The Fiscal Cost of Quantitative Easing

Adrien d'Avernas[*] Antoine Hubert de Fraisse[†] Liming Ning[‡]
Quentin Vandeweyer[§]

January, 2026

Abstract

Quantitative easing (QE) shortens the duration of the consolidated public balance sheet by swapping long-term government bonds for short, floating-rate liabilities, thereby shifting interest-rate risk onto taxpayers. In segmented bond markets, absorbing duration from the marginal investor can support real activity, but it also generates state-contingent losses that must be financed with distortionary taxes. We quantify the resulting ex ante fiscal-efficiency cost by forecasting QE-portfolio return distributions and mapping these into expected tax deadweight losses under a conservative terminal tax rule. Across all recent U.S. QE programs, the expected costs total 0.24% of GDP under risk-neutral valuation and 0.94% as an upper bound. Relative to published benefit estimates, each QE program appears to have a positive NPV at inception.

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

[*] Stockholm School of Economics

[†] London School of Economics & National Bureau of Economic Research

[‡] Northwestern University Kellogg School of Management

[§] University of Chicago Booth School of Business, quentin.vandeweyer@chicagobooth.edu, 5807 S Woodlawn Ave, Chicago, IL 60637, +17738340691

1 Introduction

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 distortive fiscal adjustments. Under monetary dominance, 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. 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?

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.[2] The consolidated perspective highlights the dual macro-fiscal implications of QE:

[1] Authors' calculations based on SOMA holdings and Treasury yield changes.

[2] An equivalent maturity shortening could be implemented by the Treasury by increasing the share of

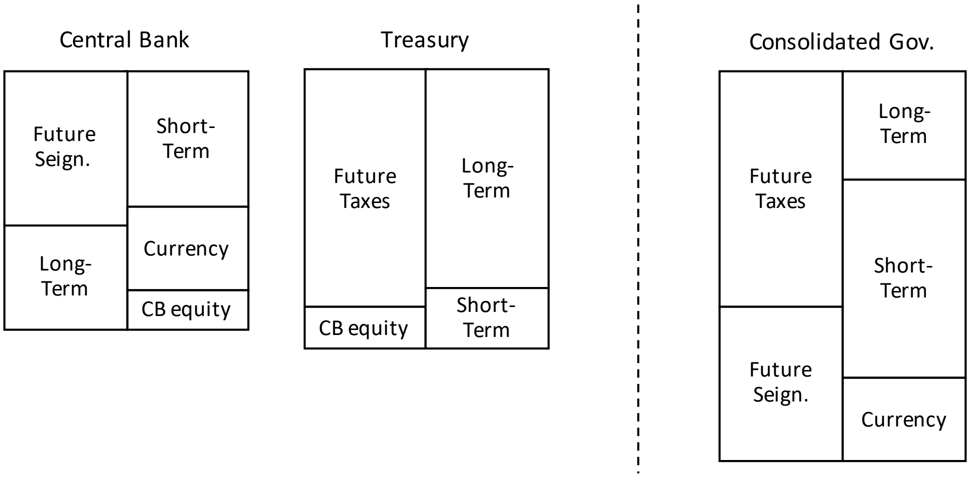


Figure 1: Sketch of Consolidated Government Balance Sheets. The figure sketches the balance sheet of the government in our model, both unconsolidated (left side) and consolidated (right side). The Treasury issues both long-term and short-term debt against agents' future tax liabilities and central bank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

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 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). By reducing the duration risk borne by optimizing bondholders, QE reduces their precautionary saving motive and stimulates aggregate demand. Under consolidated government accounting, the aggregate risk does not disappear when it moves onto the central bank's balance sheet: it is ultimately shifted onto non-participating households through future tax obligations. Because hand-to-mouth households do not offset this reallocation by assumption, the reduction in bondholders' precautionary saving raises demand and output when monetary policy is constrained at the ZLB.

The same maturity shortening, however, increases fiscal refinancing risk. 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

bills relative to bonds—an operation sometimes described as "stealth QE" or "activist Treasury issuance" (Roubini and Miran, 2024).

the dispersion of the government's financing needs and, under monetary dominance, 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 theory to the data, we introduce an implementable fiscal-cost 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. 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.[3] Our cost metric should therefore be interpreted as an upper measure 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). Given the ambiguity of the welfare-relevant pricing kernel in segmented-market environments, we provide two complementary measures: the exact Q-measure estimand evaluated under the stochastic discount factor implied by an affine term-structure model fitted to observed yield curves, and a conservative upper bound that remains valid without fully specifying the pricing kernel.

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 affine term structure model. Under our baseline horizon, we obtain a cumulative risk-neutral expected cost of 0.24% of GDP and an upper bound

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

of 0.94% of GDP. We compare these costs to cumulative output effects attributed to QE in the empirical literature: 3.28% of GDP when considering all studies and 1.17% of GDP when excluding articles written by researchers at central banking institutions (Fabo, Jančoková, Kempf, and Pástor, 2021). Within this conservative fiscal-efficiency accounting, U.S. QE programs appear to have a positive net present value at origination when benchmarked against published output effects.

At the same time, large ex post valuation moves—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, a 95th-percentile negative realization yields a total economic loss for the US economy of 6.71% of GDP across all QE programs. This distinction suggests that QE can be welfare-improving in expectation yet politically contentious when tightening episodes reveal the public sector's duration exposure.

In an extension to our model, we further show that the core trade-off is unaffected by introducing non-interest-bearing currency once the analysis is conducted at the consolidated-government level. This result clarifies recent debates about the economic interpretation of central bank losses and the policies intended to mitigate them. In particular, De Grauwe and Ji (2023) propose that losses could be circumvented by discontinuing interest payments on infra-marginal reserves. Within our framework, however, such a policy would eliminate QE's effectiveness in stimulating aggregate demand: non-interest-bearing reserves function as long-duration liabilities, so replacing one long-term claim with another does not reduce bondholders' exposure to interest-rate risk. More broadly, once the government is consolidated, countercyclical seigniorage affects the level of required distortionary finance but does not overturn the marginal refinancing-risk calculus that governs the optimality condition for QE at the ZLB.

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. In frictionless benchmarks, open-market operations are neutral (Wallace, 1981a), so QE requires limits to arbitrage and/or market segmentation to matter for asset prices and real allocations. Our mechanism is closely connected to modern preferred-habitat and "duration-risk extraction" frameworks (Greenwood and Vayanos, 2014; Vayanos and Vila, 2021; Ray, 2019; Caballero and Simsek, 2020, 2021), with classic foundations in the term-structure and portfolio-balance traditions (Culbertson, 1957; Modigliani and Sutch, 1966; Tobin, 1969). Empirically, large-scale asset purchases affect long-term yields and term premia through local-supply and related channels (Gagnon et al., 2011; Krishnamurthy and Vissing-Jorgensen, 2011; D'Amico and King, 2013; Swanson, 2011; Hamilton and Wu, 2012b). We use the program-level survey of Fabo, Jančoková, Kempf, and Pástor (2021) to summarize estimates of QE output effects.

Recent work interprets the post-2022 wave of central bank losses as having potentially important fiscal consequences that must be weighed against the macro benefits of QE (Cecchetti and Hilscher, 2024; Adrian et al., 2025).

On the theory side, our closest antecedents are papers that study QE when markets are segmented and/or when policy is constrained at the ZLB. Silva (2016) studies optimal QE and risk sharing with financial segmentation but abstracts from distortionary taxation; we instead focus on the trade-off between demand stabilization and the expected deadweight losses generated by financing stochastic QE gains/losses. In a New Keynesian setting with segmented markets but no tax distortion, Abadi (2023) shows that optimal policy uses both rate cuts and asset purchases to stabilize asset prices during downturns. In a New Keynesian setting with tax distortion but no market segmentation, Bhattarai, Eggertsson, and Gafarov (2022) emphasizes that state-contingent central-bank losses can act as a commitment device; in contrast, we study the fiscal consequences of QE-induced duration exposure when the central bank does not condition its policy rule on such losses. Finally, our framework relates to work emphasizing the benefits of maintaining a large balance sheet beyond the ZLB (Vissing-Jørgensen, 2023) and its macroprudential role through bank duration risk and resilience (Eren et al., 2024).

Our paper also relates to the long literature on the optimal maturity of government debt (Barro, 1979; Lucas Jr. and Stokey, 1983; Bohn, 1990). Angeletos (2002) and Buera and Nicolini (2004) study how debt maturity can substitute for state-contingent securities to smooth taxes in the presence of convex distortionary tax costs. Greenwood, Hanson, and Stein (2015b) characterize optimal maturity when tax-smoothing motives interact with a preference for short-term safe claims and crowding out of private liquidity transformation; they show why a government may optimally deviate from perfect tax smoothing toward shorter duration. We build on this insight by introducing a production economy with a ZLB constraint and by interpreting QE as an endogenous maturity-shortening policy that trades off output stabilization against refinancing risk. Corhay, Kind, Kung, and Morales (2023) find that maturity operations have sizable effects on expected inflation and output through a risk transmission mechanism. Our discussion is also related to policy-oriented analyses of maturity management and unconventional policy (Greenwood et al., 2015a; Garbade, 2015), and to recent work highlighting interactions between QE and Treasury issuance and auctions (Ray et al., 2024). Bigio, Nuño, and Passadore (2023) study optimal debt maturity when sovereign issuance entails liquidity costs; we abstract from such issuance frictions and focus on QE-induced maturity shortening and its fiscal-risk implications. Gomez Cram, Kung, Lustig, and Zeke (2025) study the effect of fiscal redistribution risk in a related New Keynesian model with limited participation and show that it makes Treasuries risky and can generate sizable nominal term premia.

Lastly, our paper relates to work on 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 how an insolvent central bank may lose control of inflation absent sufficient fiscal support. Christensen, Lopez, and Rudebusch (2015) propose a stress-test methodology to assess central bank balance-sheet risk following early QE programs. Instead, we maintain monetary dominance and quantify the *ex ante* fiscal-efficiency cost of QE through the expected deadweight losses generated by additional tax dispersion across states induced by QE-related duration exposure.

Relative to the existing literature, this paper makes three contributions. First, we provide a tractable model in which QE is a maturity-shortening policy for the consolidated government and derive an interior optimality condition at the ZLB that equates the marginal output benefit of risk extraction to the marginal fiscal-efficiency cost of refinancing risk. Second, we develop an implementable fiscal-cost 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—and show how it can be computed from the joint distribution of QE portfolio returns and taxes. Third, using granular SOMA holdings data, we quantify this fiscal-risk cost for each U.S. QE program and compare it to program-level output effects estimated in the existing empirical literature.

2 A Stylized 3-period Model

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. 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.[4] We derive the optimal QE scale at the ZLB and demonstrate that this result is robust to the inclusion of zero-interest-bearing currency. All formal proofs

[4] 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.

are relegated to Appendix OA.1.

2.1 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 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 when the ZLB binds. The model comprises bondholders, households, and a government financing its initial spending through debt issuance and taxation. Bondholders include domestic and foreign agents. Both domestic bondholders and households receive consumption goods each period from their capital holdings, but only domestic bondholders actively select consumption and holdings of government bonds. Households are hand-to-mouth, consuming their entire net-of-tax resources each period.

Demography We normalize the total population in the economy to 1, of which a fraction θ consists of domestic bondholders and the remaining $1 - \theta$ 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. 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 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.[5] Thus, taxes to each household are given by $\tau_t^h = \tau_t/(1-\theta)$.

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 $\frac{\alpha}{2}\frac{\tau_t^2}{1-\theta}$.

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.[6] In each period t, the government maximizes

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

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

Sticky Prices 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 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 face the same fixed nominal prices. Hence, the problem of each individual intermediate firm $i \in [0,1]$ is given by

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

[5] In Online Appendix OA.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.

