. Define $\tilde{B}_0^L = B_0^L + M$, $\tilde{D} = p_0^S B_0^S + p_0^L \tilde{B}_0^L$, $\tilde{S} = p_0^S B_0^S / \tilde{D}$, and $\tilde{G}_0 = G_0 - M(1 - p_0^L)$. The objective becomes

$$y_0 - \frac{\alpha}{2(1-\omega)} \left[(\tilde{G}_0 - \tilde{D})^2 + \tilde{D}^2 \left(m(\tilde{S} - \frac{p_0^S}{p_0^S + p_0^L})^2 + \frac{1}{p_0^S + p_0^L} \right) \right], \tag{IA.124}$$

where

$$m = \mathbb{E}_0 \left[\frac{\Lambda_1(s)}{\Lambda_0(1 + p_1^S(s))} \left(\frac{1}{p_0^S} - \frac{p_1^S(s)}{p_0^L} \right)^2 \right] = \mathbb{E}_0 \left[\frac{\Lambda_1(s)}{\Lambda_0(1 + p_1^S(s))} \left(-R_1^c(s)\right)^2 \right]. \tag{IA.125}$$

The government's problem here is isomorphic to the problem in Lemma 1. Therefore, Appendix IA.2 follows Proposition 2. □

Appendix IA.2 shows that, after accounting for seigniorage and for the fact that currency adds to the long-maturity component of public liabilities, the optimal QE rule has the same form as in Proposition 2, with (D, S) replaced by their effective counterparts $(\widetilde{D}, \widetilde{S})$. Economically, currency affects the level and timing of distortionary taxation through seigniorage and thereby shifts the full tax-smoothing benchmark. However, once the government's problem is expressed in terms of effective liabilities, the marginal trade-off that governs the optimal maturity tilt at the ZLB remains the comparison between the output gain from maturity shortening and the fiscal-efficiency cost of additional refinancing risk.

QE financed by zero-interest reserves We finally consider a counterfactual implementation in which a QE purchase is funded not by interest-bearing short liabilities, but by issuing additional zero-interest reserves to bondholders (e.g., under a regime with non-remunerated required reserves). In contrast to the baseline case, these reserves are not arbitraged with short-term debt because their return is fixed at zero. In the three-period structure of the model, a zero-interest reserve that is redeemed in period 2 has the same payoff timing as the long bond and is therefore effectively a long-duration claim from the perspective of bondholders. As a result, swapping long-term assets against such reserves does not reduce the duration risk borne by the private sector; the remaining effect operates through the seigniorage implicit in forcing bondholders to hold an asset with a below-market return.

Proposition IA.3 (Output effect of zero-interest-reserve-financed QE). *If the ZLB is binding and $p_0^L \le 1$, the impact of funding QE with zero-interest reserves M (rather than with interest-bearing short liabilities) on output y_0 is*

$$\frac{\partial y_0}{\partial M} = \frac{1-\phi}{\omega}\,\mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(1+\gamma)}\frac{-R_1^m(s)}{1+p_1^S(s)}\right] \le 0, \tag{IA.126}$$

where $R_1^m(s) \equiv \frac{p_1^S(s)}{p_0^L} - p_1^S(s) \ge 0$ is the period-1 value of a "borrow-reserves, invest-long" position.

Proof. Since the problem with currency is isomorphic to the problem without currency, we have $\tau_1(s) = \frac{B_0^S + p_1^S(s)(B_0^L+M)}{1+p_1^S(s)}$, and $\tau_0 = G_0 - D$. Since QE is financed by currency, we have $p_0^L B_0^L + M = D - p_0^S B_0^S$, where D and B_0^S are held constant. When the ZLB is binding, we get

$$y_0 = \frac{1-\phi}{\omega}D + \left(\beta\mathbb{E}_0\left[\left(a_1(s) + \frac{1-\phi}{\omega}\frac{B_0^S + p_1^S(s)\left[\frac{D-p_0^S B_0^S}{p_0^L} + M\left(1-\frac{1}{p_0^L}\right)\right]}{1+p_1^S(s)}\right)^{-\gamma}\right]\right)^{-\frac{1}{\gamma}} \tag{IA.127}$$

and output effect of QE financed by currency issuance is given by

$$\frac{\partial y_0}{\partial M} = \frac{1-\phi}{\omega}\mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)}\frac{p_1^S(s)}{1+p_1^S(s)}\left(1-\frac{1}{p_0^L}\right)\right] < 0. \tag{IA.128}$$

□

Two features deliver the sign restriction in Appendix IA.3. First, because zero-interest reserves are redeemed only in period 2, they have the same maturity as the long bond in the model. A QE swap funded by such reserves therefore does not change bondholders' exposure to period-1 interest rate risk, and the duration-risk-extraction channel that raises output in the baseline model is absent. Second, issuing zero-interest reserves at par when the market price of a period-2 claim is $p_0^L \leq 1$ generates seigniorage: relative to market funding, the consolidated government captures the wedge $1 - p_0^L$, which is transferred to households through lower required taxes. This redistribution lowers bondholders' expected future resources and increases their desired saving, reducing current demand and, hence, output.

If instead the seigniorage revenue were rebated to bondholders through non-distortionary transfers, then zero-interest-reserve-financed QE would be neutral in this framework: it would exchange two claims with the same payoff timing and the same interest-rate risk exposure, without changing risk allocation or net resources of the marginal saver.[37]

IA.3.5. Mortgage-Backed Securities

This section extends the three-period model in Section I to explicitly incorporate central-bank purchases of mortgage-backed securities (MBS). We model agency MBS as claims on a stock of long-term household mortgage loans that are default-free.[38] Under these assumptions, an MBS is a risk-free nominal claim to a unit payoff in period 2. In our three-period structure, this payoff timing coincides with that of the long bond issued by the government. As a result, MBS are perfectly spanned by long-term Treasuries, and purchases of MBS are fungible with purchases of long-term government bonds. This stylized mapping abstracts from the prepayment-driven negative convexity of actual agency MBS, whose implications for measured interest-rate exposure we address empirically in Appendix IA.5.1.

Modified setup Let B^M denote the aggregate face value of outstanding mortgage loans that mature in period 2. Households are obligated to repay B^M in period 2, and the cor-

[37]Formally, neutrality obtains if the government taxes households in periods 1 and 2 by $MR_1^m(s)/(1 + p_1^S(s))$ and rebates the proceeds to bondholders, and if these transfers do not generate additional deadweight losses.

[38]As in the data, we take the outstanding stock of mortgages as given at date 0 and abstract from the origination decision.

responding claims are initially held by bondholders. Let $B_0^{M,g}$ denote the quantity of these mortgage claims purchased by the government (equivalently, by the central bank) in period 0 and held to maturity. The government pays $p_0^M B_0^{M,g}$ in period 0 and receives $B_0^{M,g}$ in period 2. The government's budget constraints become

$$G_0 + p_0^M B_0^{M,g} = \tau_0 + p_0^S B_0^S + p_0^L B_0^L, \tag{IA.129}$$

$$0 = \tau_1(s) + p_1^S(s) B_1^S(s) - B_0^S, \tag{IA.130}$$

$$0 = \tau_2(s) + B_0^{M,g} - B_1^S(s) - B_0^L. \tag{IA.131}$$

Hand-to-mouth households now make mortgage repayments in period 2 in addition to paying taxes:

$$c_0^h = y_0 - \frac{\tau_0}{1-\omega} - \frac{\alpha}{2}\left(\frac{\tau_0}{1-\omega}\right)^2, \tag{IA.132}$$

$$c_1^h(s) = y_1(s) - \frac{\tau_1(s)}{1-\omega} - \frac{\alpha}{2}\left(\frac{\tau_1(s)}{1-\omega}\right)^2, \tag{IA.133}$$

$$c_2^h(s) = y_2 - \frac{\tau_2(s)}{1-\omega} - \frac{\alpha}{2}\left(\frac{\tau_2(s)}{1-\omega}\right)^2 - \frac{B^M}{1-\omega}. \tag{IA.134}$$

Mortgage repayments are pure transfers from households to the holders of mortgage claims and therefore do not affect the aggregate resource constraint.

Pricing Because mortgages are default-free, a unit of MBS is a risk-free claim to one unit of nominal payoff in period 2. Domestic bondholders therefore price MBS using the same stochastic discount factor as in Section I. In particular, the date-0 MBS price satisfies

$$p_0^M = \mathbb{E}_0\left[\frac{\Lambda_2(s)}{\Lambda_0}\right], \tag{IA.135}$$

which coincides with the long-bond pricing equation. Hence,

$$p_0^M = p_0^L. \tag{IA.136}$$

Similarly, conditional on the realization of s in period 1, the MBS is a one-period risk-free claim, so its date-1 price is $p_1^M(s) = p_1^S(s)$.

Reduction to effective long-maturity liabilities. Substituting (IA.136) into (IA.129)–(IA.131) shows that MBS holdings enter the consolidated public budget exactly like a reduction in long-maturity government debt. It is therefore convenient to define the effective long-maturity public position

$$\widetilde{B}_0^L \equiv B_0^L - B_0^{M,g}. \tag{IA.137}$$

Using (IA.137), the government's budget constraints can be rewritten as

$$G_0 = \tau_0 + p_0^S B_0^S + p_0^L \widetilde{B}_0^L, \tag{IA.138}$$

$$0 = \tau_1(s) + p_1^S(s) B_1^S(s) - B_0^S, \tag{IA.139}$$

$$0 = \tau_2(s) - B_1^S(s) - \widetilde{B}_0^L. \tag{IA.140}$$

These constraints are identical to the baseline constraints in Section I, with B_0^L replaced by $\widetilde{B}_0^L$.

Proposition IA.4 (Fungibility MBS and long-term Treasuries). *Absent default, $p_0^M = p_0^L$, and the consolidated public budget depends on MBS holdings only through the effective long-maturity position $\widetilde{B}_0^L = B_0^L - B_0^{M,g}$. Hence, any two policies that deliver the same pair $(B_0^S, \widetilde{B}_0^L)$ implement the same equilibrium allocation. In particular, a unit of MBS purchases is one-for-one fungible with a unit purchase of long-term government bonds.*

In particular, solving the government's period-1 problem yields full tax smoothing between periods 1 and 2:

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s)\widetilde{B}_0^L}{1 + p_1^S(s)}, \tag{IA.141}$$

and the associated short-bond issuance rule is

$$B_1^S(s) = \frac{B_0^S - \widetilde{B}_0^L}{1 + p_1^S(s)}. \tag{IA.142}$$

Effective debt variables Given $\widetilde{B}_0^L$, we define effective debt-value and short-term-share variables by

$$\widetilde{D} \equiv p_0^S B_0^S + p_0^L \widetilde{B}_0^L, \qquad \widetilde{S} \equiv \frac{p_0^S B_0^S}{\widetilde{D}}, \tag{IA.143}$$

and keep $S^\star \equiv p_0^S/(p_0^S + p_0^L)$ as in Section I.

MBS purchases and QE size In the baseline model, the maturity tilt can be measured by the deviation of the short-term share S from its full tax-smoothing benchmark $S^\star$. With MBS, the relevant long-maturity object is the effective position $\widetilde{B}_0^L$. Under full tax smoothing, $\widetilde{B}_0^L = B_0^S$, so a useful face-value measure of maturity shortening is

$$\Delta B_0 \equiv B_0^S - \widetilde{B}_0^L = (B_0^S - B_0^L) + B_0^{M,g}. \tag{IA.144}$$

Equation (IA.144) shows that a unit increase in MBS purchases raises maturity shortening one-for-one, just like a unit purchase of long-term government bonds.

Because the paper measures QE size in market-value terms, we define

$$\widetilde{Q} \equiv (\widetilde{S} - S^\star)\widetilde{D}. \tag{IA.145}$$

Using (IA.143), one obtains

$$\widetilde{Q} = \frac{p_0^S p_0^L}{p_0^S + p_0^L}\, \Delta B_0. \tag{IA.146}$$

Hence, up to the constant factor $p_0^S p_0^L/(p_0^S + p_0^L)$, QE size is proportional to the effective maturity gap ΔB_0.

Recasting the baseline propositions With $\widetilde{D}$ and $\widetilde{S}$ defined in (IA.143), the government's period-0 problem admits exactly the same reformulation as in Lemma 1, with (D, S) replaced by $(\widetilde{D}, \widetilde{S})$ and with the same refinancing-risk term $m = \mathbb{E}_0\left[\frac{\Lambda_1(s)}{\Lambda_0(1+p_1^S(s))}(R_1^c(s))^2\right]$.

Proposition IA.5 (Optimal QE with MBS). *With $\widetilde{D}$ and $\widetilde{S}$ defined in (IA.143), the optimal maturity policy is characterized exactly as in Propositions 1 and 2 after replacing*

(D, S) by $(\widetilde{D}, \widetilde{S})$. In particular, if the ZLB is slack in period 0 the government sets $\widetilde{S} = S^\star$ (equivalently, $\widetilde{Q} = 0$). If the ZLB binds and $\partial y_0/\partial \widetilde{S} > 0$, the optimal QE size is

$$\widetilde{Q}^\star = \min\left\{\frac{1-\omega}{\alpha m}\frac{1}{\widetilde{D}}\frac{\partial y_0}{\partial \widetilde{S}},\ \bar{Q}\right\}, \tag{IA.147}$$

where $\bar{Q}$ denotes the minimum maturity swap that makes the ZLB constraint slack.

Implications for empirical measurement In the minimal model of this section, an agency MBS is a risk-free nominal claim with deterministic promised cash flows. Competitive pricing therefore implies that its date-0 price equals the price of the Treasury portfolio that replicates its promised payments, so mapping MBS into the Treasury cash-flow representation is exact once one matches the payment schedule.

In the empirical implementation, security-level promised cash flows are not available for SOMA MBS holdings. Following the main text, one can proxy the MBS cash-flow schedule using the maturity distribution of Treasury and agency debt holdings in aggregate. Over our sample, standard duration measures indicate that agency MBS exhibit lower interest-rate sensitivity than Treasuries, so this proxy tends to assign MBS too much long-duration exposure. As a result, the associated interest-rate-risk component of the fiscal-efficiency cost estimate is conservative in the sense that it loads the consolidated balance sheet with more duration risk than is suggested by the duration of MBS in the data, subject to the negative-convexity caveat documented in Appendix IA.5.1: negative convexity extends duration as rates rise, which may concentrate losses in the high-rate states where tax distortions are steepest.

IA.3.6. QE Episodes in the U.S.

Figure IA.1 summarizes the evolution of the Fed's bond portfolio size in 10-year duration equivalents over the last two decades. The 10-year duration equivalent is computed as a maturity-weighted sum of promised payments (principal plus coupons), capturing the portfolio's aggregate interest-rate risk exposure.[39] The blue area depicts the 10-year duration

[39] That is, a 10-year duration equivalent is computed as $\sum_{t=1}^{30} (t/10)\, Pay_t$, where Pay_t is the principal and coupon payment accrued in t years.

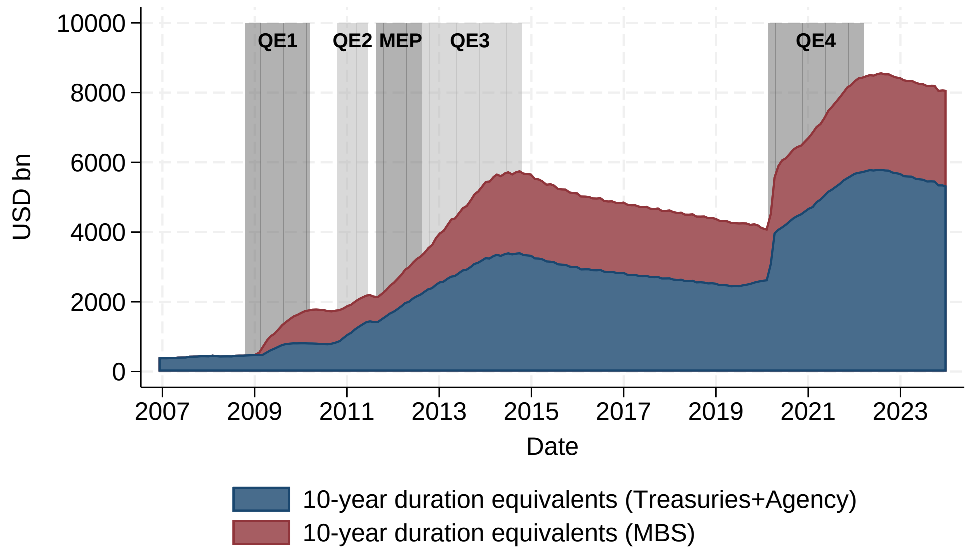


Figure IA.1: Size of the Federal Reserve's bond portfolio. This figure plots the evolution of the size of the Federal Reserve's bond portfolio measured in 10-year duration equivalents. The blue area depicts the 10-year duration equivalent values of Treasury and federal agency debt holdings, and the red area the 10-year duration equivalent values of MBS holdings. Shaded areas denote the periods of each specific asset purchase program. Section II.C details how we conservatively map MBS holdings into duration equivalents given limited information on prepayment and principal-paydown timing in SOMA releases.

equivalent values of Treasury and federal agency debt holdings, and the red area the 10-year duration equivalent values of MBS holdings. From the start of QE1 in November 2008 to the end of QE4 in March 2022, the 10-year duration equivalent of the Fed's bond portfolio increased by a factor of 17, from $0.5 trillion to $8.5 trillion.

QE1 (November 2008–March 2010) consisted of purchases of roughly $175 billion in federal agency debt, $1.25 trillion in agency MBS, and $300 billion in long-term Treasury securities. QE2 (November 2010–June 2011) consisted of $600 billion in long-term Treasury purchases. The Maturity Extension Program (MEP; September 2011–December 2012) extended the maturity of the Fed's Treasury holdings by purchasing longer-maturity Treasuries and offsetting these purchases via sales and redemptions of shorter-maturity securities totaling $667 billion.[40] QE3 (September 2012–October 2014) was an open-ended program that ultimately purchased $790 billion in Treasury securities and $823 billion in agency MBS. QE4 (March

[40] In the quantitative exercise below, we treat the MEP purchase phase as ending in August 2012 since QE3 begins in September 2012.

2020–March 2022) involved cumulative purchases of about \$4.6 trillion in long-term Treasury securities and agency MBS.

IA.4. Accounting Decomposition of the Fiscal Cost

Our quantitative exercise in Equation (23)–Proposition 4 computes the expected efficiency cost of financing QE-induced fiscal gains or losses through distortionary taxation. The object is deliberately *accounting-based*: we compare two fiscal paths—"with QE" and "without QE"—defined on the same underlying sequence of macro-financial states (e.g., bond prices, nominal output, and spending). In particular, Equation (23) is not intended to be a full general-equilibrium welfare comparison between two distinct equilibria with and without QE. Rather, it isolates how the government's tax distortions would change if the consolidated public sector were to take the QE position described in Equation (25), holding fixed other fiscal-policy rules and the state-contingent path of aggregate quantities.

This section records the consolidated budget identities that clarify (i) which fiscal channel is isolated by Equation (28), and (ii) which additional general-equilibrium channels would enter once nominal activity, primary deficits, or Treasury debt management are allowed to react endogenously. The purpose is to provide an accounting foundation for the interpretation of the fiscal-cost calculations, not to introduce a competing model of policy behavior.

Scenarios and a difference operator For any time-t object x_t, we write x_t^{QE} for its value in the QE counterfactual and x_t^{nQE} for its value absent QE, and define

$$\Delta^{\mathrm{QE}} x_t \equiv x_t^{\mathrm{QE}} - x_t^{\mathrm{nQE}}. \tag{IA.148}$$

From nominal tax revenues to tax rates Deadweight losses in Equation (23) depend on the tax take relative to the contemporaneous tax base, $\tau_t/(P_t Y_t)$. An exact identity links the QE–no-QE change in this tax rate to (i) changes in nominal revenues and (ii) changes in nominal output:

$$\Delta^{\mathrm{QE}}\left(\frac{\tau_T}{P_TY_T}\right)=\frac{1}{P_T^{\mathrm{QE}}Y_T^{\mathrm{QE}}}\left(\Delta^{\mathrm{QE}}\tau_T-\frac{\tau_T^{\mathrm{nQE}}}{P_T^{\mathrm{nQE}}Y_T^{\mathrm{nQE}}}\,\Delta^{\mathrm{QE}}(P_TY_T)\right). \tag{IA.149}$$

The second term inside parentheses isolates the *tax-base channel*: if QE raises nominal output at T, then a given change in nominal revenues corresponds to a smaller change in the tax rate. In the empirical implementation underlying Proposition 4, the path of (P_tY_t) is held fixed across counterfactuals, so $\Delta^{\mathrm{QE}}(P_TY_T)=0$ in (IA.149).

A flow identity for the Treasury at the unwind date Fix an unwind date T as in Equation (25). Let $B_t^{\mathrm{TSY},s}(n)\geq 0$ denote the nominal face value of zero-coupon Treasury debt outstanding at the end of period t with remaining maturity $n\in\{1,\dots,N\}$ in scenario $s\in\{\mathrm{QE},\mathrm{nQE}\}$. Let $p_t^s(n)$ be the time-t price of a claim to \$1 at $t+n$. Define Treasury issuance at T as

$$I_T^{\mathrm{TSY},s}(n)\equiv B_T^{\mathrm{TSY},s}(n)-B_{T-1}^{\mathrm{TSY},s}(n+1),\qquad n=1,\dots,N, \tag{IA.150}$$

with $B_{T-1}^{\mathrm{TSY},s}(N+1)\equiv 0$. Then the (nominal) consolidated flow identity at date T can be written as

$$\tau_T^s=P_T^sG_T^s+B_{T-1}^{\mathrm{TSY},s}(1)-\sum_{n=1}^{N}p_T^s(n)\,I_T^{\mathrm{TSY},s}(n)-R_T^s. \tag{IA.151}$$

Here $P_T^sG_T^s$ is nominal primary spending and R_T^s denotes net transfers from the central bank to the Treasury at T (remittances may be negative).[41]

Lemma IA.2 (Terminal tax decomposition). *Under* (IA.151)*, the QE–no-QE change*

[41] We record the accounting at a single date because the quantification in Proposition 4 maps the cumulative QE cash flow into a single fiscal adjustment date. More generally, one can choose T after the QE position has been unwound and any deferred transfers have been settled, and interpret R_T as the cumulative transfer associated with the program.

in nominal tax revenues at date T admits the exact decomposition

$$\Delta^{\mathrm{QE}}\tau_T = \Delta^{\mathrm{QE}}(P_T G_T) + \Delta^{\mathrm{QE}} B_{T-1}^{\mathrm{TSY}}(1) - \sum_{n=1}^{N} \Delta^{\mathrm{QE}} p_T(n)\, I_T^{\mathrm{TSY,nQE}}(n) - \sum_{n=1}^{N} p_T^{\mathrm{QE}}(n)\, \Delta^{\mathrm{QE}} I_T^{\mathrm{TSY}}(n) - \Delta^{\mathrm{QE}} R_T, \tag{IA.152}$$

where $\Delta^{\mathrm{QE}} p_T(n) \equiv p_T^{\mathrm{QE}}(n) - p_T^{\mathrm{nQE}}(n)$ and $\Delta^{\mathrm{QE}} I_T^{\mathrm{TSY}}(n) \equiv I_T^{\mathrm{TSY,QE}}(n) - I_T^{\mathrm{TSY,nQE}}(n)$.

If, in addition, the terminal maturity profile of Treasury debt is equalized across scenarios, $B_T^{\mathrm{TSY,QE}}(n) = B_T^{\mathrm{TSY,nQE}}(n)$ for all n, then (IA.152) simplifies to

$$\Delta^{\mathrm{QE}}\tau_T = \Delta^{\mathrm{QE}}(P_T G_T) + \Delta^{\mathrm{QE}} B_{T-1}^{\mathrm{TSY}}(1) - \sum_{n=1}^{N} \Delta^{\mathrm{QE}} p_T(n)\, I_T^{\mathrm{TSY,nQE}}(n) + \sum_{n=1}^{N-1} p_T^{\mathrm{QE}}(n)\, \Delta^{\mathrm{QE}} B_{T-1}^{\mathrm{TSY}}(n+1) - \Delta^{\mathrm{QE}} R_T. \tag{IA.153}$$

Proof. Subtract (IA.151) under $s = \mathrm{nQE}$ from (IA.151) under $s = \mathrm{QE}$. For each n, add and subtract $p_T^{\mathrm{QE}}(n) I_T^{\mathrm{TSY,nQE}}(n)$ to obtain

$$p_T^{\mathrm{QE}}(n) I_T^{\mathrm{TSY,QE}}(n) - p_T^{\mathrm{nQE}}(n) I_T^{\mathrm{TSY,nQE}}(n) = \Delta^{\mathrm{QE}} p_T(n)\, I_T^{\mathrm{TSY,nQE}}(n) + p_T^{\mathrm{QE}}(n)\, \Delta^{\mathrm{QE}} I_T^{\mathrm{TSY}}(n),$$

which gives (IA.152). If $\Delta^{\mathrm{QE}} B_T^{\mathrm{TSY}}(n) = 0$ for all n, then (IA.150) implies $\Delta^{\mathrm{QE}} I_T^{\mathrm{TSY}}(n) = -\Delta^{\mathrm{QE}} B_{T-1}^{\mathrm{TSY}}(n+1)$ for $n \le N-1$ and $\Delta^{\mathrm{QE}} I_T^{\mathrm{TSY}}(N) = 0$. Substituting into (IA.152) yields (IA.153). □

Interpreting the terms The decomposition in Lemma IA.2 isolates four accounting channels through which QE can affect the taxes required to satisfy the government's budget at T.

(i) Primary-balance channel. The term $\Delta^{\mathrm{QE}}(P_T G_T)$ captures changes in nominal primary spending (or, more generally, in the primary deficit once one augments (IA.151) to allow for non-tax revenues). This term subsumes automatic stabilizers and any discretionary fiscal responses that accompany QE-induced macroeconomic changes.

(ii) Valuation channel. The term $-\sum_n \Delta^{\text{QE}} p_T(n)\, I_T^{\text{TSY,nQE}}(n)$ captures the mechanical effect of changes in bond prices at T on the market value of a given issuance plan. Higher bond prices (lower yields) raise issuance proceeds and therefore reduce required taxes, consistent with the minus sign.

(iii) Debt-management channel. The term $-\sum_n p_T^{\text{QE}}(n)\, \Delta^{\text{QE}} I_T^{\text{TSY}}(n)$ captures active changes in issuance quantities and maturity composition. Under terminal equalization, this term reduces to the portfolio-offset term in (IA.153), which depends only on the legacy maturity vector inherited at $T-1$.

(iv) Central-bank transfer channel. The term $-\Delta^{\text{QE}} R_T$ captures how QE changes the net transfer between the central bank and the Treasury. In the consolidated-government representation used in Proposition 4, we focus on the component of this transfer driven by interest-rate risk, which is the unwind return on the QE position, R_T^{QE} in Equation (25).