[6] As discussed in 2.7 below, this assumption is made to abstract from market-completion motives.

where $\eta_t(i)$ is the capital utilization rate. Because of nominal frictions, intermediate firms cannot adjust the price of their product. The solution to their 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.

2.2 Agents' Maximization Problem

Domestic Bondholders At time 0, domestic bondholders maximize their lifetime expected value:

$$\max_{c_0^b, c_1^b(s), c_2^b(s), B_0^{S,b}, B_0^{L,b}, B_1^{S,b}(s)} \mathbb{E}_0\left[\sum_{t=0}^{2} \beta^t \frac{(c_t^b)^{1-\gamma}}{1-\gamma}\right], \tag{5}$$

subject to the budget constraint in three periods:

$$c_0^b = y_0 - B_0^{S,b} p_0^S - B_0^{L,b} p_0^L, \tag{6}$$
$$c_1^b(s) = y_1(s) - B_1^{S,b}(s) p_1^S(s) + B_0^{S,b}, \tag{7}$$
$$c_2^b(s) = y_2(s) + B_1^{S,b}(s) + B_0^{L,b}. \tag{8}$$

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. Thus, consumption by foreign bondholders c_t^f in each period is given by

$$c_0^f = -\phi(p_0^S B_0^S + p_0^L B_0^L), \tag{9}$$
$$c_1^f(s) = \phi(B_0^S - p_1^S(s) B_1^S(s)), \tag{10}$$
$$c_2^f(s) = \phi(B_1^S(s) + B_0^L). \tag{11}$$

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

$$c_0^h = y_0 - \frac{\tau_0}{1-\theta} - \frac{\alpha}{2}\left(\frac{\tau_0}{1-\theta}\right)^2, \tag{12}$$

$$c_1^h(s) = y_1(s) - \frac{\tau_1(s)}{1-\theta} - \frac{\alpha}{2}\left(\frac{\tau_1(s)}{1-\theta}\right)^2, \tag{13}$$

$$c_2^h(s) = y_2(s) - \frac{\tau_2(s)}{1-\theta} - \frac{\alpha}{2}\left(\frac{\tau_2(s)}{1-\theta}\right)^2. \tag{14}$$

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 function while satisfying its budget constraint. Importantly, we assume that the government takes bond prices as given, which allows us to abstract from the government's cost reduction motive.[7] Also, the government cannot commit to a prespecified policy rule.[8] Thus, in period 0, the government's problem is given by

$$\max_{\tau_0, B_0^S, B_0^L} \mathbb{E}_0 \left[\sum_{t=0}^{2} \frac{\Lambda_t}{\Lambda_0}\left(y_t - \frac{\alpha}{2}\frac{\tau_t^2}{1-\theta}\right)\right], \tag{15}$$

subject to the budget constraint:

$$G_0 = \tau_0 + B_0^S p_0^S + B_0^L p_0^L. \tag{16}$$

In period 1, the government solves

$$\max_{\tau_1(s), B_1^S(s)} \sum_{t=1}^{2} \frac{\Lambda_t}{\Lambda_1}\left(y_t - \frac{\alpha}{2}\frac{\tau_t^2}{1-\theta}\right), \tag{17}$$

subject to the budget constraint:

$$0 = \tau_1(s) + B_1^S(s) p_1^S(s) - B_0^S. \tag{18}$$

Finally, in period 2, taxes $\tau_2(s)$ are raised to close the budget: $0 = \tau_2(s) - B_1^S(s) - B_0^L$.

[7] 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.

[8] The absence of commitment, which follows from Greenwood et al. (2015b) and Bhattarai et al. (2022), corresponds to a more realistic assumption and yields a simpler solution.

Conventional Monetary Policy The central bank sets short-term interest rates $\{p_0^S, p_1^S(s)\}$ in order to maximize output.[9] Thus, the central bank's problem is given by

$$\max_{p_t^S} \ y_t \qquad \text{for } t = 0, 1. \tag{19}$$

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 whenever possible. In particular, this target is not always admissible when the effective lower bound is binding.

2.3 Equilibrium

We provide a definition for the sequential competitive equilibrium and derive first-order conditions.

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;

(3) Markets for consumption goods, short-term bonds, and long-term bonds clear:

  (a) *consumption*: $\theta c_t^b + (1-\theta)c_t^h + c_t^f = y_t - G_t - \alpha\tau_t^2/2(1-\theta)$;

  (b) *short-term bonds*: $\theta B_t^{S,b} = (1-\phi)B_t^S$ for $t = 0, 1$;

  (c) *long-term bonds*: $\theta B_0^{L,b} = (1-\phi)B_0^L$.

[9] 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.

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

$$p_0^S = \mathbb{E}_0\left[\frac{\Lambda_1(s)}{\Lambda_0}\right], \quad p_0^L = \mathbb{E}_0\left[\frac{\Lambda_2(s)}{\Lambda_0}\right], \quad p_1^S(s) = \frac{\Lambda_2(s)}{\Lambda_1(s)}, \tag{20}$$

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:

$$c_t^h = y_t - \frac{\tau_t}{1-\theta} - \frac{\alpha}{2}\left(\frac{\tau_t}{1-\theta}\right)^2 \qquad \forall t = 0, 1, 2. \tag{21}$$

We then obtain domestic bondholders' consumption through their budget constraints and market-clearing conditions:

$$c_t^b = y_t - \frac{1-\phi}{\theta} G_t + \frac{1-\phi}{\theta}\tau_t \qquad \forall t = 0, 1, 2. \tag{22}$$

The central bank sets short-term rates such that output is maximized in period 1 when the ZLB is assumed not to be binding. It does so by following the Wicksellian prescription of setting the interest rate equal to the natural rate. Therefore, we have $y_1(s) = a_1(s)$ and we set $y_2 = a_2$ to pin down the equilibrium. Combining equations (20) and (22) and defining the short-term rate in period 0 as $r_0 \equiv 1/p_0^S - 1$, we get

$$r_0 = \max\left\{\left(\beta \mathbb{E}_0\left[\left(\frac{\theta a_1(s) + (1-\phi)\tau_1(s)}{\theta a_0 - (1-\phi)(G_0 - \tau_0)}\right)^{-\gamma}\right]\right)^{-1} - 1, 0\right\}, \tag{23}$$

where the max operator reflects that the ZLB is potentially binding. In period 1, we get

$$r_1(s) = \left(\beta\left(\frac{\theta a_2 + (1-\phi)\tau_2(s)}{\theta a_1(s) + (1-\phi)\tau_1(s)}\right)^{-\gamma}\right)^{-1} - 1, \tag{24}$$

since the ZLB can be binding only in period 0.

2.4 The Government's Problem

Period-1 Problem We solve the problem using backward induction and first characterize the solution in period 1. In period 1, because the ZLB is never binding by assumption, monetary policy can set the short-term rate such that 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 (20) and the government's budget constraint, the solution to the government's problem (17) is characterized by tax smoothing:

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s)B_0^L}{1 + p_1^S(s)} \tag{25}$$

and the corresponding short-term bond issuance:

$$B_1^S(s) = \frac{B_0^S - B_0^L}{1 + p_1^S(s)}. \tag{26}$$

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 across states results in higher expected deadweight losses in period zero, 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 debt maturity and level. Specifically, we define the short-term bond share at period 0 as the fraction of short-term debt value, $S = p_0^S B_0^S / D$, where D represents the total debt outstanding at period 0, defined as $D = p_0^S B_0^S + p_0^L B_0^L$. This share S 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. Additionally, we introduce $R_1^c(s) = p_1^S(s)/p_0^L - 1/p_0^S$, which represents 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:*

$$\max_{S,D} \left\{ y_0 - \frac{\alpha}{2(1-\theta)} \left((G_0 - D)^2 + D^2 \left(m \left(S - S^\star\right)^2 + \frac{1}{p_0^S + p_0^L} \right) \right) \right\}, \tag{27}$$

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

In Lemma 1 above, m corresponds to the discounted variance of the bond market carry trade return, 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 by combining (26) with the definitions of (S, D):

$$B_1^S(s) = \frac{B_0^S - B_0^L}{1 + p_1^S(s)} = \frac{D}{1 + p_1^S(s)} \left(\frac{1}{p_0^S} + \frac{1}{p_0^L} \right) (S - S^\star). \tag{28}$$

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.

2.5 Demand Recession at the ZLB

When the ZLB binds in period 0, the nominal short rate cannot fall below zero. In our notation, this implies that the price of the one-period bond is at its upper bound, $p_0^S = 1$ (equivalently, $r_0 = 0$). 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, they do not trade assets and therefore do not internalize intertemporal price changes in their consumption decisions. Domestic bondholders are the marginal intertemporal decision makers, so the short-rate pricing relation (23) 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. Through goods-market clearing, this in turn pins down equilibrium output.

This observation has an immediate implication for debt management at the ZLB. A change in the short-term share S (holding the total market value of debt D fixed) changes the state's fiscal adjustment in period 1 and therefore the state-contingent resources of bondholders in period 1. When the short rate is constrained at the ZLB, equilibrium output must adjust so that bondholders' Euler equation continues to hold.

For subsequent derivations, it is useful to express bondholders' equilibrium consumption directly as a function of the maturity choice (D, S). Combining bond-market clearing with bondholders' budget constraints, the period-0 government budget constraint (16), and the period-1 tax-smoothing policy (25), bondholders' consumption can be written as

$$c_0^b = y_0 - \frac{1-\phi}{\theta} D, \tag{29}$$

$$c_1^b(s) = a_1(s) + \frac{1-\phi}{\theta} D \left[\frac{R_1^c(s)}{1+p_1^S(s)} (S^\star - S) + \frac{1}{p_0^S + p_0^L} \right]. \tag{30}$$

Equation (29) substitutes out period-0 taxes using the government budget constraint: for a given debt value D, bondholders' period-0 consumption co-moves one-for-one with output and is decreasing in the amount of newly issued debt they absorb (their domestic share $(1-\phi)/\theta$). Equation (30) expresses bondholders' period-1 consumption as 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 (29).

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

$$y_0 = \frac{1-\phi}{\theta} D + \left(\beta \mathbb{E}_0 \left[\left(c_1^b(s) \right)^{-\gamma} \right] \right)^{-\frac{1}{\gamma}} \leq a_0. \tag{31}$$

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). 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 maturity policy affects output in a ZLB recession. Differentiating the equilibrium condition in Lemma 2 with respect to the short-term share S yields:

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*

$$\frac{\partial y_0}{\partial S} = \frac{1-\phi}{\theta} D \, \mathbb{E}_0 \left[\beta \left(\frac{c_1^b(s)}{c_0^b} \right)^{-(1+\gamma)} \frac{-R_1^c(s)}{1+p_1^S(s)} \right]. \tag{32}$$

Equation (32) 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 (30)), 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.

We can decompose (32) as

$$\begin{aligned}\frac{\partial y_0}{\partial S} = \frac{1-\phi}{\theta} D \Bigg(& \underbrace{\mathbb{E}_0 \left[\beta \left(\frac{c_1^b(s)}{c_0^b} \right)^{-(1+\gamma)} \right] \mathbb{E}_0 \left[\frac{-R_1^c(s)}{1+p_1^S(s)} \right]}_{\text{intertemporal substitution}} \\ & + \underbrace{\mathrm{Cov}_0 \left(\beta \left(\frac{c_1^b(s)}{c_0^b} \right)^{-(1+\gamma)}, \frac{-R_1^c(s)}{1+p_1^S(s)} \right)}_{\text{precautionary saving}} \Bigg).\end{aligned} \tag{33}$$

The first term captures the effect of the maturity swap on expected discounted resources in period 1. For example, if the expected discounted carry on the long bond is positive, $\mathbb{E}_0\big[R_1^c(s)/(1+p_1^S(s))\big] > 0$, then $\mathbb{E}_0\big[-R_1^c(s)/(1+p_1^S(s))\big] < 0$ and the intertemporal-substitution component is negative. This case corresponds to long-duration bonds commanding a positive term premium.