Link to the fiscal-cost estimand The cost measure in Proposition 4 isolates the last channel by construction. Specifically, we hold fixed the primary fiscal stance and Treasury debt-management policy across counterfactuals (so $\Delta^{\text{QE}}(P_T G_T) = 0$ and $\Delta^{\text{QE}} I_T^{\text{TSY}}(n) = 0$), and we evaluate both scenarios along the same realized path of bond prices (so $\Delta^{\text{QE}} p_T(n) = 0$). Under these restrictions, (IA.152) reduces to $\Delta^{\text{QE}} \tau_T = -\Delta^{\text{QE}} R_T$. With the consolidated-government interpretation in Equation (26), $\Delta^{\text{QE}} R_T = R_T^{\text{QE}}$ and therefore $\Delta^{\text{QE}} \tau_T = -R_T^{\text{QE}}$. The remaining terms in Lemma IA.2 describe general-equilibrium fiscal feedbacks—through the tax base, primary deficits, debt management, and bond prices—that are intentionally held fixed when constructing the accounting-based cost measure (28).

IA.5. Measurement and Calibration

This appendix documents the measurement and calibration assumptions presented in Section II.C.

IA.5.1. Measuring the Interest-Rate Exposure of Agency MBS

For Treasury and agency-debt holdings, the SOMA releases report coupon rates and maturity dates, which allow us to construct dated nominal cash flows directly. For agency MBS, however, the SOMA releases do not report enough information on scheduled principal repayment dates and prepayment behavior to recover the timing of cash flows security by security. In the baseline implementation, we therefore approximate the maturity structure of agency MBS cash flows using the aggregate maturity profile of Treasury and agency-debt holdings.

Two features of agency MBS cut in opposite directions for the interest-rate exposure this proxy assigns to the incremental QE portfolio. The Treasury-based proxy tends to overstate that exposure, whereas the negative convexity of agency MBS extends their duration as rates rise (Hanson, 2014) and may concentrate losses in the high-rate states where tax distortions are steepest. Three pieces of evidence indicate that the overstatement is the dominant force over the observed sample range and discipline the extent of the countervailing effect. First, relative to actual agency MBS cash flows, the Treasury-based proxy implies a more back-loaded payment structure, because MBS principal is amortized over time and is additionally affected by prepayments. This tends to raise the effective duration assigned to MBS relative to their realized cash-flow profile. Second, broad Treasury indices exhibit higher modified duration than broad agency MBS indices in the data. Third, this ordering continues to hold on an option-adjusted basis: although MBS duration rises in high-rate environments because prepayments slow, Bloomberg option-adjusted duration series indicate that broad agency MBS remain below comparable Treasury portfolios even during large rate-move episodes.

Figure IA.2 illustrates this duration ordering using Bloomberg duration indices. The figure plots modified duration for U.S. Treasuries and both modified duration and option-adjusted duration for broad U.S. agency MBS indices. Over the available sample, Treasury duration remains above agency MBS duration on both measures, including during the sharp rate increases surrounding the post-COVID tightening cycle. Since our proxy assigns MBS the Treasury/agency-debt maturity profile, the net effect over the observed sample range is to overstate rather than understate the interest-rate exposure of the incremental QE portfolio, with the option-adjusted series disciplining the extent to which negative convexity pulls in the other direction.

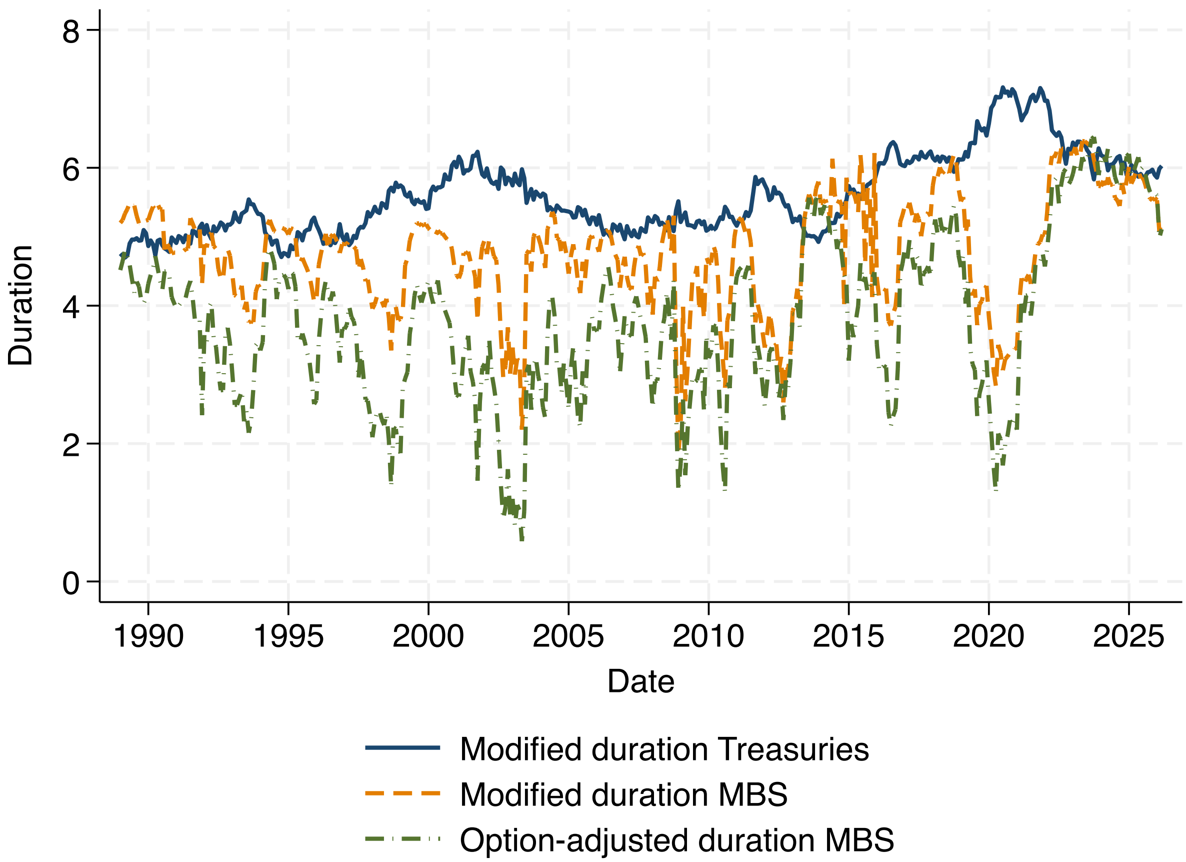


Figure IA.2: Modified and option-adjusted duration: Treasuries and MBS. The data are end-of-month series from Bloomberg US duration indices: LUATTRUU for US Treasuries, LUMSSTAT for US agency MBS modified duration, and LUMSSTAT for US agency MBS option-adjusted duration (OAD). The sample is from January 1989 (first month with non-missing OAD MBS) through February 2026.

IA.5.2. Term Structure Model

Our term-structure model follows the standard setting as in Duffie and Kan (1996); Dai and Singleton (2000), and we adopt the methodology in Hamilton and Wu (2012b).

IA.5.2.1. Basic Framework

Consider a $(L\times 1)$ vector of (latent) factors that are Markov and have dynamics characterized by

$$\mathbf{F}_{t+1}=\mathbf{c}+\rho\mathbf{F}_t+\mathbf{\Sigma}\mathbf{u}_{t+1}, \tag{IA.154}$$

where $u_t \sim$ i.i.d. $N(0, I_L)$. This implies that $\mathbf{F}_{t+1}|\mathbf{F}_t \sim N(\mathbf{c}+\rho\mathbf{F}_t, \mathbf{\Sigma}\mathbf{\Sigma}')$. Further assume that there exists a (nominal) pricing kernel $M_{t,t+1}$ that prices all possible cashflows and takes the following functional form:

$$M_{t,t+1}=\exp\left(-r_t-\frac{1}{2}\varsigma_t'\varsigma_t-\varsigma_t'\mathbf{u}_{t+1}\right), \tag{IA.155}$$

where r_t is the risk-free one-period interest rate, and ς_t is the market price of risk. Both are assumed to be an affine function of $\mathbf{F}_t$ only and are specified as follows:

$$r_t=\delta_0+\delta_1'\mathbf{F}_t, \tag{IA.156}$$

$$\varsigma_t=\varsigma+\Phi\mathbf{F}_t. \tag{IA.157}$$

IA.5.2.2. Asset Pricing Implication

Consider an asset that has payoff X_{t+1} at time $t+1$, where the payoff is possibly dependent on information up to t, $\mathcal{F}_t$, factor shocks, $\mathbf{u}_{t+1}$, and other shocks, $\mathbf{v}_{t+1}$. That is, $X_{t+1}=X(\mathbf{u}_{t+1},\mathbf{v}_{t+1};\mathcal{F}_t)$. Then its price at t can be calculated according to

$$P_t^X=\mathbb{E}_t[M_{t,t+1}X_{t+1}]=\int M_{t,t+1}X_{t+1}f_t(\mathbf{u}_{t+1},\mathbf{v}_{t+1})d\mathbf{u}_{t+1}d\mathbf{v}_{t+1}, \tag{IA.158}$$

where $f_t(\mathbf{u}_{t+1},\mathbf{v}_{t+1})$ is the probability density function of $\mathbf{u}_{t+1},\mathbf{v}_{t+1}$ conditional on $\mathcal{F}_t$. Next, we define the probability density function under the equivalent $\mathbb{Q}$-measure in this specific environment.[42] Define $f_t^{\mathbb{Q}}(\mathbf{u}_{t+1},\mathbf{v}_{t+1})\equiv\frac{M_{t,t+1}}{\mathbb{E}_t[M_{t,t+1}]}f_t(\mathbf{u}_{t+1},\mathbf{v}_{t+1})$. We can confirm that

[42] Formally, the equivalent $\mathbb{Q}$-measure is defined using the Radon-Nikodym derivative $\frac{d\mathbb{Q}}{d\mathbb{P}}$.

$f_t^{\mathbb{Q}}(\mathbf{u}_{t+1}, \mathbf{v}_{t+1})$ is a probability density function. Then

$$
\begin{aligned}
P_t^X &= \int M_{t,t+1} X_{t+1} f_t(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) d\mathbf{u}_{t+1} d\mathbf{v}_{t+1} \\
&= \int \mathbb{E}_t[M_{t,t+1}] X_{t+1} f_t^{\mathbb{Q}}(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) d\mathbf{u}_{t+1} d\mathbf{v}_{t+1} \triangleq \mathbb{E}_t^{\mathbb{Q}}[\mathbb{E}_t[M_{t,t+1}] X_{t+1}]. \qquad \text{(IA.159)}
\end{aligned}
$$

Since $M_{t,t+1}$ is pinned down by past information $\mathcal{F}_t$ and current factor shocks $\mathbf{u}_{t+1}$, we can integrate over $\mathbf{v}_{t+1}$ to derive the marginal probability density function under the $\mathbb{Q}$-measure for $\mathbf{u}_{t+1}$: $f_t^{\mathbb{Q}}(\mathbf{u}_{t+1}) = \int_{\mathbf{v}_{t+1}} \frac{M_{t,t+1}}{\mathbb{E}_t[M_{t,t+1}]} f_t(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) d\mathbf{v}_{t+1} = \frac{M_{t,t+1}}{\mathbb{E}_t[M_{t,t+1}]} f_t(\mathbf{u}_{t+1})$. We further plug in the probability density function for $\mathbf{u}_{t+1}$:

$$
\begin{aligned}
f_t^{\mathbb{Q}}(\mathbf{u}_{t+1}) &= \frac{M_{t,t+1}}{\mathbb{E}_t[M_{t,t+1}]} f_t(\mathbf{u}_{t+1}) && \text{(IA.160)} \\
&= \frac{1}{(2\pi)^{\frac{L}{2}}} \exp(-\frac{1}{2}\varsigma_t'\varsigma_t - \varsigma_t'\mathbf{u}_{t+1}) \exp\left(-\frac{1}{2}\mathbf{u}_{t+1}'\mathbf{u}_{t+1}\right) && \text{(IA.161)} \\
&= \frac{1}{(2\pi)^{\frac{L}{2}}} \exp\left(-\frac{1}{2}(\mathbf{u}_{t+1} + \varsigma_t)'(\mathbf{u}_{t+1} + \varsigma_t)\right). && \text{(IA.162)}
\end{aligned}
$$

Therefore, $\mathbf{u}_{t+1} \sim N(-\varsigma_t, I_L)$ under the $\mathbb{Q}$-measure, so the factor dynamics under the $\mathbb{Q}$-measure is $\mathbf{F}_{t+1} = \mathbf{c} + \rho\mathbf{F}_t + \mathbf{\Sigma}\mathbf{u}_{t+1}$, where $\mathbf{u}_{t+1} \sim N(-\varsigma_t, I_L)$. Equivalently,

$$
\mathbf{F}_{t+1} = \mathbf{c}^{\mathbb{Q}} + \rho^{\mathbb{Q}}\mathbf{F}_t + \mathbf{\Sigma} u_{t+1}^{\mathbb{Q}}, \qquad \text{(IA.163)}
$$

where $\mathbf{c}^{\mathbb{Q}} = c - \mathbf{\Sigma}\varsigma$, $\rho^{\mathbb{Q}} = \rho - \mathbf{\Sigma}\Phi$, and $u_{t+1}^{\mathbb{Q}} \sim N(0, I_L)$.

Any other shock $\mathbf{v}_{t+1}$ that is independent of the factor shocks—in particular the idiosyncratic macro disturbances of Appendix IA.5.3—retains its physical distribution under $\mathbb{Q}$. Because $M_{t,t+1}$ depends on the factor shocks $\mathbf{u}_{t+1}$ alone, integrating the tilted joint density over $\mathbf{u}_{t+1}$ gives $f_t^{\mathbb{Q}}(\mathbf{v}_{t+1}) = \int_{\mathbf{u}_{t+1}} \frac{M_{t,t+1}}{\mathbb{E}_t[M_{t,t+1}]} f_t(\mathbf{u}_{t+1}, \mathbf{v}_{t+1})\, d\mathbf{u}_{t+1}$; when $\mathbf{v}_{t+1}$ is independent of $\mathbf{u}_{t+1}$, so that $f_t(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) = f_t(\mathbf{u}_{t+1})\, f_t(\mathbf{v}_{t+1})$, this equals $f_t(\mathbf{v}_{t+1})$. The change of measure therefore leaves the distribution of $\mathbf{v}_{t+1}$ unchanged and preserves its independence from the factors.