The second term 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 desired 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 $\partial y_0/\partial S > 0$. Online Appendix OA.2 provides a sufficient condition under which the precautionary-saving channel dominates:

$$\frac{D(S-S^\star)}{a_2} \leq \frac{4\theta}{\gamma(1-\phi)\left(\frac{1}{p_0^S}+\frac{1}{p_0^L}\right)}. \tag{34}$$

Effectively, condition (34) restricts the maturity deviation relative to terminal resources, ensuring that the fundamental component of bondholders' period-1 consumption dominates the endogenous fiscal component induced by refinancing risk (so $c_1^b(s)$ remains increasing in the state). Under this restriction, the precautionary-saving term in (33) is positive and sufficiently large to imply $\partial y_0/\partial S > 0$.[10] In the next section, we assume this condition is satisfied.

2.6 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. Conditional on the level of debt D, it minimizes the dispersion of future taxes across states and therefore minimizes expected deadweight losses from tax distortions. Deviating from $S^\star$ exposes the fiscal authority to refinancing risk—taxes must adjust more strongly to interest-rate realizations—which raises expected distortionary costs.

At the ZLB, however, a maturity deviation also affects equilibrium output through the aggregate-demand channel characterized in Lemma 3. Since p_0^S is pinned down at the ZLB, changes in the state-contingent fiscal adjustment induced by maturity policy shift bondholders' intertemporal consumption choice, and output adjusts to satisfy their Euler equation. The government's optimal maturity therefore balances a fiscal-efficiency cost against a demand-stabilization benefit. Taking the first-order condition with respect to S yields

$$\underbrace{\frac{\alpha m}{1-\theta}\left(S-S^\star\right)D}_{\text{marginal refinancing-risk cost}} \quad = \quad \underbrace{\frac{1}{D}\frac{\partial y_0}{\partial S}}_{\text{marginal output benefit}} \quad . \tag{35}$$

[10] 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.

The left-hand side is the marginal increase in expected tax-distortion costs from a further shortening of maturity relative to $S^\star$. Because tax distortions are quadratic, this marginal cost is proportional to the existing deviation $(S - S^\star)$. It is scaled by α, which governs the curvature of the deadweight-loss function, and by the outstanding debt level D, which determines the fiscal exposure to refinancing shocks. The factor m summarizes how uncertainty along the yield curve translates into dispersion of the government's financing needs when maturity is shortened. Finally, the denominator $(1 - \theta)$ reflects that taxes are levied on households, which implies that a smaller tax base generates larger per-capita tax movements and hence larger quadratic distortionary costs.

The right-hand side is the marginal demand benefit at the ZLB. Increasing S by $\mathrm{d}S$ corresponds to increasing 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. Accordingly, $\frac{1}{D}\frac{\partial y_0}{\partial S}$ 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 $y_0 = a_0$, so the marginal output benefit is zero. In this case, (35) 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 (34)), the optimal deviation from full tax smoothing (or, equivalently, the optimal size of a QE program) is given by*

$$Q^\star \equiv (S - S^\star)D = \min\left\{\frac{1-\theta}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S}, \overline{Q}\right\}, \tag{36}$$

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). Although the model is agnostic about the institutional implementation, we 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 Section 3, we apply Proposition 2 to evaluate the five U.S. QE programs quantitatively. Several modeling choices are made to isolate the paper's central trade-off; we discuss their role in Section 2.7.

2.7 Discussion of the Model's Assumptions

This section clarifies the role of five modeling choices in our results. Together, they isolate the paper's central trade-off: QE-induced maturity shortening stabilizes output at the ZLB by extracting interest rate risk from private balance sheets but increases the dispersion of future fiscal outcomes and thus the expected efficiency cost of distortionary taxation.

Segmentation and limited participation. In frictionless complete-market benchmarks, open-market operations are neutral (Wallace, 1981b). To obtain real effects of maturity transformation, we introduce segmentation by assuming that a subset of agents are hand-to-mouth: they do not trade bonds and have a high marginal propensity to consume out of current income. This assumption provides a tractable way to ensure that shifting interest rate risk away from optimizing bondholders is not undone by offsetting portfolio adjustments of all agents. The key property required is limited participation, rather than the specific hand-to-mouth structure itself.

Government objective function. Our focus is on the trade-off between output stabilization and the fiscal-efficiency cost of rollover risk. Under segmentation, a fully utilitarian planner would also have an additional motive to adjust government risk exposure, providing implicit risk-sharing between households and bondholders through state-contingent tax liabilities. To abstract from this market-completion motive and isolate the maturity-choice margin specifically related to tax smoothing, we evaluate outcomes using bondholders' stochastic discount factor (the valuation kernel of agents pricing government bonds). We refer to Silva (2016) for a detailed analysis of optimal QE policy when such risk-sharing considerations are central. Online Appendix OA.3.2 generalizes this assumption by allowing for general government welfare weights and re-derives our core propositions, explicitly nesting the baseline case.

Price-taking debt policy We assume the government acts as a price taker when deciding on its debt policy. This assumption abstracts from strategic "tax-reduction" motives (assumption similar to Greenwood, Hanson, and Stein, 2015b) that would incentivize the central bank to manipulate bond prices to reduce the government tax burden, thereby introducing additional political-economy considerations beyond our focus. Importantly, this assumption does not preclude equilibrium price responses to maturity changes; it merely removes strategic internalization of these responses. Online Appendix OA.3.3 relaxes price-taking by allowing the government to partially internalize issuance-induced price effects and re-derives our main propositions, recovering the baseline as a limiting case.

Tax incidence and risk redistribution We assume all taxes are collected from households, insulating bondholders from direct tax risk. Maturity shortening thus transfers interest rate risk away from bondholders toward households through the government budget constraint, reinforcing the output-stabilization mechanism. If bondholders bear part of the tax incidence, this redistribution mechanism weakens. Online Appendix OA.3.1 relaxes this assumption by considering general tax incidence distributions. It re-derives the government's problem and associated optimality conditions under dispersed tax burdens, explicitly nesting the baseline assumption. The main insights remain valid provided that domestic bondholders' share of bondholdings exceeds their tax incidence.

Interest-bearing reserves and zero-interest liabilities The baseline model treats the consolidated public sector's short-maturity liabilities as a single instrument and does not distinguish between Treasury bills and reserve balances. This abstraction is appropriate in operating regimes with remunerated reserves (as in the post-2008 U.S. floor system), under which interest-bearing reserves and short Treasury securities are close substitutes and are priced as equivalent short-duration claims. Because recent policy discussions have emphasized non-interest-bearing central bank liabilities—either through currency-driven seigniorage or through proposals to eliminate interest on (required) reserves (De Grauwe and Ji, 2023; Faria-e-Castro and Jordan-Wood, 2023)—Online Appendix OA.3.4 relaxes this equivalence. The extension shows that an exogenous stock of non-interest-bearing currency changes the implementation of tax smoothing through seigniorage but leaves the optimal maturity tilt at the ZLB unchanged once the problem is written in terms of effective liabilities. By contrast, if QE purchases are financed by non-remunerated reserves that bondholders are required to hold, the liability issued to fund purchases is effectively long duration in our three-period environment; the maturity transformation underlying the duration-risk-extraction channel is therefore absent, and the resulting output effect is weakly non-positive.

3 Quantification

In this section, we quantify the fiscal-efficiency cost component of quantitative easing from the consolidated-government perspective formalized in (36). 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 monetary dominance—any realized net income shortfall at the central bank 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. We then compare these fiscal-efficiency cost estimates to program-level output effects from the empirical literature. We study the five U.S. maturity-transformation episodes conducted by the Federal Reserve since 2008: QE1 (2008), QE2 (2010), the Maturity Extension Program (MEP, 2011), QE3 (2012), and QE4 (2020).

3.1 Fiscal-efficiency Cost Framework

Our empirical implementation follows directly from consolidated government accounting and requires only a small set of transparent, maintained assumptions. The guiding consideration is that these assumptions are chosen to (i) isolate the remittances-based fiscal-efficiency cost that is central to our mechanism and (ii) remain conservative.

QE programs Consistent with the stock view of QE (e.g., Bernanke, 2020), QE increases the stock of longer-maturity securities held on the central bank balance sheet; this stock is typically maintained for some time through rollover and reinvestment of principal payments and is financed by short-term liabilities (reserves) (e.g., Board of Governors of the Federal Reserve System, 2011). Our fiscal-cost calculation focuses on the consequences of maintaining this long–short position after net purchases end. Because remittances (and associated fiscal adjustments) are pinned down by realized net returns on the incremental portfolio, an ex-ante fiscal-efficiency evaluation requires specifying how the incremental position is held and financed thereafter.

Let date 0 denote the end of net purchases. We define the incremental QE position at date 0 as the realized change in SOMA holdings over the purchase phase (with corresponding short-term funding on the liability side). From date 0 onward, we assume a passive holding rule that maintains a constant incremental maturity-transformation wedge: principal payments are reinvested and short-term funding is adjusted so that both the *size* and *maturity composition* of the incremental position remain fixed until an exogenous unwind date T. At date T, the position is unwound in a single-date sale and the associated net portfolio value is realized. We report results for alternative values of T to show how fiscal-efficiency costs scale with the length of exposure; our baseline fixes T and uses a single-date exit, while recognizing that alternative exit policies (e.g., gradual runoff) could raise or lower costs depending on the timing and speed of normalization.[11] We argue below that the single-date exit rule is in general conservative.

Terminal financing rule of QE programs To make the remittances channel operational, we adopt a terminal financing assumption: for each QE episode, the consolidated government rolls over short-term funding during the life of the program and then offsets the cumulative net return on the QE portfolio by a one-time tax adjustment at a horizon T. This assumption maximizes the deadweight loss for a given realized fiscal impact because it eliminates tax smoothing. In practice, fiscal adjustments and remittances are typically smoothed (including via deferred-asset accounting), so our estimates should be interpreted as conservative fiscal-efficiency costs under an extreme financing rule.[12]

Two complementary cost measures Within this framework, we report two complementary measures of fiscal-efficiency costs. First, we compute a pricing-kernel cost estimate, which evaluates the expected change in the present value of tax deadweight losses using the stochastic discount factor delivered by an arbitrage-free term structure model. Operationally, this object can be computed under the $\mathbb{Q}$-measure via a change of measure (details below and in Appendix A.1, Appendix OA.6.1). Second, we report a conservative inequality-based upper bound that does not require fully specifying the welfare-relevant pricing kernel. Instead, it only depends on second-order moments of the QE portfolio payoff and the counterfactual fiscal environment.

Scope of the exercise We quantify the fiscal-efficiency cost of QE that operates through the consolidated balance sheet and the remittances channel, that is, the expected efficiency loss from financing QE-induced fiscal gains or losses with distortionary

[11] Such alternatives can be represented as a sequence of dated reductions in the incremental position; their implied fiscal-efficiency costs can be constructed by aggregating the costs of the corresponding passive "stock" positions with different unwind horizons.

[12] This precaution may, however, be warranted as fiscal consolidations may sometimes occur in short periods and through exceptional measures (Leigh, Pescatori, Devries, and Guajardo, 2011).

taxation under a conservative adjustment rule. This object 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. Our focus is therefore narrower but also more directly implementable and tightly disciplined: it isolates a component that any broader welfare evaluation must confront once QE reallocates interest-rate risk onto the public sector. Appendix OA.5 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.

3.2 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.

Government objective and deadweight losses To accommodate the long horizon of QE programs, we extend the notation of Section 2 to an infinite-horizon setting with inflation. Define Y_t as the 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 $\theta_t = \tau_t/(P_tY_t)$ as the tax rate. The period-0 objective of the consolidated government is

$$\max \ \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \Lambda_{0,t} Y_t \left(1 - \xi\left(\theta_t\right)\right) \right]. \tag{37}$$

We take $\xi(\theta_t)$ to represent the real efficiency cost of taxation as a fraction of the real output. In the baseline implementation, we use the quadratic specification

$$\xi\left(\theta_t\right) = \frac{\alpha}{2}\left(\theta_t\right)^2, \tag{38}$$

which is naturally interpreted as a second-order (local) approximation around the average tax rate; we discuss its calibration below.

Counterfactual and notation Let superscript QE denote the path under a QE program and nQE denote the no-QE counterfactual. For any object x_t, define $\Delta^{\text{QE}}x_t \equiv$

$x_t^{\mathrm{QE}} - x_t^{\mathrm{nQE}}$. We define the change in expected tax deadweight losses induced by QE as

$$\Delta^{\mathrm{QE}} L_0 \equiv \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \Lambda_{0,t} Y_t \xi \left(\theta_t^{\mathrm{QE}} \right) \right] - \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \Lambda_{0,t} Y_t \xi \left(\theta_t^{\mathrm{nQE}} \right) \right]. \tag{39}$$

QE as an incremental long–short 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 $\Delta^{\mathrm{QE}} B_0(n)$ and finances the position by issuing one-period zero-coupon bonds $B_0(1)$ in quantity

$$-\Delta^{\mathrm{QE}} B_0(1) = \sum_{n=2}^{N} \Delta^{\mathrm{QE}} B_0(n) \times \frac{p_0(n)}{p_0(1)}, \tag{40}$$

where $p_t(n)$ denotes the date-t price of an n-period nominal zero-coupon bond.[13]

Portfolio rollover rule 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 $\Delta^{\mathrm{QE}} B_t(1)$ (short-term funding) is then chosen each period to finance these trades and roll over outstanding short-term funding. Given this rollover rule, the time-t short-term financing requirement evolves according to

$$-\Delta^{\mathrm{QE}} B_t(1) = -\frac{\Delta^{\mathrm{QE}} B_{t-1}(1)}{p_t(1)} - \Delta^{\mathrm{QE}} B_{t-1}(2) + \sum_{n=2}^{N} \frac{p_t(n)}{p_t(1)} \left(\Delta^{\mathrm{QE}} B_0(n) - \Delta^{\mathrm{QE}} B_{t-1}(n+1) \right). \tag{41}$$

Portfolio unwind, terminal financing, and the fiscal-efficiency cost of QE At date T, the position is unwound in a single-date sale after rebalancing under the portfolio rule. The resulting cash flow from liquidation is R_T^{QE}, the realized time-T net nominal value of the incremental QE portfolio (assets plus the position for one-period):

$$R_T^{\mathrm{QE}} = \sum_{n=1}^{N} p_T(n) \, \Delta^{\mathrm{QE}} B_T(n). \tag{42}$$

[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 $\Delta^{\mathrm{QE}} B_0(n) - \Delta^{\mathrm{QE}} B_{t-1}(n+1)$ additional n-year zero-coupon bonds $B_t(n)$ for all $n > 1$, such that $\Delta^{\mathrm{QE}} B_t(n) = \Delta^{\mathrm{QE}} B_0(n)$ for every period $t \leq T$.

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

$$\Delta^{\mathrm{QE}}\tau_T = -R_T^{\mathrm{QE}}, \qquad \Delta^{\mathrm{QE}}\tau_t = 0 \text{ for all } t \neq T. \tag{43}$$

For the quadratic distortion function, $\xi(\theta) = \frac{\alpha}{2}\theta^2$, the fiscal-efficiency cost admits the following exact decomposition, obtained by expanding $\xi\left(\theta_T^{\mathrm{nQE}} - r_T^{\mathrm{QE}}\right)$ around θ_T^{nQE}:

$$\Delta^{\mathrm{QE}}L_0 = \frac{\alpha}{2}\,\mathbb{E}_0\left[\Lambda_{0,T}Y_T\left(r_T^{\mathrm{QE}}\right)^2\right] - \alpha\,\mathbb{E}_0\left[\Lambda_{0,T}Y_T\,\theta_T^{\mathrm{nQE}}\,r_T^{\mathrm{QE}}\right], \tag{44}$$

where $r_T^{\mathrm{QE}} = R_T^{\mathrm{QE}}/P_TY_T$.

Moreover, we compute the pricing-kernel cost estimate by rewriting (44) under the equivalent $\mathbb{Q}$-measure associated with the money-market numeraire.

Proposition 3 (Pricing-kernel cost estimate). *Let $m_{0,0} = 1$ and define the nominal money-market account discount factor by $m_{0,t+1} \equiv m_{0,t}\,p_t(1)$, where $p_t(1)$ is the one-period nominal zero-coupon bond price defined above. Let $\mathbb{Q}$ be the equivalent martingale (pricing) measure associated with the nominal discount factor $m_{0,t}$. Define $\lambda_{0,T} \equiv m_{0,T} \cdot P_T/P_0$, which is the real discount factor induced by the nominal numeraire.*[15] *Then the fiscal-efficiency cost in (44) can be written as*

$$\Delta^{\mathrm{QE}}L_0 = \frac{\alpha}{2}\,\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_T\left(r_T^{\mathrm{QE}}\right)^2\right] - \alpha\,\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_T\,\theta_T^{\mathrm{nQE}}\,r_T^{\mathrm{QE}}\right]. \tag{45}$$

Equation (45) provides a convenient pricing-kernel representation of the fiscal-efficiency cost of QE, $\Delta^{\mathrm{QE}}L_0$, implied by the terminal financing rule in (43). 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 and $\Delta^{\mathrm{QE}}\tau_T = -R_T^{\mathrm{QE}}$, the QE tax rate satisfies $\theta_T^{\mathrm{QE}} = \theta_T^{\mathrm{nQE}} - r_T^{\mathrm{QE}}$, so positive portfolio returns ($r_T^{\mathrm{QE}} > 0$) reduce the required tax rate at date T, whereas negative returns increase it. In (45), 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 expressed as a fraction of contemporaneous real output.

The first component in (45) is a pure convexity (or Jensen) term arising from QE return dispersion. 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.

[15] For any real payoff X_T (in date-T real units), its date-0 value in real units is $\mathbb{E}_0^{\mathbb{Q}}\left[m_{0,T}\frac{P_T}{P_0}X_T\right]$. See Appendix A.1.2 for more details, or Duffie (2001); Björk (2019) for a general treatment of numeraire invariance.

Intuitively, the marginal excess burden of taxation rises with the tax rate. As a result, a tax increase of a given size in a "bad" state (when taxes are already high) generates more additional distortion than the distortion that is undone by an equally sized tax cut in a "good" state (when taxes are already low). 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 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 expected present value of tax distortions.

The second term of (45) is a state-contingent interaction (or covariance) term. 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 (high θ_T^{nQE} coinciding with high r_T^{QE}), 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 (so 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.

Overall, (45) 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 mechanically costly under convex taxation, whereas the second can attenuate, or even offset, that cost if QE returns provide fiscal insurance.

A key advantage of (45) relative to (44) is that it can be implemented directly from an arbitrage-free term structure model via the risk-neutral measure $\mathbb{Q}$, without simulating the physical-measure stochastic discount factor. Working under $\mathbb{Q}$ also mitigates well-known numerical instability in long-horizon Monte Carlo evaluation of SDF-weighted moments in affine models.

Empirically, we evaluate (45) using an arbitrage-free term structure model estimated using historical Treasury prices as described in Section 3.4. Conceptually, this implies we use the kernel that prices Treasury cash flows as a proxy for the relevant welfare discount factor.

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, this implies that 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 profits.

Conservative upper-bound In addition to the pricing-kernel estimate, we derive a conservative upper bound which follows from Cauchy–Schwarz, and that requires only (i) an upper envelope $\bar{\lambda}$ for the pricing-based real discount factor and (ii) second moments of the QE payoff and counterfactual tax rate, without committing to a particular welfare-relevant asset-pricing model.

Proposition 4 (Conservative bound). *Keeping assumptions in Proposition 3, if in addition $\lambda_{0,T} \le \bar{\lambda}$ state-by-state, then $\Delta^{\mathrm{QE}} L_0 \le \overline{\Delta^{\mathrm{QE}} L_0}$, where the conservative bound is*

$$\overline{\Delta^{\mathrm{QE}} L_0} \equiv \frac{\alpha\bar{\lambda}}{2}\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[r_T^{\mathrm{QE}}\right]\mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]} + \alpha\bar{\lambda}\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[\theta_T^{\mathrm{nQE}}\right]\mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]}. \quad (46)$$

The key observation to derive Proposition 4 is that the incremental QE strategy is constructed to have zero net worth at inception. Under no-arbitrage, this implies that its $\mathbb{Q}$-priced real payoff has zero mean: $\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right] = 0$. Using this identity, the two expectations in (45) can be written as covariances with the priced payoff $\lambda_{0,T} Y_T r_T^{\mathrm{QE}}$:

$$\Delta^{\mathrm{QE}} L_0 = \frac{\alpha}{2}\,\mathrm{Cov}_0^{\mathbb{Q}}\left(r_T^{\mathrm{QE}},\, \lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right) - \alpha\,\mathrm{Cov}_0^{\mathbb{Q}}\left(\theta_T^{\mathrm{nQE}},\, \lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right).$$

The envelope $\lambda_{0,T} \le \bar{\lambda}$ implies

$$\mathrm{Var}_0^{\mathbb{Q}}(\lambda_{0,T} Y_T r_T^{\mathrm{QE}}) = \mathbb{E}_0^{\mathbb{Q}}[(\lambda_{0,T} Y_T r_T^{\mathrm{QE}})^2] \le \bar{\lambda}^2 \mathbb{E}_0^{\mathbb{Q}}[(Y_T r_T^{\mathrm{QE}})^2].$$

Applying Cauchy–Schwarz to each covariance term then yields (46).

Equation (46) is, therefore, an upper-bound in a precise sense: it delivers the maximum fiscal-efficiency cost attainable given the observed second moments and the restriction that $\lambda_{0,T}$ never exceeds $\bar{\lambda}$. The first term in (46) corresponds to the maximal convexity cost from QE-induced tax dispersion, obtained when the priced payoff $\lambda_{0,T} Y_T r_T^{\mathrm{QE}}$ comoves as strongly as possible with the terminal tax shock r_T^{QE}. The second

[16] I.e., $\mathbb{E}_0^{\mathbb{Q}}\left[m_{0,T} \cdot R_T^{\mathrm{QE}}\right] = 0$ (equivalently, in real units, $\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} r_T^{\mathrm{QE}} Y_T\right] = 0$).

term eliminates any potential fiscal-insurance gains by imposing the most adverse alignment between baseline marginal distortions and QE payoffs. In particular, because the cost loads on $-\mathrm{Cov}_0^{\mathbb{Q}}(\theta_T^{\mathrm{nQE}}, \lambda_{0,T} Y_T r_T^{\mathrm{QE}})$, the worst case occurs when this covariance is as negative as allowed by second moments, 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 "amplifying" scenario in which the maturity-transformation payoff provides no insurance against fiscal stress and instead amplifies it to the maximum extent.

The condition $\lambda_{0,T} \leq \bar{\lambda}$ is a tail-regularity restriction on long-horizon real state prices. 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 (45). For our baseline $T = 10$-year horizon, we set $\bar{\lambda} = 2$ so that it trims only extreme tail draws. In the Monte Carlo paths estimation used for our program-level forecasts, approximately 99.5% of paths satisfy this bound (Appendix B.4). This choice is also conservative relative to historical realizations. For example, in historical data, the real return of one-year nominal T-bills is by large higher than -3% over the relevant window, except for the extreme cases in 2021 (-6.7%), 2022 (-5.9%) and 1974 (-4.4%) when the inflation spiked. This worst-case scenario implies a $\lambda_{0,10} \leq \exp(0.067 + 0.059 + 0.03 \times 8) \approx 1.44 < 2$.[17]

3.3 Data

SOMA holdings We measure each program's incremental asset purchases using the Federal Reserve's System Open Market Account (SOMA) holdings data, which report weekly security-level positions in U.S. Treasury securities, agency debt, and agency mortgage-backed securities (MBS).[18] In Appendix OA.4, we provide a brief institutional description of these programs and their timing, which defines the QE episodes used in the empirical analysis.