Note that $\mathbb{E}_t[M_{t,t+1}] = e^{-r_t}$ using the functional form of $M_{t,t+1}$. We then have $P_t^X = e^{-r_t}\mathbb{E}_t^{\mathbb{Q}}[X_{t+1}]$: any one-period ahead cashflow is priced with the risk-free rate after the change of measure, as if marginal investors are risk-neutral. The equivalent $\mathbb{Q}$-measure is therefore also known as the (one-period) risk-neutral measure.

By induction, an asset that pays X_{t+T} at time $t+T$ has time-t price

$$P_t^X = \mathbb{E}_t[M_{t,t+1}P_{t+1}^X] = \mathbb{E}_t[M_{t,t+T}X_{t+T}], \tag{IA.164}$$

where $M_{t,t+T} = \Pi_{s=0}^{T-1} M_{t+s,t+s+1}$ is the T-period pricing kernel. Equivalently, $P_t^X = \mathbb{E}_t^{\mathbb{Q}}[\Pi_{s=0}^{T-1} e^{-r_{t+s}} X_{t+T}]$, where $\Pi_{s=0}^{T-1} e^{-r_{t+s}}$ is now the T-period pricing kernel under the one-period $\mathbb{Q}$-measure.

IA.5.2.3. The Nominal Yield Curve

We apply the Gaussian affine term structure model above to the pricing of the nominal yield curve.[43] Conjecture that the time-t yield of a nominal risk-free n-period pure-discount bond is

$$y_t^n = a_n + \mathbf{b}_n' \mathbf{F}_t, \tag{IA.165}$$

and the corresponding price under the continuous discounting rule is

$$p_t(n) = \exp(-ny_t^n) = \exp\left(-n(a_n + \mathbf{b}_n'\mathbf{F}_t)\right). \tag{IA.166}$$

On the other hand, the bond is also priced by the one-period risk-free bonds under the $\mathbb{Q}$-measure:

$$p_t(n) = e^{-r_t}\mathbb{E}_t^{\mathbb{Q}}[p_{t+1}(n-1)]. \tag{IA.167}$$

[43] In our main analysis, the real pricing kernel under either measure is calculated as the nominal pricing kernel divided by the corresponding inflation $\pi_{t,t+T}$.

Substituting (IA.166), (IA.156) and (IA.163) into (IA.167), we can calculate $a_n, \mathbf{b}_n$ recursively according to

$$\mathbf{b}_n = \frac{1}{n}\sum_{i=1}^{n}[(\rho^{\mathbb{Q}})^{i-1}]'\delta_1, \tag{IA.168}$$

$$a_n = \delta_0 + \frac{(\mathbf{c}^{\mathbb{Q}})'}{n}(\sum_{i=1}^{n-1} i\mathbf{b}_i) - \frac{1}{2n}\sum_{i=1}^{n-1} i^2 \mathbf{b}_i' \mathbf{\Sigma}\mathbf{\Sigma}' \mathbf{b}_i. \tag{IA.169}$$

The estimation strategy follows Hamilton and Wu (2012b). Let $\mathbf{Y}_t^1$ be the L yields in the data that are perfectly spanned by latent factors, and $\mathbf{Y}_t^2$ be other yields that may contain an error. Let $\mathbf{A}_1, \mathbf{A}_2$ be vectors and $\mathbf{B}_1, \mathbf{B}_2$ be matrices such that $\mathbf{Y}_t^1 = \mathbf{A}_1 + \mathbf{B}_1\mathbf{F}_t$ and $\mathbf{Y}_t^2 = \mathbf{A}_2 + \mathbf{B}_2\mathbf{F}_t$. Similar to Ang and Piazzesi (2003), we adopt the identifying restrictions that $\mathbf{c} = 0, \mathbf{\Sigma} = I_L$, and ρ is lower triangular. From $\mathbf{Y}_t^1 = \mathbf{A}_1 + \mathbf{B}_1\mathbf{F}_t$ and $\mathbf{F}_{t+1} = \rho\mathbf{F}_t + \mathbf{u}_{t+1}$, we have

$$(\mathbf{B}_1)^{-1}(\mathbf{Y}_{t+1}^1 - \mathbf{A}_1) = \rho(\mathbf{B}_1)^{-1}(\mathbf{Y}_t^1 - \mathbf{A}_1) + \mathbf{u}_{t+1}^1, \tag{IA.170}$$

$$\mathbf{Y}_{t+1}^1 = (I_L - \mathbf{B}_1\rho(\mathbf{B}_1)^{-1})\mathbf{A}_1 + \mathbf{B}_1\rho(\mathbf{B}_1)^{-1}\mathbf{Y}_t^1 + \mathbf{u}_{t+1}^1, \tag{IA.171}$$

where $\mathbf{u}_{t+1}^1 = \mathbf{B}_1\mathbf{u}_{t+1}$. For $\mathbf{Y}_{t+1}^2$, we have

$$\mathbf{Y}_{t+1}^2 = \mathbf{A}_2 + \mathbf{B}_2\mathbf{F}_{t+1} + \mathbf{u}_{t+1}^2 \tag{IA.172}$$

$$= \mathbf{A}_2 - \mathbf{B}_2(\mathbf{B}_1)^{-1}\mathbf{A}_1 + \mathbf{B}_2\mathbf{B}_1^{-1}\mathbf{Y}_{t+1}^1 + \mathbf{u}_{t+1}^2, \tag{IA.173}$$

where $\mathbf{u}_{t+1}^2 \sim N(0, \mathbf{\Sigma}_2\mathbf{\Sigma}_2')$ for lower-triangular $\mathbf{\Sigma}_2$ and is independent of $\mathbf{u}_{t+1}^1$. It is equivalent to the following restricted VAR:

$$\mathbf{Y}_{t+1} = \hat{\mathbf{A}} + \hat{\rho}\mathbf{Y}_t + \varepsilon_{t+1}, \tag{IA.174}$$

with

$$\hat{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} = \begin{bmatrix} \left(I_L - \mathbf{B}_1 \rho \mathbf{B}_1^{-1}\right) \mathbf{A}_1 \\ \mathbf{A}_2 - \mathbf{B}_2 \rho \mathbf{B}_1^{-1} \mathbf{A}_1 \end{bmatrix}, \tag{IA.175}$$

$$\hat{\rho} = \begin{bmatrix} \hat{\rho}_{11} & 0 \\ \hat{\rho}_{21} & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \rho \mathbf{B}_1^{-1} & 0 \\ \mathbf{B}_2 \rho \mathbf{B}_1^{-1} & 0 \end{bmatrix}, \tag{IA.176}$$

$$\varepsilon_{t+1} = \begin{bmatrix} \varepsilon_{t+1}^1 \\ \varepsilon_{t+1}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{u}_{t+1}^1 \\ \mathbf{B}_2 \mathbf{B}_1^{-1} \mathbf{u}_{t+1}^1 + \mathbf{u}_{t+1}^2 \end{bmatrix}. \tag{IA.177}$$

By comparing the number of parameters identified in the restricted VAR with the number of parameters in the term structure model, we conclude that taking $\mathbf{Y}_t^2$ to be one-dimensional yields exact identification (Hamilton and Wu, 2012b).

IA.5.3. Forecasting Macro Variables

We use the latent factors from the estimated term-structure model to forecast the macro variables of interest. Specifically, we assume the following data-generating process for real GDP growth, inflation, and tax rates, $X_t^j \in X_t = (\log Y_t/Y_{t-1}, \log P_t/P_{t-1}, \log \theta_t)$:

$$X_t^j = \alpha^j + \sum_i \beta^{ji} F_t^i + u_t^{X,j}, \tag{IA.178}$$

where F_t^i is the ith latent term-structure factor, and $u_t^{X,j}$ is an idiosyncratic disturbance that is independent across equations and follows an AR(1) process: $u_t^{X,j} = \rho^{X,j} u_{t-1}^{X,j} + \epsilon_t^j$, $\mathbb{E}_{t-1}[\epsilon_t^j] = 0$, and α^j, β^{ji} are parameters in this linear model. Because, in our specification, these residual disturbances are independent of the factor shocks—jointly Gaussian with, and uncorrelated with, them—and the Radon–Nikodym derivative for the risk-neutral measure depends on the latent factors alone, the change of measure tilts only the factor-shock distribution and leaves the conditional law of the residual innovations unchanged (see Appendix IA.5.2.2). The simulated macro paths therefore retain their unspanned variation under $\mathbb{Q}$, even though their factor-driven component follows the $\mathbb{Q}$ factor dynamics.

For each QE program, we estimate this system using data from November 1971 through the end of the program's purchase phase. We then simulate 1,000,000 paths for the macro variables conditional on the simulated factor paths and compute the pricing-kernel and upper-bound cost estimates from the resulting joint distribution of returns, output, prices, and tax rates. In Appendix IA.7.1, we show that the results are robust to replacing these separate forecasting equations with a VAR that includes the latent term-structure factors as exogenous regressors.

IA.5.4. Baseline Unwinding Horizon

Our baseline specification sets the unwinding horizon to $T = 10$ years. We interpret this horizon as a conservative exposure window for the incremental QE portfolio.

Two pieces of evidence motivate this choice. First, duration-equivalent comparisons with realized quantitative tightening suggest that a horizon of roughly this length is consistent with the time required for subsequent balance-sheet runoff to offset earlier QE expansions. Appendix Figure IA.1 shows that cumulative quantitative tightening (QT) reductions matched the QE1 stock (in 10-year duration equivalents) by August 2018, approximately 8.5 years after QE1 net purchases ended, and that cumulative QT1 and QT2 reductions matched the combined QE1 and QE2 stock by February 2023, approximately 11.5 years after QE2 net purchases ended.

Second, survey-based expectations contemporaneous to the implementation of the QE programs typically implied reinvestment and runoff horizons below 7 years. New York Fed Survey of Primary Dealers (SPD) vintages imply a median expected time from the end of net purchases to the start of runoff of about three years for QE2, roughly two years for QE3, and only a few months for QE4, with a median implied runoff duration of about three years for QE4 and about one year for QE3.[44]

Relative to these benchmarks, the $T = 10$ baseline is conservative in our framework because the quantitative exercise assumes a discrete unwind at date T rather than gradual passive runoff. By maintaining the maturity-transformation wedge until the unwind date,

[44]We use the November 2011 SPD SOMA path for QE2, the September 2015 SPD questions on the most likely end of reinvestments and the associated phase-out duration for QE3, and the March 2022 SPD projections for net monthly changes and the quarter in which SOMA "ceases to decline" for QE4. For QE2, the published SOMA path ends in 2015 while the median remains declining, so this vintage does not pin down an end-of-runoff horizon.

this convention preserves the full exposure of the incremental QE portfolio until T, whereas under passive runoff the exposure would decline gradually over time.

IA.5.5. Measuring Output Effects

Literature surveyed by Fabo et al. (2021) To benchmark our cost estimates against the output benefits of QE, we draw on the 18 U.S. studies surveyed by Fabo et al. (2021). Our measure of the output effect uses the cumulative estimates reported for each study in Appendix Table A.4 of Fabo et al. (2021). We take each estimate at the end of the horizon considered by the authors and express it as a percentage of GDP in the base year.

Six of these studies report cumulative effects for at least two Fed QE programs. For each such paper, we treat each program as a separate observation. For a given paper–program observation, we allocate the study's reported cumulative effect across the programs it considers in proportion to each program's size, measured in 10-year duration equivalents. This procedure yields 28 paper–program observations. For each paper–program observation, we then compute a size-normalized effect, defined as the cumulative output effect per 1% of GDP in 10-year duration-equivalent purchases.

We then exclude paper–program observations for which the size-normalized effect is below the 5th percentile or above the 95th percentile of the size-normalized effect distribution across papers. This trimming leaves 24 paper–program observations, drawn from 16 of the 18 surveyed papers, which we report in Appendix Table IA.I. In this sample, the unweighted average (median) cumulative output effect per 1% of 10-year duration equivalent to GDP is 0.07 (0.03) percentage points of GDP, with a standard deviation of 0.08 percentage points of GDP. Fabo et al. (2021) show that central-bank-affiliated authors report larger QE effects on both output and inflation than academic authors. In our trimmed sample, 7 of the 24 paper–program observations come from papers with no authors affiliated with central banks. For this restricted sample, the average (median) cumulative output effect per 1% of 10-year duration equivalent to GDP is 0.02 (0.02) percentage points of GDP, and the standard deviation is 0.01 percentage points of GDP.

The studies surveyed by Fabo et al. (2021) cover QE1, QE2, and QE3 only. To obtain an output effect for MEP and QE4, we first compute the average size-normalized effect across the 24 (7 for the restricted sample) paper–program observations per 1% program size measured in 10-year duration equivalents to GDP, which is 0.07 percentage points of GDP. We then extrapolate the output effects for MEP and QE4 by multiplying this average size-normalized effect by the respective sizes of the MEP and QE4 programs.

Reference	Central Bank Affiliation	QE Program	Method	Output Effect, GDP %	Output Effect per 10-y Dur. Equiv. pp of GDP, GDP %
Chung et al. (2012)	1	LSAP1	VAR	0.40	0.05
Balatti et al. (2017)	0	LSAP1	VAR	0.13	0.02
Balatti et al. (2017)	0	LSAP2	VAR	0.05	0.02
Baumeister and Benati (2013)	1	LSAP1	VAR	0.98	0.12
Chen, Cúrdia, and Ferrero (2012)	1	LSAP2	DSGE	0.07	0.02
Dahlhaus, Hess, and Reza (2018)	1	LSAP1	VAR	0.86	0.11
Dahlhaus, Hess, and Reza (2018)	1	LSAP2	VAR	0.34	0.11
Dahlhaus, Hess, and Reza (2018)	1	LSAP3	VAR	1.60	0.11
Del Negro et al. (2017)	1	LSAP1	DSGE	0.25	0.03
Del Negro et al. (2017)	1	LSAP2	DSGE	0.10	0.03
Engen, Laubach, and Reifschneider (2015)	1	LSAP1	DSGE	0.28	0.03
Engen, Laubach, and Reifschneider (2015)	1	LSAP2	DSGE	0.11	0.03
Engen, Laubach, and Reifschneider (2015)	1	LSAP3	DSGE	0.52	0.03
Falagiarda (2014)	0	LSAP2	DSGE	0.01	0.00
Fuhrer and Olivei (2011)	1	LSAP2	VAR	0.78	0.25
Gambacorta, Hofmann, and Peersman (2014)	1	LSAP1	VAR	0.12	0.02
Gertler and Karadi (2013)	1	LSAP1	DSGE	0.01	0.00
Haldane et al. (2016)	1	LSAP2	VAR	0.94	0.30
Hesse, Hofmann, and Weber (2018)	1	LSAP2	VAR	0.54	0.17
Popescu (2015)	0	LSAP2	VAR	0.04	0.01
Weale and Wieladek (2016)	1	LSAP1	VAR	0.34	0.04
Wu and Xia (2016)	0	LSAP1	VAR	0.23	0.03
Wu and Xia (2016)	0	LSAP2	VAR	0.09	0.03
Wu and Xia (2016)	0	LSAP3	VAR	0.44	0.03

Table IA.I: QE studies from Fabo et al. (2021). This table presents the effects of the QE program studied on cumulative output for each study-program observation in our sample. We report the effects on the output level in percent. The effects are reported as a percentage of real GDP at the end of each program's purchase phase, Y_0. Standardized effects are obtained by dividing the total program effects by the size of each program over a 10-year duration, equivalent to GDP.

Table IA.II presents the average cumulative effect of each QE program on the level of GDP implied by the conversion, trimming, and imputation steps described above.

	QE1	QE2	MEP	QE3	QE4	All	1pp of GDP
Output Effect (all)	0.36	0.28	0.71	0.85	1.18	3.38	0.06
Output Effect (academia)	0.18	0.05	0.21	0.44	0.34	1.21	0.02
Output Effect (DSGE)	0.18	0.07	0.26	0.52	0.43	1.45	0.03
Output Effect (VAR)	0.44	0.40	0.94	1.02	1.55	4.34	0.08

Table IA.II: Output effects from the literature surveyed in Fabo et al. (2021). The table presents the effects of the QE program studied on cumulative output averaged by program from our sample of paper-program observations. The effects are reported as a percentage of real GDP at the end of each program's purchase phase, Y_0.

The first column reports the average effect on output for each QE program. The second column reports the average effect on output for each QE program for paper-program observations in the sample of academic authors, i.e., where none of the authors of the paper is affiliated with a central bank.

Discounting convention For comparability with the fiscal-efficiency cost measure, we convert the cumulative output effects reported in the literature into date-0 present values using

$$\underline{p}_0^r(T) \equiv \min_{t\in\{1,\ldots,T\}} p_0^r(t),$$

where $p_0^r(t)$ denotes the date-0 price of a real risk-free claim paying one unit of consumption at date t.

As a result, any benefit that accrues before T is assigned a weakly lower present value than under horizon-specific discounting. We view this discounting convention as being conservative because QE is typically viewed as a short- to medium-run stabilization tool: estimated macro responses in VAR/DSGE-style exercises are often hump-shaped, peaking within a few quarters to a couple of years and then fading as financial conditions normalize and the economy exits the ZLB; see, e.g., Weale and Wieladek (2016) and Kim, Laubach, and Wei (2020).

Output Effects from Recent Studies This appendix translates findings from recent studies into output-effect estimates that are comparable to those from the Fabo et al. (2021) collection exercise.

For each study, we express the reported change in output as a share of GDP and normalize it by the size of the corresponding QE intervention, measured in percentage points of GDP in 10-year duration equivalents. The resulting estimates represent the change in output (as a percentage of GDP) implied by a purchase of 1% of GDP in 10-year duration equivalents.

Fieldhouse, Mertens, and Ravn (2018) identify exogenous shocks to U.S. federal expansions in mortgage lending support from 1968-2014 using a narrative identification strategy. Figure X in their paper indicates that a 1 pp increase in expected future agency commitments as a ratio of mortgage originations raises cumulative personal consumption by 0.15% after 24 months. Given that consumption averages 63.9% of GDP over their sample period,[45] this translates into a 0.095% increase in GDP. Since originations account for about 20% of mortgage debt and mortgage debt for roughly 20% of GDP (upper bound given Figure I), a 1 pp increase in commitments corresponds to 0.04 pp of GDP in purchases. Scaling the estimated consumption effect accordingly yields an implied conservative output effect of 2.4% of GDP per 1% of GDP in agency purchases. Taking a 3-year average duration for agency purchases—reflecting the high-interest-rate environment and faster prepayments over their sample period—1% of GDP in face value corresponds to 0.3% of GDP in 10-year duration equivalents, so this implies an output effect of approximately 8% of GDP per 1% of GDP in 10-year duration equivalent purchases.[46]

Di Maggio, Kermani, and Palmer (2020) exploit the differential impact of QE1 on conforming versus jumbo mortgages to estimate the refinancing-driven consumption response. They report that the decision to purchase MBS rather than Treasuries during QE1 increased aggregate consumption by approximately $13.5 billion, or 0.111% of GDP at the time.[47] Given total MBS purchases of 5.53 pp of GDP in 10-year duration equivalents, this implies an output effect of about 0.02% of GDP per 1% of GDP in 10-year duration equivalent purchases.

Ray, Droste, and Gorodnichenko (2024) identify auction-surprise demand shocks and cali-

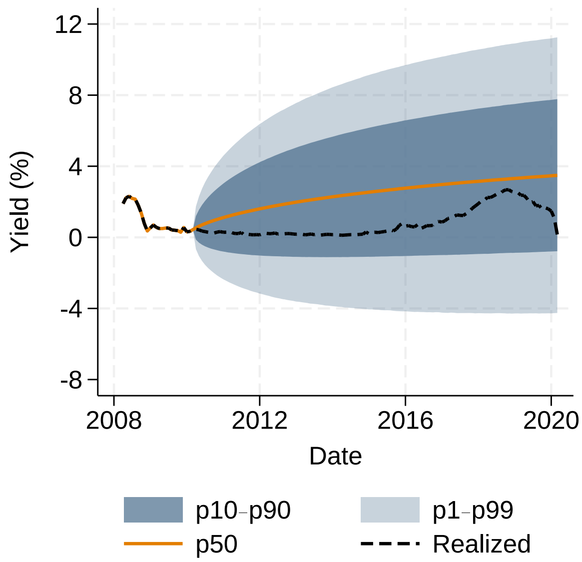
[45] See series DPCERE1Q156NBEA from FRED.

[46] For comparison, Krishnamurthy and Vissing-Jorgensen (2011) report a duration of about 7 years for agency MBS in low-rate environments, which would imply a smaller converted effect of approximately 3.4%.

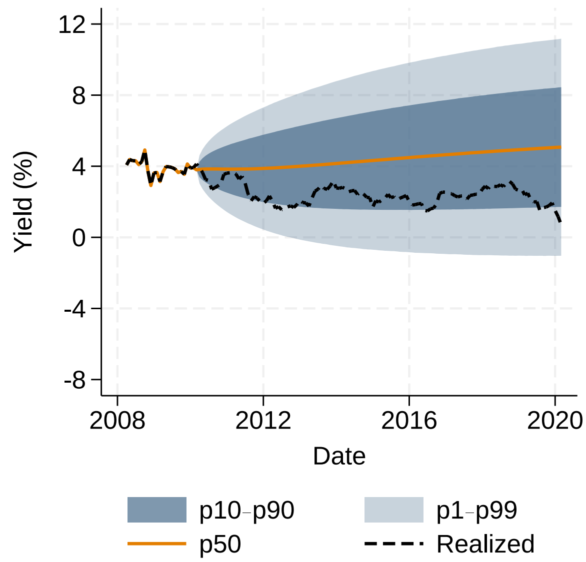
[47] GDP at the end of 2010Q1 is $12.161 trillion.

brate a preferred-habitat DSGE model that matches yield-curve responses. In their baseline calibration, QE1 has a cumulative effect on the output of 0.56 pp. Note that QE1 purchases total 7.90 pp of GDP in 10-year duration equivalents. Dividing the cumulative output response by the purchase size implies an output effect of approximately 0.071% of GDP per 1% of GDP in 10-year duration equivalent purchases.

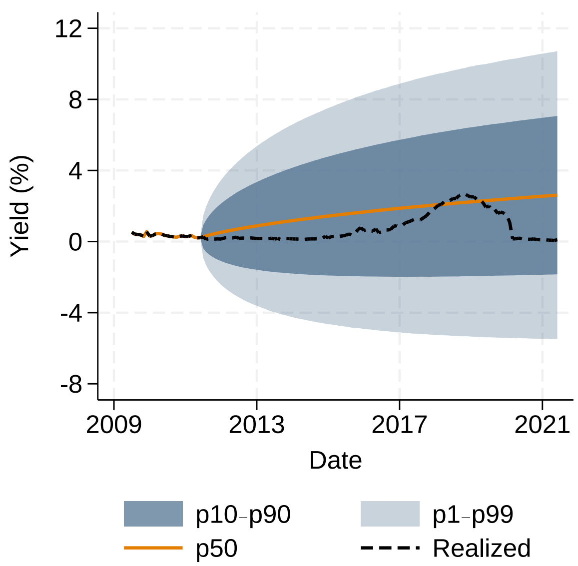
IA.6. Term-Structure Model Forecasts



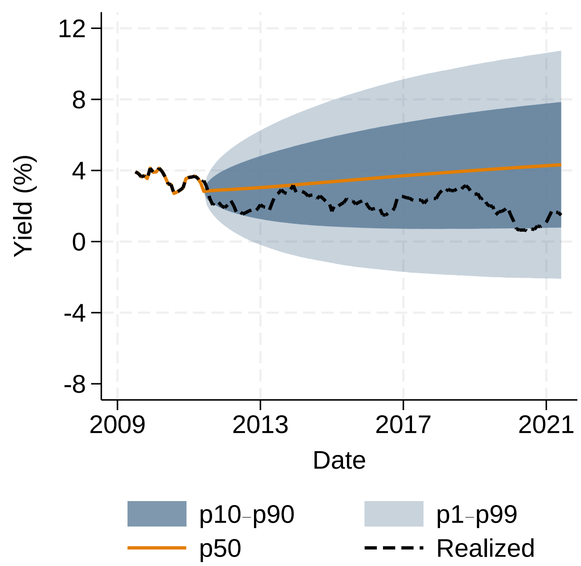
(a) 1-year yield forecasts for QE1



(b) 10-year yield forecasts for QE1

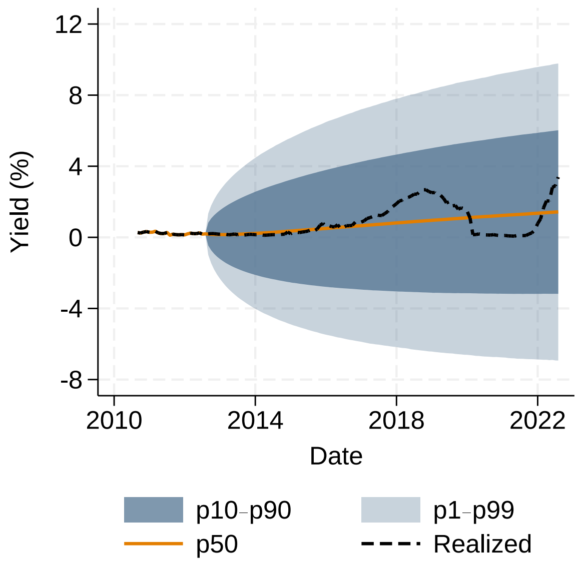


(c) 1-year yield forecasts for QE2

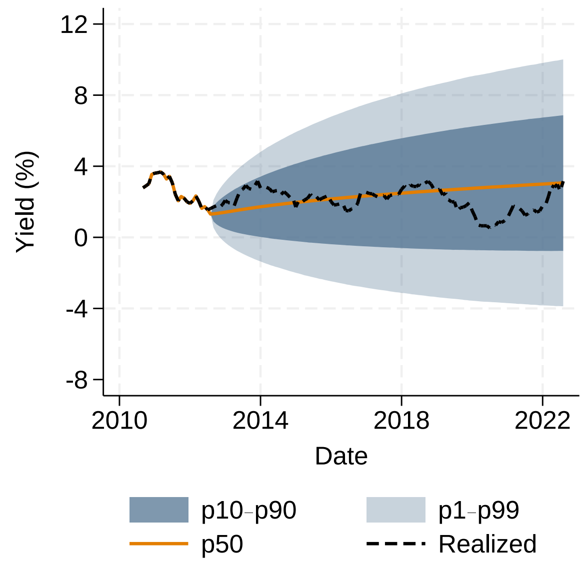


(d) 10-year yield forecasts for QE2

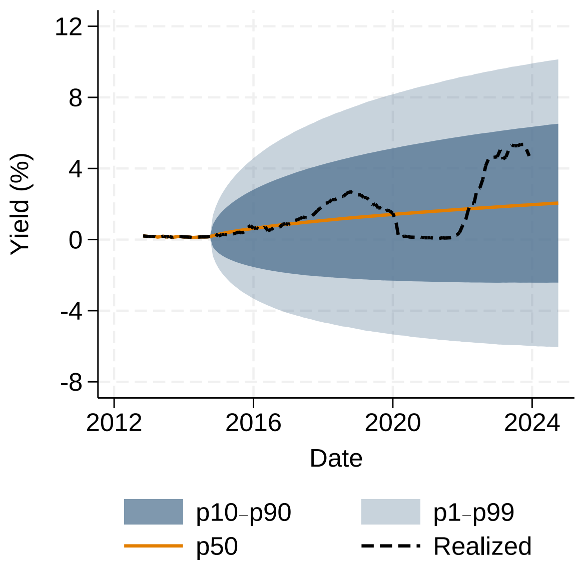
Figure IA.3: Term structure forecasts for QE1 and QE2. The figure plots the distribution of 10-year-ahead forecasts for the 1-year and 10-year yields (1st, 10th, 50th, 90th, and 99th percentiles across 1,000,000 simulated paths under physical measure) and compares them with realized yields.