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 and report smaller pricing errors compared

[17] The real return of one-year nominal T-bills is calculated as the one-year nominal T-bill yield minus the realized inflation (measured as log CPI growth) over the next year. The nominal yield data is from Liu and Wu (2021), and the CPI data is from the U.S. Bureau of Labor Statistics.

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

to Gürkaynak, Sack, and Wright (2007). 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 percent of GDP from FRED (series `FYFRGDA188S`). For scaling and robustness checks, we also retrieve nominal annual GDP from FRED (series `GDPA`) and CPI from U.S. Bureau of Labor Statistics (All Urban Consumers, Current Series).

3.4 Measurement

To compute the pricing-kernel cost estimate (45) and the upper bound (46), we construct, for each episode, the date-0 conditional joint distribution of the terminal QE payoff and the fiscal environment at the unwind date. Concretely, for a holding horizon T, we require the joint distribution of $\left(r_T^{\text{QE}}, P_T, Y_T, \theta_T^{\text{nQE}}\right)$ as well as the convexity parameter α. 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 (required for the pricing-kernel cost estimate and the upper bound). Third, we calibrate our convex cost function using standard parameter values from the public finance literature. Finally, we also compile estimates from existing literature on the output-stabilization effects of QE.

Construction of QE portfolios For each QE program, the initial QE portfolio is a zero-net-worth long–short position: the central bank is long the long-maturity securities acquired during the program's purchase phase, $\Delta^{\text{QE}}B_0(n)$, and short the one-year bonds issued in quantity $-\Delta^{\text{QE}}B_0(1)$ to finance these acquisitions.[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

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

[20] Although our model 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.

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 measure 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 contain comparable information on the timing of principal payments and prepayment behavior. We therefore conservatively approximate the maturity structure of MBS cash flows using that of Treasury and agency-debt holdings in aggregate, which treats MBS cash flows as more back-loaded than in reality and hence overstates portfolio duration and interest-rate exposure.[21]

Table 1 reports program size in face value, total promised payments (principal plus coupons), and 10-year duration equivalents. QE1, QE3, and QE4 combine Treasury, agency, and MBS net purchases, whereas QE2 consists only of Treasury purchases; the MEP is close to zero in face-value terms but sizable in duration-equivalent units, reflecting the maturity swap.[22]

Term structure model and simulations To measure the ex-ante fiscal-efficiency cost of each QE episode, we estimate a three-factor arbitrage-free affine term structure model 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 (2012a), 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 A.1.[23] Starting from the end-of-purchases state, we then simulate 1,000,000 paths of future factor realizations and construct the implied yield-curve paths needed to evaluate holding-horizon outcomes up to 20 years. The estimated affine model delivers both the physical transition of the factors and the risk-neutral dynamics under the nominal money-market numeraire. Unless otherwise noted, we compute the

[21] MBS cash flows are typically front-loaded because of amortization and embedded prepayment options. Over November 2008–March 2022, Bloomberg's effective-duration measures average 4.58 years for MBS and 5.9 years for Treasuries.

[22] Small differences between program size (in face-value terms) and the purchase amounts described above arise because part of the announced amounts was used to replace maturing securities rather than to expand net holdings.

[23] 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 (45) and (46).

	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 1: 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 share 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.

moments entering (45) and (46) under the model-implied $\mathbb{Q}$-measure; we use the physical dynamics only for forecast diagnostics. Appendix B.3 illustrates the physical-measure distribution of 10-year-ahead forecasts for the 1-year and 10-year yields at the end of QE1 and QE4 (our baseline holding horizon is 10 years), together with realized yields. The model implies slow-moving median paths with substantial tail dispersion.

For macroeconomic variables—including tax rates, real GDP growth, and inflation—we build forecasting models based on the latent factors of the term structure model and then generate 1,000,000 paths for these variables using the corresponding factor paths. These simulations deliver the horizon-T nominal GDP level P_TY_T used to form $r_T^{\text{QE}} = R_T^{\text{QE}}/(P_TY_T)$ and the no-QE tax rate θ_T^{nQE}. Details of the time-series specifications are relegated to Appendix OA.6.1.

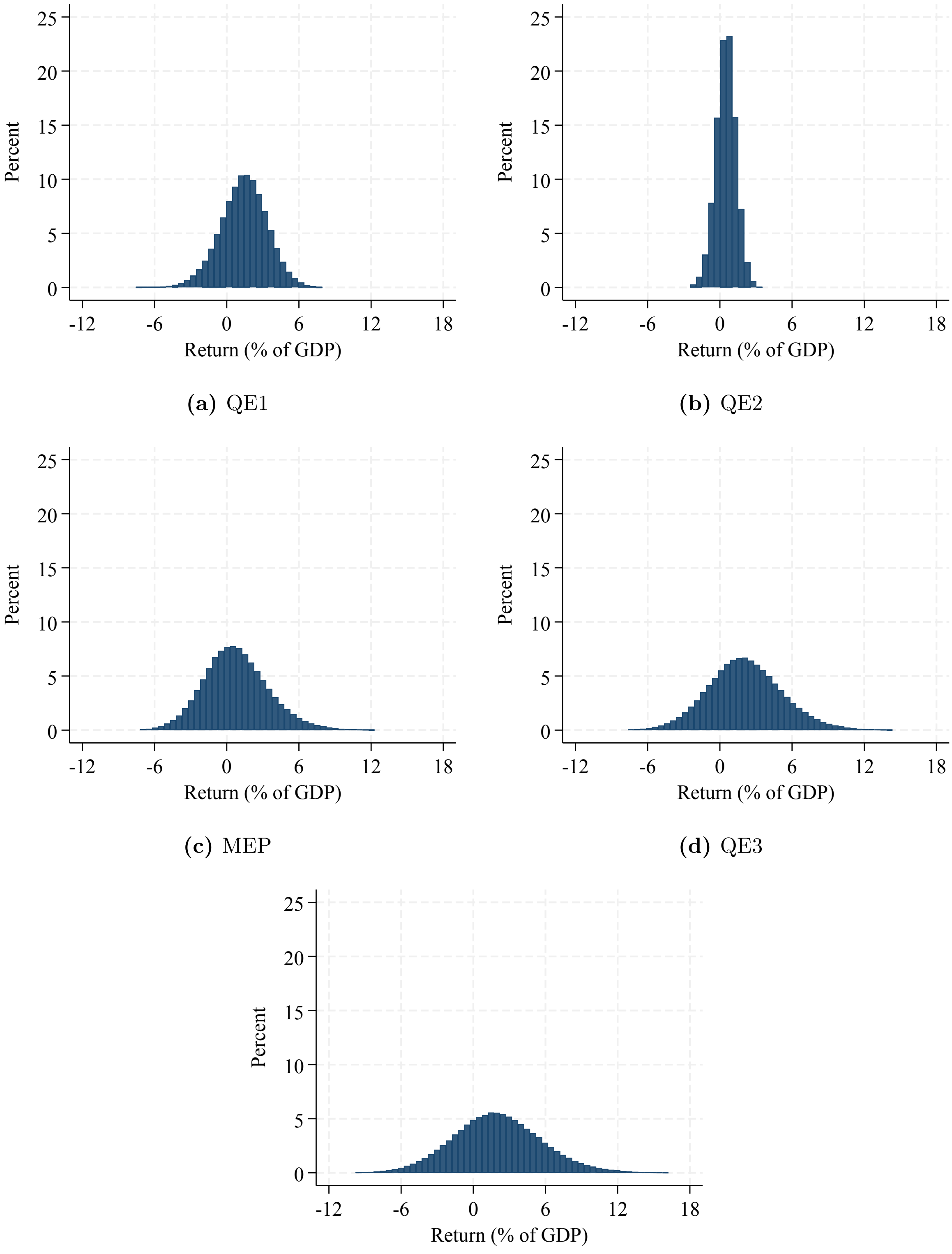
Implementation on the model maturity grid In the baseline specification, the longest maturity consistently available over our estimation sample is 15 years. We therefore estimate the term structure model on maturities up to 15 years and implement QE portfolio cash flows on the same annual grid. Each 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. We compute the associated short-term financing requirement $-\Delta^{\mathrm{QE}}B_t(1)$ using (41). 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.

QE portfolio returns We compute the terminal QE payoff R_T^{QE} by applying (42) (with the rollover rule implemented via (41)) to each simulated yield-curve path. For interpretability, we report R_T^{QE} scaled by nominal GDP at the end of the purchase phase, P_0Y_0. We set our baseline unwinding horizon to $T = 10$ years, which we interpret as a conservative exposure window. In duration-equivalent terms, Appendix Figure OA.4.1 implies that cumulative QT1 reductions matched the QE1 stock (in 10-year duration equivalents) by August 2018 (about 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 (about 11.5 years after QE2 net purchases ended). Survey-based expectations contemporaneous to the implementation of each program typically implied shorter reinvestment and runoff horizons.[24] Our baseline assumes a discrete unwind at T rather than gradual passive runoff; by maintaining the maturity-transformation wedge until the unwind date, this convention tends to widen the distribution of R_T^{QE} relative to more gradual exit policies. For robustness, we also report results for alternative holding horizons below. Figure 2 illustrates the resulting distribution for the five programs at our baseline unwinding of $T = 10$.

Distortionary tax function calibration In the baseline, we use a quadratic function for ξ, which can be interpreted as a second-order (local) approximation to the efficiency cost of taxation around the steady state. To pin down α, we target a standard benchmark estimate of the marginal excess burden (MEB) of raising an additional dollar of revenue through income taxation.[25] Because taxes are modeled as a proportional wedge $\theta_t = \tau_t/(P_tY_t)$, a marginal change $d\theta$ raises revenue by $P_tY_t\, d\theta$ and raises deadweight losses by

[24] 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. 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.

[25] Throughout, we use "deadweight loss" in the standard public-finance sense of *excess burden*: the welfare cost of raising a given amount of revenue with distortionary taxes relative to a non-distorting (lump-sum) benchmark. While higher taxes may increase leisure, the excess burden reflects that choices are made at distorted prices, so raising revenue through a wedge entails a welfare loss over and above the transferred dollars.



(a) QE1

(b) QE2

(c) MEP

(d) QE3

(e) QE4

Figure 2: Distribution of QE portfolio returns. This figure plots the physical distribution of incremental nominal QE portfolio returns at a 10-year holding horizon ($T = 10$) as a share of nominal GDP measured at the year-end following the end of net purchases (P_0Y_0) for each program. For each episode, the unwind payoff R_T^{QE} is computed from (42) along each of 1,000,000 simulated yield-curve paths under physical measure. Returns are trimmed at the 0.1% level in both tails.

$P_tY_t\,\xi'(\theta_t)\,d\theta$. Thus, the marginal excess burden per additional dollar of revenue is locally $\xi'(\theta)$, which we discipline using an empirical benchmark.

Saez, Slemrod, and Giertz (2012) provide a sufficient-statistics expression for the MEB of a small increase in the *marginal* tax rate in the top bracket. In their framework, MEB depends on (i) the elasticity of taxable income e, (ii) the combined top-bracket marginal tax rate θ^m, and (iii) the Pareto parameter a that characterizes the thickness of the top tail of the income distribution. They show that the marginal excess burden per dollar of additional revenue is

$$\text{MEB} \equiv -\frac{dB}{dR} = \frac{e\,a\,\theta^m}{1-\theta^m-e\,a\,\theta^m}, \tag{47}$$

and the corresponding marginal efficiency cost of funds is $\text{MECF} = 1 + \text{MEB}$.

Using U.S. data for the top 1% income cutoff, Saez, Slemrod, and Giertz (2012) take $a \simeq 1.5$, a combined marginal tax rate of $\theta^m = 42.5\%$ (federal, state, Medicare, and typical sales taxes, net of deductibility interactions), and an elasticity estimate of $e = 0.25$, which they describe as mid-range in the taxable-income elasticity literature. Under these values, behavioral responses offset 27.7% of the mechanical revenue gain, implying $\text{MECF} \approx 1.38$ and hence $\text{MEB} \approx 0.38$: raising an additional \$1 of revenue entails an efficiency cost of roughly 38 cents.

While θ^m is a top-bracket marginal rate, our reduced-form wedge θ_t summarizes the effective distortionary adjustment used to finance the consolidated government budget. We therefore use the Saez–Slemrod–Giertz estimate as a benchmark for the marginal efficiency cost of raising revenue and impose it as the local slope of $\xi(\theta)$ at the steady state:

$$\xi'(\bar{\theta}) = 0.38. \tag{48}$$

Given our steady-state calibration $\bar{\theta} = 0.25$, this implies $\alpha = 0.38/\bar{\theta} = 1.52$.

Two considerations guide this choice. First, the Saez–Slemrod–Giertz benchmark is anchored in the top bracket where marginal tax rates are high; as a result, it plausibly provides an upper-end (conservative) benchmark for the marginal excess burden of broad-based fiscal adjustments. Second, the taxable-income elasticity aggregates real behavioral responses and avoidance margins. To the extent that part of the taxable-income response reflects re-timing or shifting with limited real resource costs, the welfare-relevant marginal excess burden could be smaller (see, e.g., Chetty, 2009). At the same time, the sufficient-statistics calculation abstracts from marginal administrative and compliance costs, which would push in the opposite direction. For these reasons, we treat $\xi'(\bar{\theta}) = 0.38$ as a disciplined conservative benchmark and examine alternative values in sensitivity analysis.

Output effect We measure QE-program-level output effects using the 18 U.S. studies surveyed by Fabo, Jančoková, Kempf, and Pástor (2021). For each study, we take the cumulative output response at the horizon reported by the authors and express it as a percent of base-year GDP (see Appendix D). When a study reports a joint output effect for multiple QE episodes, we allocate the total effect across programs in proportion to program size measured in 10-year duration equivalents. We then compute a size-normalized effect (output per unit of duration-equivalent purchases), trim the size-normalized estimates below the 5th percentile and above the 95th percentile, and average the remaining estimates for QE1, QE2, and QE3 separately. Because the MEP and QE4 are not directly covered in the surveyed set, we impute their output effects by assuming the pooled size-normalized effect applies and scaling by the respective program size. Details and the full survey list are in Appendix D.[26]

Fabo, Jančoková, Kempf, and Pástor (2021) document systematic differences across study characteristics: papers with central-bank-affiliated authors report larger QE effects than academic-only papers, although both groups imply positive output effects. For QE1, the average estimated output effect across all studies is 0.36% of GDP, compared with 0.18% when restricting to academic-only studies. We therefore report both benchmarks. We also report results by methodology as classified by Fabo, Jančoková, Kempf, and Pástor (2021). For QE1, the average output effect is 0.18% of GDP in DSGE-based studies and 0.44% of GDP in VAR-based studies.

Overall, the cross-study average implies that purchases equal to 1% of GDP in 10-year duration equivalents increase output by 0.6% of GDP. The implied effect is smaller in academic-only studies (0.2%) and in DSGE studies (0.3%), and larger in VAR studies (0.8%). More recent papers published after the Fabo, Jančoková, Kempf, and Pástor (2021) survey deliver effects of comparable (or even larger) magnitude once expressed in the same units; we summarize back-of-the-envelope conversions in Appendix D.

Timing and discounting Our fiscal-efficiency cost is a terminal-T object: under (43), QE affects distortionary taxation only at the unwind date T, and we report the associated date-0 present value. By contrast, the output effects reported in the QE literature surveyed by Fabo, Jančoková, Kempf, and Pástor (2021) describe gains along the output

[26] The output effects reported in the empirical literature are predominantly short-horizon effects identified from near-term yield and activity responses around purchase announcements (and hence are conceptually tied to periods near the ZLB), whereas the fiscal-efficiency cost we quantify is a long-horizon object that mechanically grows with the holding horizon through the accumulation of interest-rate risk. Moreover, the studies surveyed by Fabo, Jančoková, Kempf, and Pástor (2021) largely predate 2018. We therefore treat these estimates as disciplined benchmarks and examine sensitivity across study classifications.

path and may accrue over a range of horizons. To make the cost–benefit comparison conservative when output responses are front-loaded relative to T (e.g., $T = 10$ years in our baseline), we discount output gains using the most stringent real risk-free discount factor up to maturity T, $\underline{p}_0^r(T) \equiv \min_{t\in\{1,\ldots,T\}} p_0^r(t)$, where $p_0^r(t)$ is the date-0 price of a real risk-free claim paying one unit of consumption at date t. This choice ensures that any benefit accruing at $t \leq T$ is assigned a (weakly) lower present value than discounting at its effective horizon.[27]

Accordingly, we compare present values at date 0: fiscal-efficiency costs are evaluated as expected distortions at date T discounted back to date 0, while output benefits are discounted using $\underline{p}_0^r(T)$. In our extreme-realization analysis, we compare tail outcomes for R_T^{QE} to output-effect benchmarks without additional discounting; because fiscal impacts occur at horizon T whereas output gains tend to occur earlier, this convention mechanically tilts the comparison toward larger measured costs when output effects are concentrated before T.

3.5 Cost estimates and cost–benefit analysis

This section reports fiscal-efficiency costs of U.S. QE programs under the terminal financing rule in (43) and benchmarks them against program-level output effects from the empirical literature. The fiscal-efficiency object is the ex ante change in the present value of deadweight losses from distortionary taxation implied by financing QE gains and losses. Throughout, we express both costs and output-effect benchmarks as a percent of real GDP at the end of the purchase phase (Y_0). We first present baseline estimates for a 10-year holding horizon ($T = 10$), then decompose costs into a return-dispersion component and a tax-covariance (fiscal-insurance) component, and finally examine how costs scale with the unwind horizon.

Baseline magnitudes Table 2 reports fiscal-efficiency costs for each QE episode. For $T = 10$, the pricing-kernel estimate ranges from 0.01% to 0.09% of GDP across programs, with an aggregate cost of 0.24% of GDP. The conservative upper bound is larger by construction—0.06% to 0.29% program-by-program and 0.94% in the aggregate—because it imposes both a worst-case envelope on the pricing weight and a worst-case treatment of comovement. Rescaling those estimates to a standard purchase size of 10 percentage points of GDP in 10-year duration equivalents implies an aggregate pricing-kernel cost

[27] Empirically and conceptually, 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 (2016a) and Kim et al. (2020).

	QE1	QE2	MEP	QE3	QE4	All
Panel A: Estimated Cost (% of GDP)						
Pricing-Kernel Estimate	0.03	0.01	0.06	0.06	0.09	0.24
Upper-Bound Estimate	0.16	0.06	0.21	0.23	0.29	0.94
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 2: Fiscal-efficiency costs and output effects of QE programs. This table reports the pricing-kernel fiscal-efficiency cost estimate (45), the conservative upper bound (46), and output-effect benchmarks from Fabo, Jančoková, Kempf, and Pástor (2021), all expressed as a percent of real GDP at the end of the 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 3.4. QE return $\mathbb{Q}$-measure distributions are computed from 1,000,000 term-structure simulations. The upper bound implements the state-by-state envelope $\lambda_{0,10} \leq 2$.

of about 5 basis points of GDP (0.05%) and an aggregate conservative upper bound of about 18 basis points of GDP (0.18%).

Cost heterogeneity across programs is economically meaningful. It reflects differences in purchase size and duration composition, as well as differences in the interest-rate environment at origination (in particular, the distribution of future rate changes and the degree of term-premium compensation embedded in long-duration claims).

Benchmarking against output effects The costs in Table 2 are small relative to output-effect benchmarks. Aggregating across programs, the output effects implied by Fabo, Jančoková, Kempf, and Pástor (2021) range from 1.17% of GDP (academic-only studies) to 4.22% (VAR studies), with 3.28% in the full-study average. Thus, at the baseline horizon, even the conservative upper bound on the fiscal-efficiency cost of the historical programs remains below the lower end of the discounted output-effect range. This comparison should be interpreted as a benchmark using comparable units rather than as a complete welfare calculation; our fiscal-efficiency measure isolates the distortionary-financing implications of QE under the terminal financing rule, but does not account for potential additional channels operating through broader macroeconomic feedback mechanisms (see Appendix OA.5).

Return dispersion versus tax covariance As discussed above, the pricing-kernel cost in (45) can be decomposed into a *return-dispersion* component and a *tax-covariance* component. The return-dispersion component captures the efficiency cost created by dispersion in the unwind payoff r_T^{QE} when tax distortions are convex in the tax rate. In other words, it captures fluctuations in the terminal tax adjustment that raise expected distortions through Jensen's inequality. The tax-covariance component captures the extent to which QE payoffs provide fiscal insurance. If QE pays off in high-tax states (positive covariance with the counterfactual tax rate θ_T^{nQE}), it reduces distortions by delivering resources when the counterfactual tax rate is high; if QE losses occur in high-tax states (negative covariance), it increases distortions by forcing larger tax adjustments precisely when the counterfactual tax rate is already high.

Table 3 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 (0.15% of GDP from return dispersion and 0.09% from tax covariance). The positive tax-covariance contribution means that the covariance term in (45) is unfavorable in sign: because it enters as $-\mathrm{Cov}(\theta_T^{\mathrm{nQE}}, \lambda_{0,T} Y_T r_T^{\mathrm{QE}})$, a positive contribution corresponds to a *negative* underlying covariance. Economically, QE exhibits "wrong-way" fiscal exposure in this benchmark: payoffs are lower in states in which the counterfactual tax rate is higher.

Because the conservative bound applies Cauchy–Schwarz, it is tight only under an extreme comovement configuration. In particular, the tax-covariance piece of (46) corresponds to the worst-case alignment in which QE payoffs are lowest exactly when the counterfactual tax rate is highest (i.e., $\mathrm{Corr}(\theta_T^{\mathrm{nQE}}, \lambda_{0,T} Y_T r_T^{\mathrm{QE}}) = -1$). This worst-case assumption mechanically amplifies the tax-covariance contribution and explains why the bound is dominated by that component (0.66% of GDP versus 0.28% from return dispersion). Accordingly, the dominance of the covariance term in the bound should be read as an inequality-driven envelope rather than as a statement about the most likely pattern of comovement.

Costs over alternative holding horizons Figure 3 plots aggregate fiscal-efficiency costs as a function of the unwind horizon T (the date at which the terminal tax adjustment is applied under (43)). Holding the initial QE portfolio fixed, a longer 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 comove-

	QE1	QE2	MEP	QE3	QE4	All
Panel A: Pricing-Kernel Estimate (% of GDP)						
Estimate	0.03	0.01	0.06	0.06	0.09	0.24
Return Dispersion Term	0.02	0.00	0.03	0.04	0.06	0.15
Tax Covariance Term	0.01	0.01	0.02	0.02	0.03	0.09
Panel B: Upper-Bound Estimate (% of GDP)						
Estimate	0.16	0.06	0.21	0.23	0.29	0.94
Return Dispersion Term	0.04	0.01	0.06	0.07	0.11	0.28
Tax Covariance Term	0.12	0.05	0.15	0.16	0.18	0.66

Table 3: Decomposition of fiscal-efficiency costs. This table decomposes the pricing-kernel cost estimate (45) and the conservative upper bound (46) into a *return-dispersion* component and a *tax-covariance* component, expressed as a percent of real GDP at the end of the purchase phase (Y_0). The holding horizon is $T = 10$ years. Moments are computed from 1,000,000 term-structure simulations under $\mathbb{Q}$-measure. The upper bound uses the envelope $\lambda_{0,10} \leq 2$ and Cauchy–Schwarz inequalities, which are tight under extreme comovement.

ment. For comparison, the dashed line reports the discounted output-effect benchmark from Fabo, Jančoková, Kempf, and Pástor (2021). This benchmark is constant in T because the underlying estimates account for the entire discounted benefits at T_0, whereas the remittances-based fiscal exposure accumulates over the duration during which the maturity mismatch is maintained. Up to horizons of approximately 15 years, even the conservative bound remains below the lower range of these output-effect benchmarks; for longer horizons, the two become comparable, with the upper bound crossing the benefit estimate around the 16-year mark.