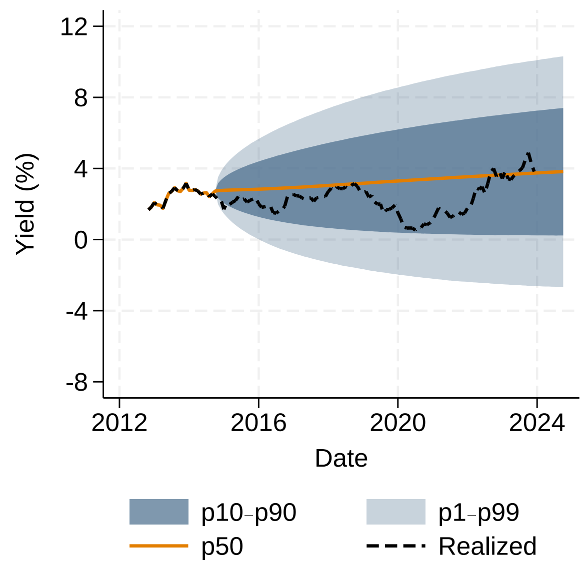
(a) 1-year yield forecasts for MEP



(b) 10-year yield forecasts for MEP

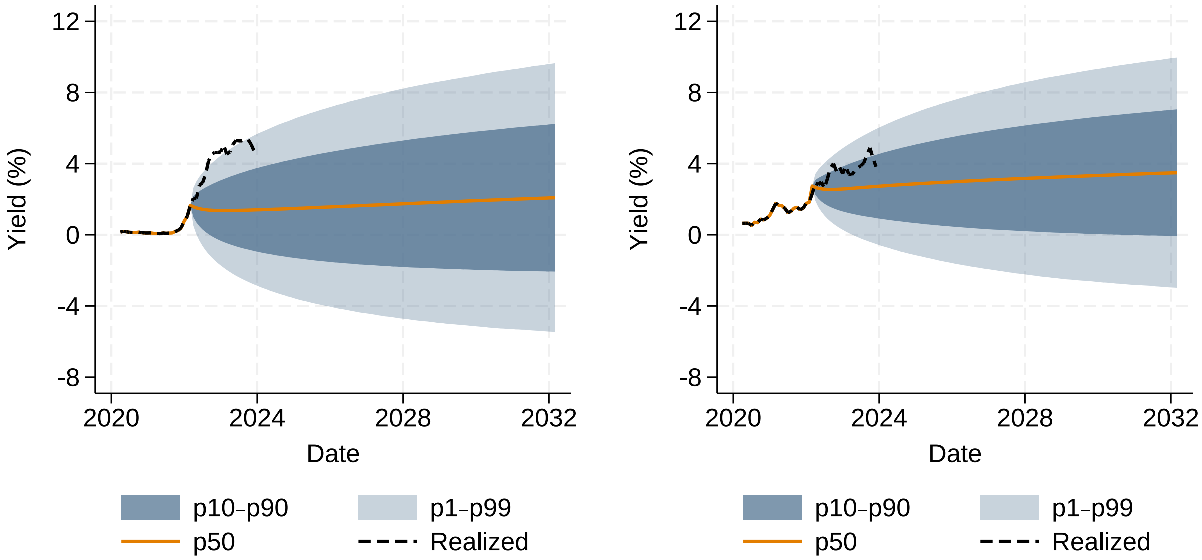


(c) 1-year yield forecasts for QE3



(d) 10-year yield forecasts for QE3

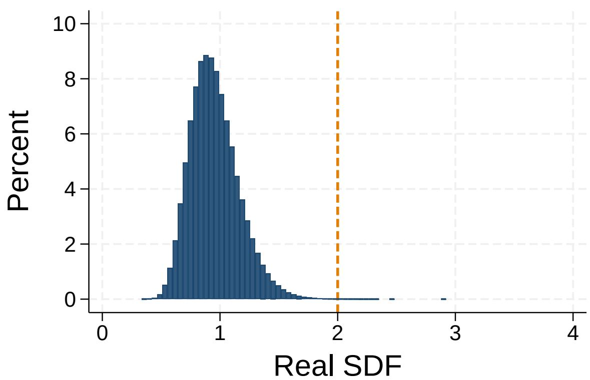
Figure IA.4: Term structure forecasts for MEP and QE3. The figure plots the distribution of 10-year-ahead forecasts for the 1-year and 10-year yields (1st, 10th, 50th, 90th, and 99th percentiles across 1,000,000 simulated paths under physical measure) and compares them with realized yields.



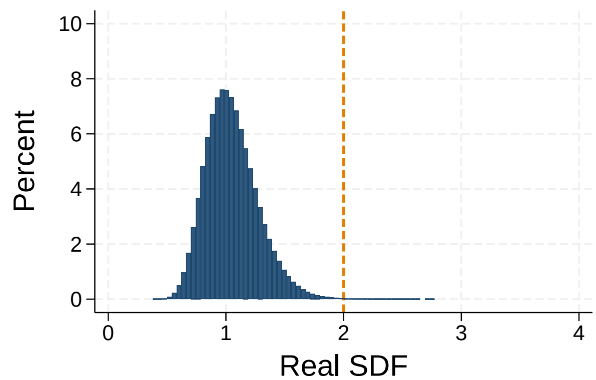
(a) 1-year yield forecasts for QE4

(b) 10-year yield forecasts for QE4

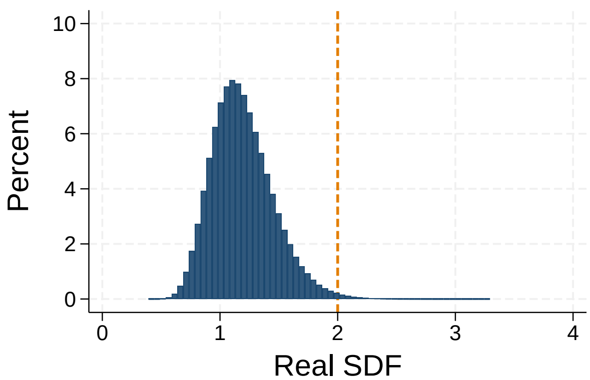
Figure IA.5: Term structure forecasts for QE4. The figure plots the distribution of 10-year-ahead forecasts for the 1-year and 10-year yields (1st, 10th, 50th, 90th, and 99th percentiles across 1,000,000 simulated paths under physical measure) and compares them with realized yields.



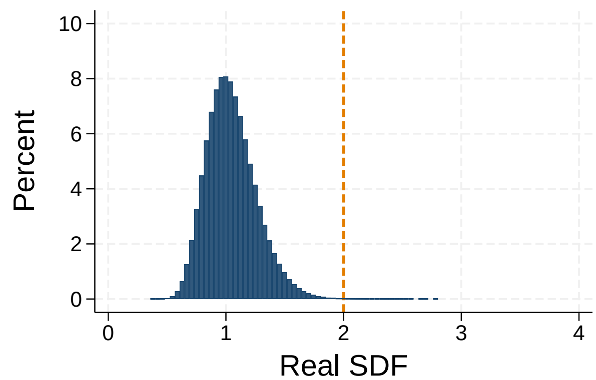
(a) QE1



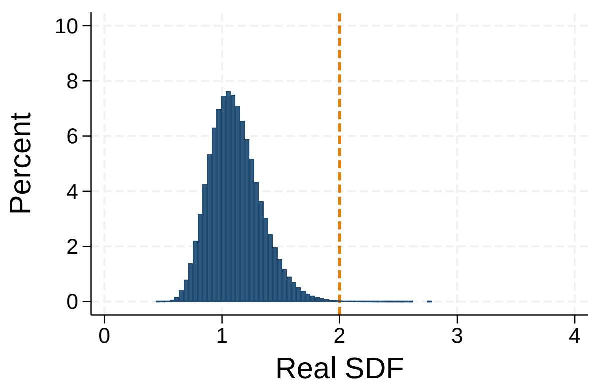
(b) QE2



(c) MEP



(d) QE3



(e) QE4

Figure IA.6: Cap on long-horizon real state prices by program. The figure plots the Monte Carlo distribution of the long-horizon real state-price variable $\lambda_{0,T}$ (real $\mathbb{Q}$-discount factor) used in the valuation-weighted moments in (28), separately for each QE program. The horizontal line indicates the cap $\lambda_{0,T} \leq \bar{\lambda}$ with $\bar{\lambda} = 2$ at the baseline horizon $T = 10$. This tail-regularity restriction trims only very extreme draws: in the program-level Monte Carlo used for our forecasts, approximately 99.5% of simulated paths satisfy the bound.

IA.7. Sensitivity Analyses

IA.7.1. Alternative Term-Structure Model Specifications

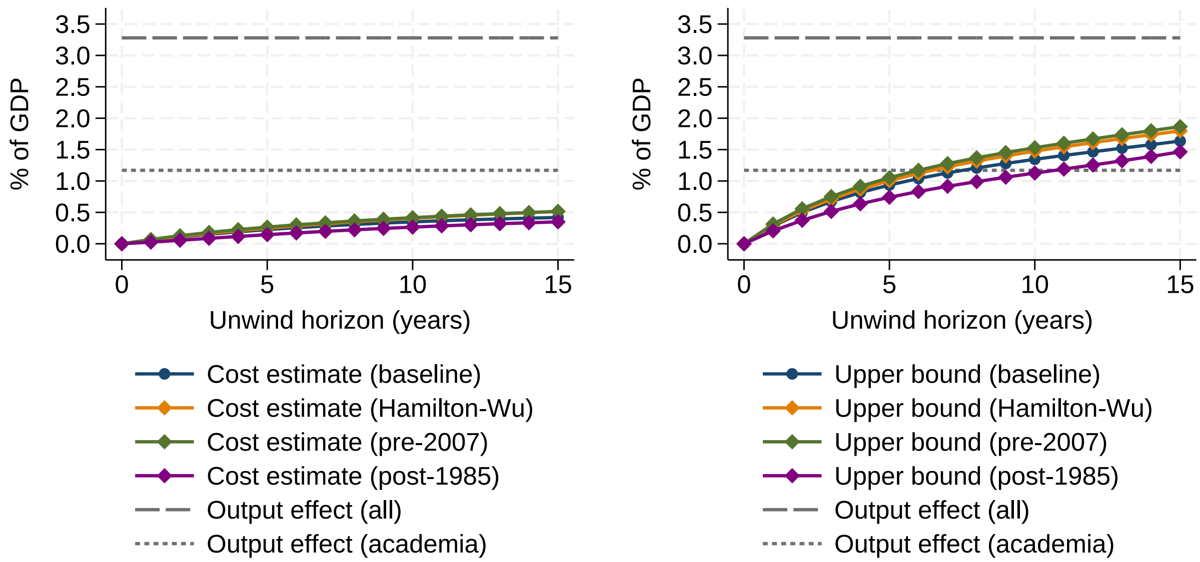
Term-structure model estimates Due to limited availability of yield data with maturities up to 15 years before 1971, the estimation sample for our baseline term-structure model begins in 1971. For each QE program, we estimate the model using yield data through the end of the purchase phase, simulate 1,000,000 future factor paths, and construct the corresponding yield-curve and macroeconomic paths up to 15 years ahead. We then compute QE portfolio returns along each simulated yield-curve path, matching cash flows beyond 15 years by converting them into 15-year-maturity equivalents, that is, by multiplying the amount by a factor of $m/15$.

In this subsection, we consider three alternative term-structure specifications as robustness checks: *Hamilton–Wu*, *Post-1985*, and *Pre-2007*.

First, we consider the parameterization in Hamilton and Wu (2012b) (hereafter, *Hamilton–Wu*). Their model is estimated on yield data from 1952 to 2000 using maturities up to 5 years. As a robustness check, we repeat our cost analysis using these term-structure parameters. For each QE program, we reconstruct latent factors at the end of the purchase phase using the Hamilton–Wu parameters, simulate 1,000,000 future factor, yield-curve, and macroeconomic paths up to 15 years ahead, and compute QE portfolio returns along each simulated yield-curve path, again matching cash flows beyond 15 years by converting them into 15-year-maturity equivalents.

Second, we estimate our term-structure model using yields with maturities up to 30 years rather than 15 years (hereafter, *Post-1985*). Because long-maturity yields are available only from 1985 onward, this specification uses a shorter estimation sample. For each QE program, we estimate the model through the end of the purchase phase and simulate 1,000,000 future factor, yield-curve, and macroeconomic paths up to 15 years ahead. When computing QE portfolio returns, we use the full QE cash-flow structure rather than converting cash flows beyond 15 years into 15-year-maturity equivalents.

Third, we estimate the term-structure model using yield data through January 2007 with maturities up to 15 years in order to exclude the crisis and post-crisis period from estimation



(a) Pricing-kernel cost estimate

(b) Upper-bound on cost estimate

Figure IA.7: Aggregate fiscal-efficiency costs: Estimation samples. The figure plots the aggregate exact pricing-kernel cost and conservative upper bound estimate as a function of the unwind horizon T, expressed as a percentage of real GDP at the end of each program's purchase phase (Y_0). The baseline series uses our preferred term-structure specification. *Hamilton–Wu* series uses term-structure model parameters from Hamilton and Wu (2012b) specification (HW). *Pre-2007* series uses a term-structure model estimated with yield-curve data before 2007. *Post-1985* series uses a term-structure model estimated with yield-curve data after 1985. The dashed horizontal lines report the output-effect benchmarks from Fabo et al. (2021). Estimates are computed from 1,000,000 term-structure simulations under the $\mathbb{Q}$-measure.

(hereafter, *Pre-2007*). For each QE program, we reconstruct latent factors at the end of the purchase phase using these pre-2007 parameters, simulate 1,000,000 future factor, yield-curve, and macroeconomic paths up to 15 years ahead, and compute QE portfolio returns along each simulated yield-curve path, matching cash flows beyond 15 years by converting them into 15-year-maturity equivalents.

The costs of QE are estimated analogously to the baseline analysis. The results are reported in Figure IA.7.

Relative to the baseline specification, the cost estimates are slightly larger under the *Hamilton–Wu* and *Pre-2007* specifications, which exclude the low-interest-rate period from estimation, and slightly smaller under the *Post-1985* specification, which excludes the volatile interest-rate environment of the 1970s.

A VAR specification for macro variable forecasts In our baseline specification, we assume the following data-generating process for real GDP growth, inflation, and tax rates, $X_t^j \in X_t = (\log Y_t/Y_{t-1}, \log P_t/P_{t-1}, \log \theta_t)$:

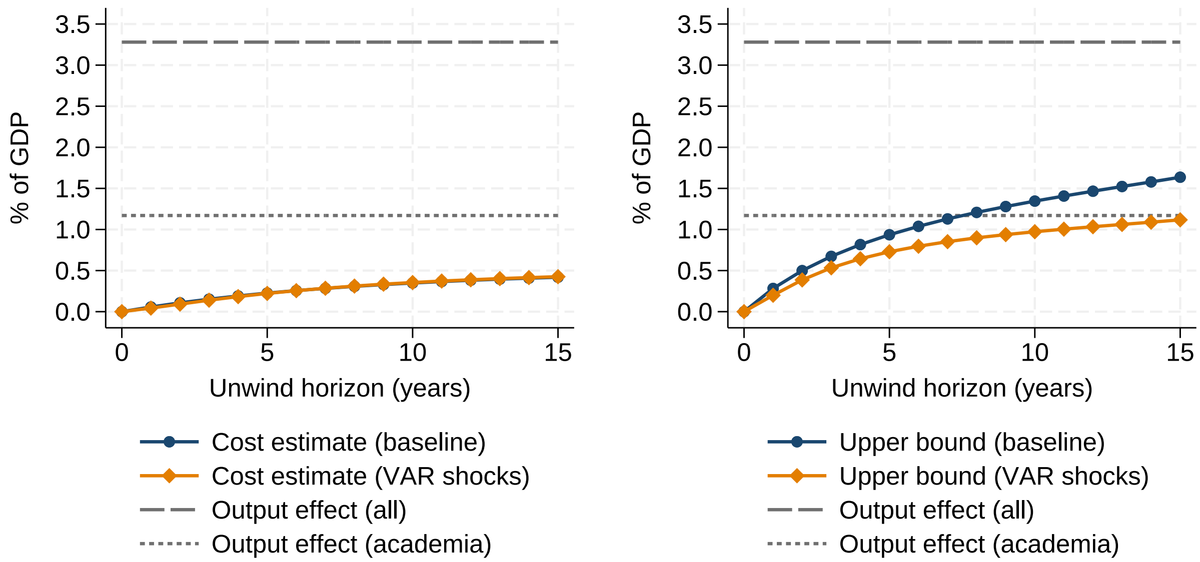
$$X_t^j = \alpha^j + \sum_i \beta^{ji} F_t^i + u_t^{X,j}, \tag{IA.179}$$

where F_t^i is the i-th term-structure factor we identify, $u_t^{X,j}$ independent of $\mathbf{u}_t^{X,-j}$ and follows an AR(1) process: $u_t^{X,j} = \rho^{X,j} u_{t-1}^{X,j} + \eta_t^j$, $\mathbb{E}_{t-1}[\eta_t^j] = 0$, and α^j, β^{ji} are parameters in this linear model. This is equivalent to a VAR specification that restricts all common shocks to be captured by the term structure factors. To relax this structure, we use an alternative full-VAR specification for macro variable dynamics in this section and repeat our cost analysis. To be specific, we consider the following VAR:

$$\mathbf{X}_t = \alpha + \beta \mathbf{F}_t + \gamma \mathbf{X}_{t-1} + \Sigma^u \mathbf{u}_t^X, \tag{IA.180}$$

where α is a 3×1 vector, β, γ are both 3×3 matrices, Σ^u is a 3×3 lower-triangular matrix, and $\mathbf{u}_t^X \sim N(0, I_3)$ is i.i.d. We first estimate the VAR model taking the term structure factors as exogenous variables, and then follow the same procedure as in the baseline specification to generate 1,000,000 paths for macro variables corresponding to the term structure paths.

We calculate our cost estimates for each QE program accordingly and present the total cost estimates across programs in Figure IA.8. Relative to the baseline, the total pricing-kernel estimate is nearly unchanged, while the total upper-bound estimate is slightly lower.



(a) Pricing-kernel cost estimate

(b) Upper-bound on cost estimate

Figure IA.8: Aggregate fiscal-efficiency costs: VAR specification. The figure plots the aggregate exact pricing-kernel cost and conservative upper bound estimate as a function of the unwind horizon T, expressed as a percentage of real GDP at the end of each program's purchase phase (Y_0). The baseline series assumes all common shocks are captured by the term structure factors. The alternative series specifies a full-VAR model for macro variables with the term structure factors as additional predictors. The dashed horizontal lines report the output-effect benchmarks from Fabo et al. (2021). Estimates are computed from 1,000,000 term-structure simulations under the $\mathbb{Q}$-measure.

IA.7.2. Alternative Financing Rules

We relax the terminal financing rule of QE portfolios in this section. We assume that the cumulative net return on the QE portfolio will be offset by $K \geq 1$ consecutive tax adjustments starting from horizon T. Still denote R_T^{QE} the return on the whole QE portfolio at T, and denote $R_t^{\text{QE}}(K)$ the part of return that needs to offset at t under the K-period financing rule. Assume $R_t^{\text{QE}}(K) = 0$ for $t \leq T-1$ and $t \geq T+K$, and $R_{T+k}^{\text{QE}}(K) = \frac{1}{K} R_T^{\text{QE}} \times \frac{1}{p_T(k)}$, that is, QE portfolios unwind at T, and K consecutive tax adjustments are made to offset an equal proportion of the cumulative return. When the return is offset in $k > 0$ periods, this portion is assumed to be invested in the corresponding risk-free bond that matures in k periods and the return associated is $1/p_T(k)$. Then the total cost with tax adjustments over K periods can be calculated as

$$\Delta^{\mathrm{QE}} L_0(K) = \sum_{k=0}^{K-1} \Delta^{\mathrm{QE}} L_{0,k}(K), \tag{IA.181}$$

where

$$\Delta^{\mathrm{QE}} L_{0,k}(K) = \frac{\alpha}{2}\, \mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T+k} Y_{T+k} \left(r_{T+k}^{\mathrm{QE}}(K)\right)^2\right] - \alpha\, \mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T+k} Y_{T+k}\, \theta_{T+k}^{\mathrm{nQE}}\, r_{T+k}^{\mathrm{QE}}(K)\right], \tag{IA.182}$$

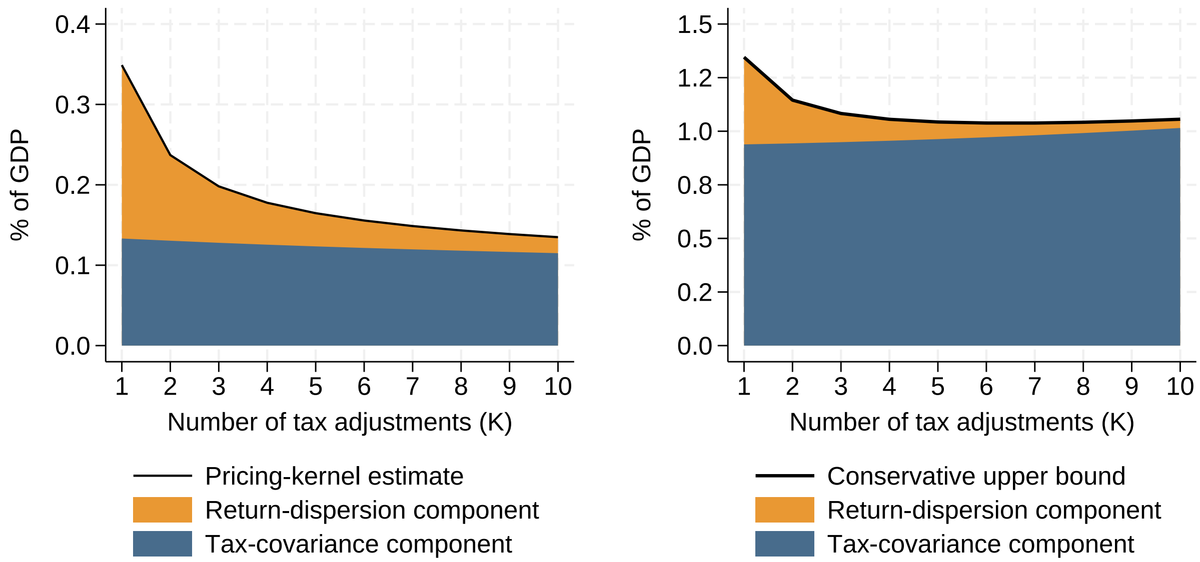
and $r_{T+k}^{\mathrm{QE}}(K) = R_{T+k}^{\mathrm{QE}}(K)/(P_{T+k} Y_{T+k})$.

For the upper-bound estimate, similarly we have

$$\begin{aligned}\Delta^{\mathrm{QE}} L_{0,k}(K) \leq & \alpha\bar{\lambda}\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[\theta_{T+k}^{\mathrm{nQE}}\right] \mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_{T+k} r_{T+k}^{\mathrm{QE}}(K)\right)^2\right]} \\ & + \frac{\alpha\bar{\lambda}}{2}\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[r_{T+k}^{\mathrm{QE}}(K)\right] \mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_{T+k} r_{T+k}^{\mathrm{QE}}(K)\right)^2\right]},\end{aligned} \tag{IA.183}$$

and the upper-bound estimate for $\Delta^{\mathrm{QE}} L_0(K)$ is derived by summing up the upper-bound estimates for $\Delta^{\mathrm{QE}} L_{0,k}(K)$.

As can be seen in Appendix Figure IA.9, the cost estimates decline when tax adjustments are spread over more future periods. This is mainly driven by the reduction of the return dispersion term due to convexity of the cost function and Jensen's inequality, which reflects the incentive to smooth QE-induced tax adjustments over time so as to equalize marginal losses. The tax covariance term is nearly unchanged because the correlations between the cyclical component of counterfactual taxes and tax adjustments induced by QE in each period are nearly the same when these adjustments are spread evenly over periods.



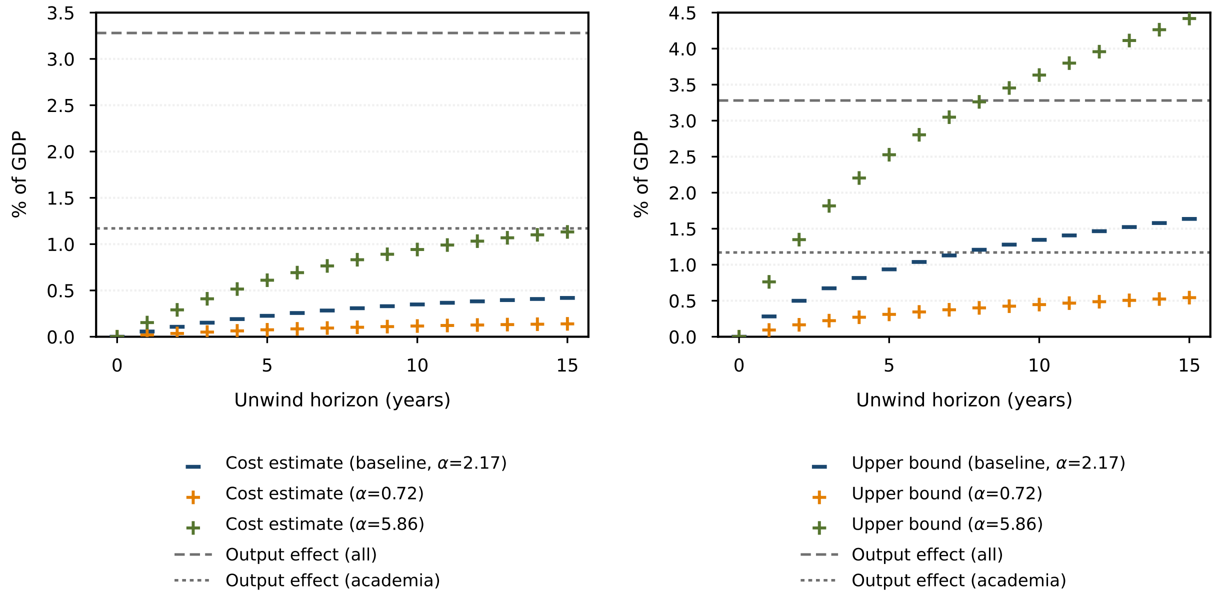
(a) Pricing-kernel cost estimate

(b) Upper-bound on cost estimate

Figure IA.9: Aggregate fiscal-efficiency costs: Smoothing tax adjustments. The figure decomposes aggregate fiscal-efficiency costs into *return dispersion* and *tax covariance* components as a function of the number of tax adjustments K, given $T = 10$. (a) reports the decomposition of the pricing-kernel estimate (28). (b) reports the decomposition of the conservative upper bound (29). Both panels are based on 1,000,000 term-structure simulations under the $\mathbb{Q}$-measure. The conservative bound is computed under the envelope $\lambda_{0,T+k} \leq 2$ for each relevant horizon.