Figure 4 decomposes the aggregate cost into *return dispersion* and *tax covariance* components. The decomposition is broadly stable across horizons: return dispersion rises mechanically with T, while the covariance component remains economically meaningful, reflecting persistent comovement between QE payoffs and the counterfactual fiscal environment.

Interpretation Figures 3–4 summarize how remittances-driven fiscal exposure evolves with the unwind horizon and how it splits into return dispersion and payoff–tax comovement. For the historical U.S. programs and unwind horizons on the order of a decade, the implied fiscal-efficiency costs are modest relative to standard output-effect benchmarks. Taking these external output-effect estimates at face value, the historical programs there-

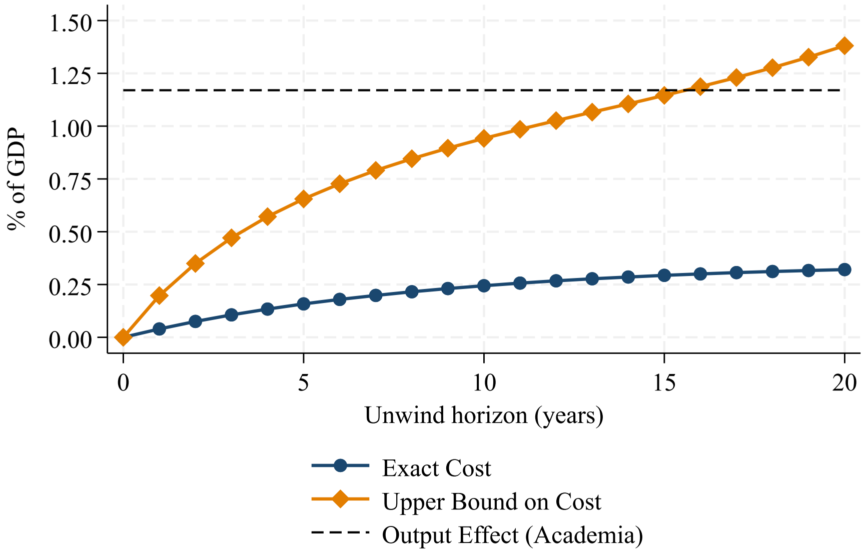
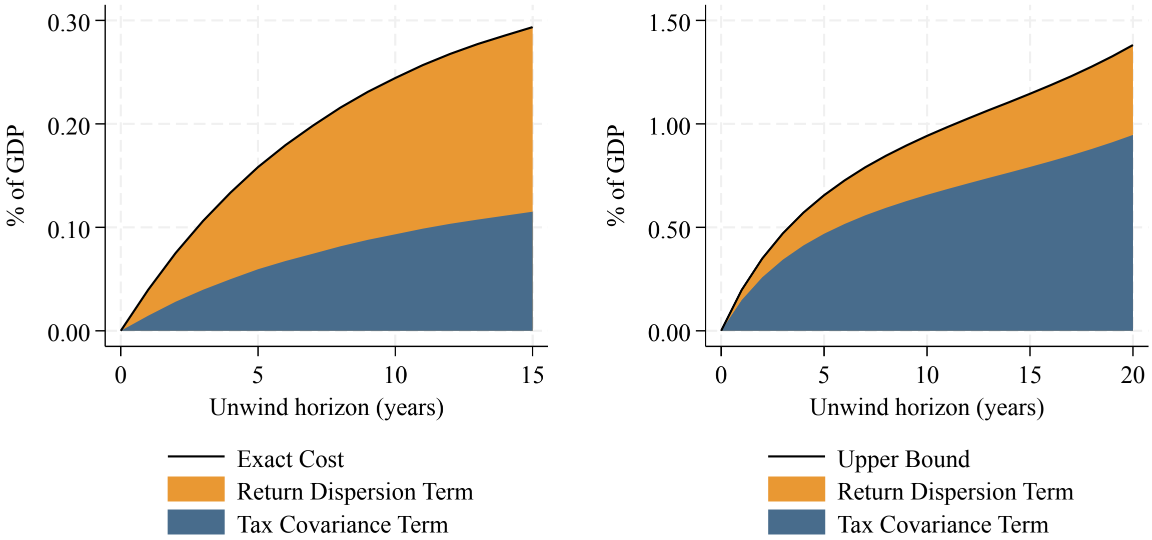


Figure 3: Aggregate fiscal-efficiency costs versus unwind horizon. The figure plots the aggregate pricing-kernel estimate (45) and the conservative upper bound (46) as a function of the unwind horizon T, expressed as a percent of real GDP at the end of the purchase phase (Y_0). The dashed horizontal line reports the discounted output-effect benchmark from Fabo, Jančoková, Kempf, and Pástor (2021) described in Section 3.4. Costs are computed from 1,000,000 term-structure simulations under $\mathbb{Q}$-measure. The conservative bound is computed under the envelope $\lambda_{0,T} \leq 2$.



(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. Panel (a) reports the decomposition of the pricing-kernel estimate (45). Panel (b) reports the decomposition of the conservative upper bound (46). Both panels are based on 1,000,000 term-structure simulations under $\mathbb{Q}$-measure. The conservative bound is computed under the envelope $\lambda_{0,T} \leq 2$.

fore pass a simple cost–benefit check within the consolidated-government criterion in (36). At the same time, the small number of U.S. QE episodes implies that our evidence is best interpreted as speaking to average costs and benefits across the historical programs, rather than to the marginal welfare effect of scaling QE at the margin.

The gap between the pricing-kernel estimate and the conservative bound clarifies what is identified. The bound is constructed to remain valid under a worst-case envelope for discounting and an adverse comovement configuration, and it is therefore tight only under an extreme alignment in which QE payoffs are systematically lowest in high-tax states. A substantially smaller pricing-kernel estimate indicates that the joint distribution implied by the data is far from this worst-case configuration; in particular, payoff–tax comovement is not close to maximally adverse.

More broadly, the estimates highlight that remittances-based fiscal-efficiency costs are governed by two empirical objects. The first is the exposure of the incremental portfolio to interest-rate risk—summarized by the unwind horizon T and by the size and duration of the incremental portfolio (the extent of maturity transformation)—which drives the return-dispersion component.

Sensitivity analyses We perform sensitivity analyses along two dimensions. First, we recompute the aggregate pricing-kernel cost using alternative term-structure parameters, replacing our preferred calibration with the model parameterization from Liu and Wu (2021). The resulting cost profile is somewhat higher across unwind horizons, but the overall shape is very similar to the baseline (Appendix Figure B.5). Second, we vary the tax-loss convexity parameter α when computing deadweight-loss costs (and the conservative bound $\overline{\Delta^{\mathrm{QE}} L_0}$). Holding the term-structure objects fixed, α enters multiplicatively in (45) and (46), so costs scale one-for-one with α; Appendix Figure B.6 illustrates this proportional rescaling for $\alpha \in \{1, 2\}$ around the baseline calibration $\alpha = 1.52$.

3.6 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 realized 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 realized fiscal losses under the maintained terminal-financing rule.

From portfolio payoffs to realized losses We map a given payoff realization into two components with distinct interpretations. The first is a net real transfer to foreign bondholders. The second is an efficiency loss due to distortionary taxation. We report both because the transfer component captures a leakage of resources from domestic residents to foreign holders in adverse states, while the deadweight-loss component captures the efficiency cost of financing QE gains and losses via taxes.[28]

Transfers to foreigners QE can generate net transfers to foreign bondholders because the central bank absorbs duration risk that, absent QE, would have been borne in part by foreign investors when U.S. interest rates rise. Let ϕ denote the fraction of the incremental position effectively absorbed on behalf of foreign holders (an incidence parameter rather than the literal fraction of trades executed with foreign counterparties).[29] The implied terminal net real transfer to foreigners is

$$\Delta^{\mathrm{QE}} F_T \;=\; Y_T\big(-\phi\, r_T^{\mathrm{QE}}\big). \tag{49}$$

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 expansion of the deadweight-loss function around the counterfactual tax rate θ_T^{nQE}, terminal deadweight losses are

$$\Delta^{\mathrm{QE}} L_T \;=\; Y_T\left(-\xi'\left(\theta_T^{\mathrm{nQE}}\right) r_T^{\mathrm{QE}} + \frac{1}{2}\xi''\left(\theta_T^{\mathrm{nQE}}\right)\left(r_T^{\mathrm{QE}}\right)^2\right). \tag{50}$$

Under the quadratic specification $\xi(\theta) = \frac{\alpha}{2}\theta^2$, we have $\xi'(\theta) = \alpha\theta$ and $\xi''(\theta) = \alpha$. Total realized losses are the sum of the two components:

$$\Delta^{\mathrm{QE}} TC_T \;=\; \Delta^{\mathrm{QE}} F_T + \Delta^{\mathrm{QE}} L_T. \tag{51}$$

Foreign-share calibration We set ϕ to the share of foreign holdings of U.S. Treasury securities around each episode's purchase phase, and apply this share to the incremental QE position as a maintained approximation.[30] Table 4 reports the resulting values and

[28] From a global-welfare perspective, transfers net out; in that case the efficiency component $\Delta^{\mathrm{QE}} L_T$ is the relevant welfare cost. We nevertheless report $\Delta^{\mathrm{QE}} F_T$ because the consolidated-government perspective in our framework is naturally interpreted as domestic.

[29] By construction, the incremental QE portfolio has zero initial value, so its pricing-kernel discounted expected payoff is zero under no arbitrage. This implies that the transfer component does not enter our expected deadweight-loss cost measures, but it can be quantitatively important in realized adverse scenarios.

[30] Because the incremental portfolio includes agency debt and MBS, this calibration abstracts from asset-class differences in the foreign investor base. The goal is not to attribute transfers precisely by asset

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

Table 4: 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.

documents the data construction.

Stress-test results For each program, we evaluate (49)–(51) 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 3.4. Panel A (Panel B) of Table 5 reports outcomes at the 95th (75th) percentile of the portfolio-loss distribution, defined as the distribution of $-r_T^{\mathrm{QE}}$ (equivalently, the 5th and 25th percentiles of r_T^{QE}).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 realized loss magnitudes against the program-level output-effect estimates from Fabo, Jančoková, Kempf, and Pástor (2021) and report the break-even percentile at which $\Delta^{\mathrm{QE}}TC_T$ equals the output-effect benchmark. As discussed in Section 3.4, we do not apply additional discounting in this tail-outcome comparison, which is conservative because fiscal impacts are realized at T while output gains typically accrue earlier.

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 realized fiscal exposure implied by the duration risk embedded in the incremental QE portfolios.

class, but to quantify the order of magnitude of the transfer channel under a transparent benchmark.