IA.7.3. Alternative Tax Distortion Calibrations

Alternative calibrations for the convexity parameter To relax the baseline calibration assumptions, we compute the fiscal-efficiency costs under alternative calibrations of the tax-loss convexity parameter α. Because α enters multiplicatively in (28) and (29), the cost estimates scale one-for-one with α. Appendix Figure IA.10 reports cost profiles across the interval $\alpha \in [0.72, 5.86]$, obtained by mapping the range of taxable-income elasticities that Saez, Slemrod, and Giertz (2012) consider most reliable ($e \in [0.12, 0.40]$) through the marginal-excess-burden formula used in the baseline calibration; the interval is thus anchored in estimated elasticities rather than chosen values. Across this entire interval, the pricing-kernel cost estimate remains below the academic-only output-effect benchmark at the baseline horizon; the conservative upper bound reaches the mean benchmark only near the top of the elasticity range.



(a) Pricing-kernel cost estimate

(b) Upper-bound on cost estimate

Figure IA.10: Aggregate fiscal-efficiency costs: Tax distortion parameters. The figure plots the aggregate exact pricing-kernel cost and conservative upper bound estimate as a function of the unwind horizon T, expressed as a percentage of real GDP at the end of each program's purchase phase (Y_0). The baseline series uses the calibration $\alpha = 2.17$. The alternative series rescale the baseline estimate proportionally to α across the elasticity-derived interval $[0.72, 5.86]$. The dashed horizontal lines report the output-effect benchmarks from Fabo et al. (2021). Estimates are computed from 1,000,000 term-structure simulations under the $\mathbb{Q}$-measure.

Third-order approximation of the tax-distortion function In our baseline specification, we approximate the loss function with its second-order Taylor expansion around a benchmark tax rate θ_0. We consider a third-order Taylor expansion in this section as a robustness check. For all θ, let

$$\xi(\theta) = \xi(\theta_0) + \kappa(\theta - \theta_0) + \frac{\alpha}{2}(\theta - \theta_0)^2 + \frac{\nu}{6}(\theta - \theta_0)^3,$$

where $\kappa = \xi'(\theta_0)$, $\alpha = \xi''(\theta_0)$, and $\xi'''(\theta_0) = \nu$. Then the difference of loss functions $\xi\left(\theta_t^{\text{QE}}\right) - \xi\left(\theta_t^{\text{nQE}}\right)$ can be expressed as

$$
\begin{aligned}
&\kappa(\theta_t^{\text{QE}} - \theta_t^{\text{nQE}}) + \frac{\alpha}{2}(\theta_t^{\text{QE}} - \theta_t^{\text{nQE}})(\theta_t^{\text{QE}} + \theta_t^{\text{nQE}} - 2\theta_0) \\
&+\frac{\nu}{6}(\theta_t^{\text{QE}} - \theta_t^{\text{nQE}})\left((\theta_t^{\text{QE}} - \theta_0)^2 + (\theta_t^{\text{QE}} - \theta_0)(\theta_t^{\text{nQE}} - \theta_0) + (\theta_t^{\text{nQE}} - \theta_0)^2\right) \qquad \text{(IA.184)} \\
=&(\kappa - \alpha\theta_0)\Delta^{\text{QE}}\theta_t + \alpha\theta_t^{\text{nQE}}\Delta^{\text{QE}}\theta_t + \frac{\alpha}{2}(\Delta^{\text{QE}}\theta_t)^2 \\
&+\frac{\nu}{6}\Delta^{\text{QE}}\theta_t\left(3(\theta_t^{\text{nQE}} - \theta_0)^2 + 3\Delta^{\text{QE}}\theta_t(\theta_t^{\text{nQE}} - \theta_0) + (\Delta^{\text{QE}}\theta_t)^2\right). \qquad \text{(IA.185)}
\end{aligned}
$$

We assume $\Delta^{\text{QE}}\tau_T = -R_T^{\text{QE}}$ and $\Delta^{\text{QE}}\tau_t = 0$ for any $t \neq T$. Also note that $\mathbb{E}_0\left[\Lambda_{0,T}Y_T\Delta^{\text{QE}}\theta_T\right] = -\mathbb{E}_0\left[\Lambda_{0,T}R_T^{\text{QE}}/P_T\right] = 0$ according to the Euler equation. Therefore, the expression for $\Delta^{\text{QE}}L_0$ can be rewritten as

$$
\begin{aligned}
\Delta^{\text{QE}}L_0 =& -\alpha\mathbb{E}_0\left[\Lambda_{0,T}Y_T\left(\theta_T^{\text{nQE}}r_T^{\text{QE}}\right)\right] + \frac{\alpha}{2}\mathbb{E}_0\left[\Lambda_{0,T}Y_T\left(r_T^{\text{QE}}\right)^2\right] \\
&+\frac{\nu}{6}\mathbb{E}_0\left[\Lambda_{0,T}Y_T\left(-3(\theta_T^{\text{nQE}} - \theta_0)^2 + 3(\theta_T^{\text{nQE}} - \theta_0)r_T^{\text{QE}} - (r_T^{\text{QE}})^2\right)r_T^{\text{QE}}\right], \qquad \text{(IA.186)}
\end{aligned}
$$

where $r_T^{\text{QE}} = R_T^{\text{QE}}/P_TY_T$. We can write an equivalent expression for equation above using the risk-neutral $\mathbb{Q}$-measure, as specified in Appendix IA.5.2.2:

$$
\begin{aligned}
\Delta^{\text{QE}}L_0 =& -\alpha\,\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_T\left(\theta_T^{\text{nQE}}\,r_T^{\text{QE}}\right)\right] + \frac{\alpha}{2}\,\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_T\left(r_T^{\text{QE}}\right)^2\right] \\
&+\frac{\nu}{6}\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_T\left(-3(\theta_T^{\text{nQE}} - \theta_0)^2 + 3(\theta_T^{\text{nQE}} - \theta_0)r_T^{\text{QE}} - (r_T^{\text{QE}})^2\right)r_T^{\text{QE}}\right], \qquad \text{(IA.187)}
\end{aligned}
$$

where $\lambda_{0,T} = m_{0,T}\frac{P_T}{P_0}$. $1/\lambda_{0,T}$ denotes the date-T real value of a strategy that rolls over one-period nominal risk-free bonds up to date T: the strategy invests in one unit of consumption good and gets $\frac{P_0}{m_{0,T}}$ in terms of money at T, which is equivalent to $\frac{1}{m_{0,T}}\frac{P_0}{P_T}$ units of consumption good at T.

Similar to Appendix IA.1, we can derive the expression for an upper bound of $\Delta^{\text{QE}}L_0$.

Consider in general the expression $\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T \, x_T \, r_T^{\mathrm{QE}}\right]$ for some x_T. Notice that

$$\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T \, x_T \, r_T^{\mathrm{QE}}\right] = \mathrm{Cov}_0^{\mathbb{Q}}\left[x_T, \lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right] \leq \sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[x_T\right] \mathrm{Var}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right]} \tag{IA.188}$$

because $\mathbb{E}_0^{\mathbb{Q}}[\lambda_{0,T} \cdot Y_T r_T^{\mathrm{QE}}] = \mathbb{E}_0^{\mathbb{Q}}[\lambda_{0,T} \cdot R_T^{\mathrm{QE}}/P_T] = 0$ according to the real Euler equation and the absolute value of a correlation is not greater than 1. Assume that $\lambda_{0,T} \leq \bar{\lambda}$. Under the assumption, we have

$$\mathrm{Var}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right] = \mathbb{E}_0^{\mathbb{Q}}\left[\left(\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right)^2\right] - \left(\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right]\right)^2 \leq (\bar{\lambda})^2 \mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]. \tag{IA.189}$$

Suppose $\nu \geq 0$. Applying this to the expression of $\Delta^{\mathrm{QE}} L_0$, we have

$$\begin{aligned}\Delta^{\mathrm{QE}} L_0 \leq & \alpha \bar{\lambda} \sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[\theta_T^{\mathrm{nQE}}\right] \mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]} + \frac{\alpha \bar{\lambda}}{2} \sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[r_T^{\mathrm{QE}}\right] \mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]} \\ & + \frac{\nu \bar{\lambda}}{6} \mathcal{V} \sqrt{\mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]},\end{aligned} \tag{IA.190}$$

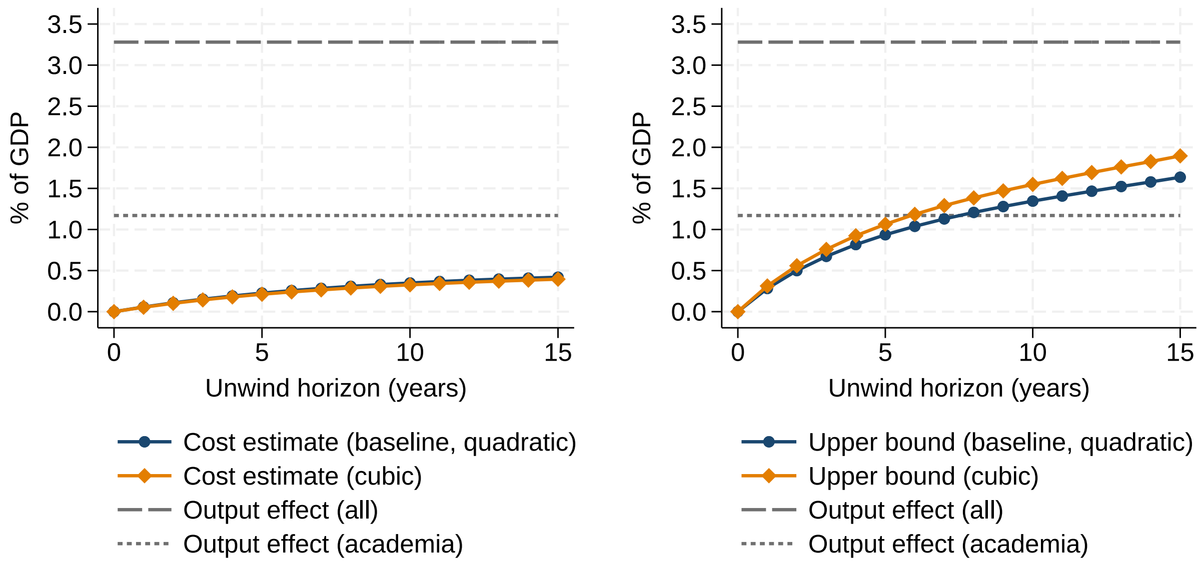
where

$$\mathcal{V} = \left(3\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[(\theta_T^{\mathrm{nQE}} - \theta_0)^2\right]} + 3\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[(\theta_T^{\mathrm{nQE}} - \theta_0) r_T^{\mathrm{QE}}\right]} + \sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[(r_T^{\mathrm{QE}})^2\right]}\right). \tag{IA.191}$$

We calibrate the parameters based on the calibrated marginal excess burden function (MEB) from Saez, Slemrod, and Giertz (2012) and its first- and second-order derivatives. The marginal excess burden (MEB) function of raising revenue through income taxation in the top bracket takes the following form:

$$\mathrm{MEB}(\theta^m) = \frac{e \, a \, \theta^m}{1 - \theta^m - e \, a \, \theta^m}.$$

Under their baseline U.S. calibration ($a = 1.5$, $e = 0.25$, $\theta^m = 0.425$), this implies $\mathrm{MEB}(0.425) \approx 0.38$, meaning that raising an additional dollar of revenue at that margin generates an addi-



(a) Pricing-kernel cost estimate

(b) Upper-bound on cost estimate

Figure IA.11: Aggregate fiscal-efficiency costs: Tax-distortion functions. The figure plots the aggregate exact pricing-kernel cost and conservative upper bound estimate as a function of the unwind horizon T, expressed as a percentage of real GDP at the end of each program's purchase phase (Y_0). The baseline series approximates the loss function with a quadratic polynomial. The alternative series approximates the loss function with a cubic polynomial. The dashed horizontal lines report the output-effect benchmarks from Fabo et al. (2021). Estimates are computed from 1,000,000 term-structure simulations under the $\mathbb{Q}$-measure.

tional efficiency loss of roughly 38 cents. We calibrate the curvature of $\xi(\cdot)$ using the slope of this schedule at the same benchmark rate:

$$\kappa = \xi'(\theta_0) = \mathrm{MEB}(\theta^m), \tag{IA.192}$$

$$\alpha = \xi''(\theta_0) = \mathrm{MEB}'(\theta^m), \tag{IA.193}$$

$$\nu = \xi'''(\theta_0) = \mathrm{MEB}''(\theta^m), \tag{IA.194}$$

This yields $\kappa = 0.38, \alpha = 2.17, \nu = 14.36$. Appendix Figure IA.11 compares the pricing-kernel cost and conservative upper bound estimates in the baseline quadratic loss function specification versus the alternative cubic loss function specification. Our results are robust to the order of approximation we use.