	QE1	QE2	MEP	QE3	QE4	All
Panel A: Cost at 95th Percentile						
QE Portfolio Losses	1.18	0.55	2.29	1.88	2.18	8.08
Tax Deadweight Losses	0.32	0.15	0.63	0.53	0.62	2.25
Transfer to Foreign Investors	0.68	0.33	1.41	1.19	0.85	4.46
Total Economic Losses	1.00	0.48	2.04	1.72	1.48	6.71
Panel B: Cost at 75th Percentile						
QE Portfolio Losses	-0.13	0.01	0.73	-0.05	0.25	0.81
Tax Deadweight Losses	-0.03	0.00	0.18	-0.01	0.07	0.21
Transfer to Foreign Investors	-0.07	0.01	0.45	-0.03	0.10	0.45
Total Economic Losses	-0.10	0.01	0.63	-0.05	0.17	0.66
Panel C: Realized Cost						
QE Portfolio Losses	-3.01	-0.83	-0.41	0.56	0.42	-3.26
Tax Deadweight Losses	-0.68	-0.21	-0.12	0.14	0.11	-0.76
Transfer to Foreign Investors	-1.73	-0.50	-0.26	0.36	0.17	-1.96
Total Economic Losses	-2.40	-0.71	-0.37	0.50	0.27	-2.72
Panel D: Benefit (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)	85.98	88.89	76.57	87.56	92.53	86.31
Breakeven Percentile (Academia)	82.19	77.28	65.57	82.33	78.94	77.26

Table 5: 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 *portfolio-loss* distribution across the 1,000,000 term-structure simulations under physical measure (i.e., percentiles of $-Y_{10}r_{10}^{\text{QE}}$). The remaining entries in Panels A and B report the 95th (75th) percentile of tax deadweight losses, transfers to foreigners, and total realized losses across corresponding simulations. Tax deadweight losses, transfers to foreigners, and total realized losses are calculated using (50), (49), and (51), 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, Jančoková, Kempf, and Pástor (2021). Panel E reports the break-even percentile of QE programs, that is, the quantile of the distribution of total economic losses ($\Delta^{\text{QE}}TC_{10}$) which equals the corresponding output-effect benchmark.

4 Conclusion

This paper proposes a framework to characterize and measure the fiscal-efficiency cost of quantitative easing, defined as the ex ante change in expected deadweight losses from taxation induced by the increased interest-rate refinancing risk associated with the policy. When this fiscal-efficiency cost is active, it naturally introduces a trade-off between QE's output-stabilization benefits and its fiscal-efficiency costs. We illustrate this trade-off within a simple three-period model featuring market segmentation, tax distortions, nominal frictions, and a zero lower bound constraint. Utilizing SOMA portfolio holdings and standard term-structure model simulations, we quantify the fiscal-efficiency cost of each QE program implemented in the US. We compare our cost estimates against existing output-stabilization estimates from the literature, concluding that US QE programs were likely all positive net-present-value policies from a cost-efficiency standpoint at inception. These conclusions were reached when isolating the cost-efficiency channel, which is only one channel among others, given the multiple additional feedback effects that a large policy like QE can have on the macroeconomy. We leave it to future research to address those additional channels—such as, among many others, how fiscal authorities respond to changes in central bank duration and how this may erode the monetary dominance regime, with possible effects on sovereign spreads—and to study how these interact with our cost-efficiency measure.

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.

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.

Bernanke, Ben S., 2020, The new tools of monetary policy, *American Economic Review* 110, 943–983.

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).

Board of Governors of the Federal Reserve System, 2011, Federal reserve issues fomc statement, Press release.

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.

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.

Chetty, Raj, 2009, Is the taxable income elasticity sufficient to calculate deadweight loss? the implications of evasion and avoidance, *American Economic Journal: Economic Policy* 1, 31–52.

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, J. M., 1957, The Term Structure of Interest Rates, *The Quarterly Journal of Economics* 71, 485–517.

Dahlahaus, 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.

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.

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, Brian Sack, Michael Fleming, Jeremy Forster, Joshua Frost, Allen Harvey, Spence Hilton, Warren Hrung, Frank Keane, Karin Kimbrough, David Lucca, Brian Madigan, Patricia Mosser, Asani Sarkar, Lisa Stowe, Richard Wagreich, Carl Walsh, and Jonathan Wright, 2011, The Financial Market Effects of the Federal Reserve's Large-Scale Asset Purchases * .

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* (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.

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.

Gürkaynak, Refet S., Brian Sack, and Jonathan H. Wright, 2007, The US Treasury yield curve: 1961 to the present, *Journal of Monetary Economics* 54, 2291–2304.

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, Identification and estimation of Gaussian affine term structure models, *Journal of Econometrics* 168, 315–331.

Hamilton, James D., and Jing Cynthia Wu, 2012b, The Effectiveness of Alternative Monetary Policy Tools in a Zero Lower Bound Environment, *Journal of Money, Credit and Banking* 44, 3–46.

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.

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

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.

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.

Leigh, Daniel, Andrea Pescatori, Pete Devries, and Jaime Guajardo, 2011, *A new action-based dataset of fiscal consolidation* (International Monetary Fund).

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.

Modigliani, F., and R. Sutch, 1966, Innovations in interest rate policy, *The American Economic Review* 178–197.

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.

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, *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, *2011 Meeting Papers, Society for Economic Dynamics* .

Tobin, James, 1969, A General Equilibrium Approach To Monetary Theory, *Journal of Money, Credit and Banking* 1, 15.

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

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

Wallace, Neil, 1981a, A Modigliani-Miller Theorem for Open-Market Operations 71, 267–274.

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

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

Weale, Martin, and Tomasz Wieladek, 2016b, 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.

Appendices

A Measurement Procedure Details

A.1 Gaussian Affine Term Structure Models

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 (2012a).

A.1.1 Basic Framework

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

$$F_{t+1} = c + \rho F_t + \Sigma u_{t+1}, \tag{52}$$

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

$$M_{t,t+1} = \exp[-r_t - \frac{1}{2}\varsigma_t'\varsigma_t - \varsigma_t' u_{t+1}], \tag{53}$$

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 F_t only and are specified as follows:

$$r_t = \delta_0 + \delta_1' F_t \tag{54}$$

$$\varsigma_t = \varsigma + \Phi F_t. \tag{55}$$

A.1.2 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 u_{t+1} and other shocks v_{t+1}, that is, $X_{t+1} = X(\mathcal{F}_t, u_{t+1}, v_{t+1})$. Then its price at t can be calculated according to

$$P_t^X = \mathbb{E}_t[M_{t,t+1}X_{t+1}] = \int M_{t,t+1}X_{t+1}f_t(v_{t+1}|u_{t+1})f_t(u_{t+1})du_{t+1}dv_{t+1}. \tag{56}$$

Next, we define the equivalent $\mathbb{Q}$-measure.[31] Define $f_t^{\mathbb{Q}}(u_{t+1}) = \frac{M_{t,t+1}}{\mathbb{E}_t[M_{t,t+1}]}f_t(u_{t+1})$. We can confirm that $f_t^{\mathbb{Q}}(u_{t+1})$ is a probability density function. Since $M_{t,t+1}$ is $\sigma(u_{t+1}) \vee \mathcal{F}_t$-

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

measurable, we can also conveniently define $f_t^{\mathbb{Q}}(v_{t+1}|u_{t+1}) = f_t(v_{t+1}|u_{t+1})$. Therefore,

$$P_t^X = \mathbb{E}_t[M_{t,t+1}X_{t+1}] = \int M_{t,t+1}X_{t+1}f_t(v_{t+1}|u_{t+1})f_t(u_{t+1})du_{t+1}dv_{t+1} \tag{57}$$

$$= \int \mathbb{E}_t[M_{t,t+1}]X_{t+1}f_t^{\mathbb{Q}}(u_{t+1}, v_{t+1})du_{t+1}dv_{t+1} \overset{\triangle}{=} \mathbb{E}_t^{\mathbb{Q}}[\mathbb{E}_t[M_{t,t+1}]X_{t+1}]. \tag{58}$$

The factor dynamics under the equivalent $\mathbb{Q}$-measure are calculated using the corresponding probability distribution function. Since

$$f_t^{\mathbb{Q}}(u_{t+1}) = \frac{M_{t,t+1}}{\mathbb{E}_t[M_{t,t+1}]}f_t(u_{t+1}) \tag{59}$$

$$= \frac{1}{(2\pi)^{\frac{L}{2}}}\exp(-\frac{1}{2}\varsigma_t'\varsigma_t - \varsigma_t'u_{t+1})\exp\left(-\frac{1}{2}u_{t+1}'u_{t+1}\right) \tag{60}$$

$$= \frac{1}{(2\pi)^{\frac{L}{2}}}\exp\left(-\frac{1}{2}(u_{t+1}+\varsigma_t)'(u_{t+1}+\varsigma_t)\right). \tag{61}$$

Therefore, the factor dynamics under the $\mathbb{Q}$-measure is $F_{t+1} = c + \rho F_t + \Sigma u_{t+1}$, where $u_{t+1} \sim N(-\varsigma_t, I)$. Equivalently,

$$F_{t+1} = c^{\mathbb{Q}} + \rho^{\mathbb{Q}}F_t + \Sigma u_{t+1}^{\mathbb{Q}}, \tag{62}$$

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

$$P_t^X = \mathbb{E}_t[M_{t,t+1}P_{t+1}^X] = \mathbb{E}_t[M_{t,t+T}X_{t+T}], \tag{63}$$

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

A.1.3 The Nominal Yield Curve

We apply the Gaussian affine term structure model above to the pricing of the nominal yield curve.[32] Conjecture that the time-t yield of a nominal risk-free n-period pure-discount bond is

$$y_t^n = a_n + b_n'F_t, \tag{64}$$

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

and the corresponding price under the continuous discounting rule is

$$P_t^n = \exp(-ny_t^n) = \exp\left(-n(a_n + b_n' F_t)\right). \tag{65}$$

On the other hand, the bond is also priced by the one-period nominal pricing kernel under the $\mathbb{Q}$-measure:

$$P_t^n = e^{-r_t}\mathbb{E}_t^{\mathbb{Q}}[P_{t+1}^{n-1}]. \tag{66}$$

Substituting (65), (54) and (62) into (66), we can calculate a_n, b_n recursively according to

$$b_n = \frac{1}{n}\sum_{i=1}^{n}[(\rho^Q)^{i-1}]'\delta_1, \tag{67}$$

$$a_n = \delta_0 + \frac{c^{\mathbb{Q}}}{n}(\sum_{i=1}^{n-1} ib_i) - \frac{1}{2n}\sum_{i=1}^{n-1} i^2 b_i'\Sigma\Sigma' b_i. \tag{68}$$

The estimation strategy follows Hamilton and Wu (2012a). Let Y_t^1 be the L yields in the data that are perfectly spanned by latent factors, and Y_t^2 be other yields that may contain an error. Let A_1, A_2 be vectors and B_1, B_2 be matrices such that $Y_t^1 = A_1 + B_1 F_t$ and $Y_t^2 = A_2 + B_2 F_t$. Similar to Ang and Piazzesi (2003), we adopt the identifying restrictions that $c = 0, \Sigma = I_L$, and ρ is lower triangular. From $Y_t^1 = A_1 + B_1 F_t$ and $F_{t+1} = \rho F_t + u_{t+1}$, we have

$$(B_1)^{-1}(Y_{t+1}^1 - A_1) = \rho(B_1)^{-1}(Y_t^1 - A_1) + u_{t+1}, \tag{69}$$

$$Y_{t+1}^1 = (I - B_1\rho(B_1)^{-1})A_1 + B_1\rho(B_1)^{-1}Y_t^1 + u_{t+1}^1, \tag{70}$$

where $u_{t+1}^1 = B_1 u_{t+1}$. For Y_{t+1}^2, we have

$$Y_{t+1}^2 = A_2 + B_2 F_{t+1}^2 + u_{t+1}^2 \tag{71}$$

$$= A_2 - B_2(B_1)^{-1}A_1 + B_2 B_1^{-1} Y_{t+1}^1 + u_{t+1}^2, \tag{72}$$

where $u_{t+1}^2 \sim N(0, \Sigma_2\Sigma_2')$ for lower-triangular Σ_2 and is independent of u_{t+1}^1. It is equivalent to the following restricted VAR:

$$Y_{t+1} = \hat{A} + \hat{\rho} Y_t + \varepsilon_{t+1}, \tag{73}$$

with

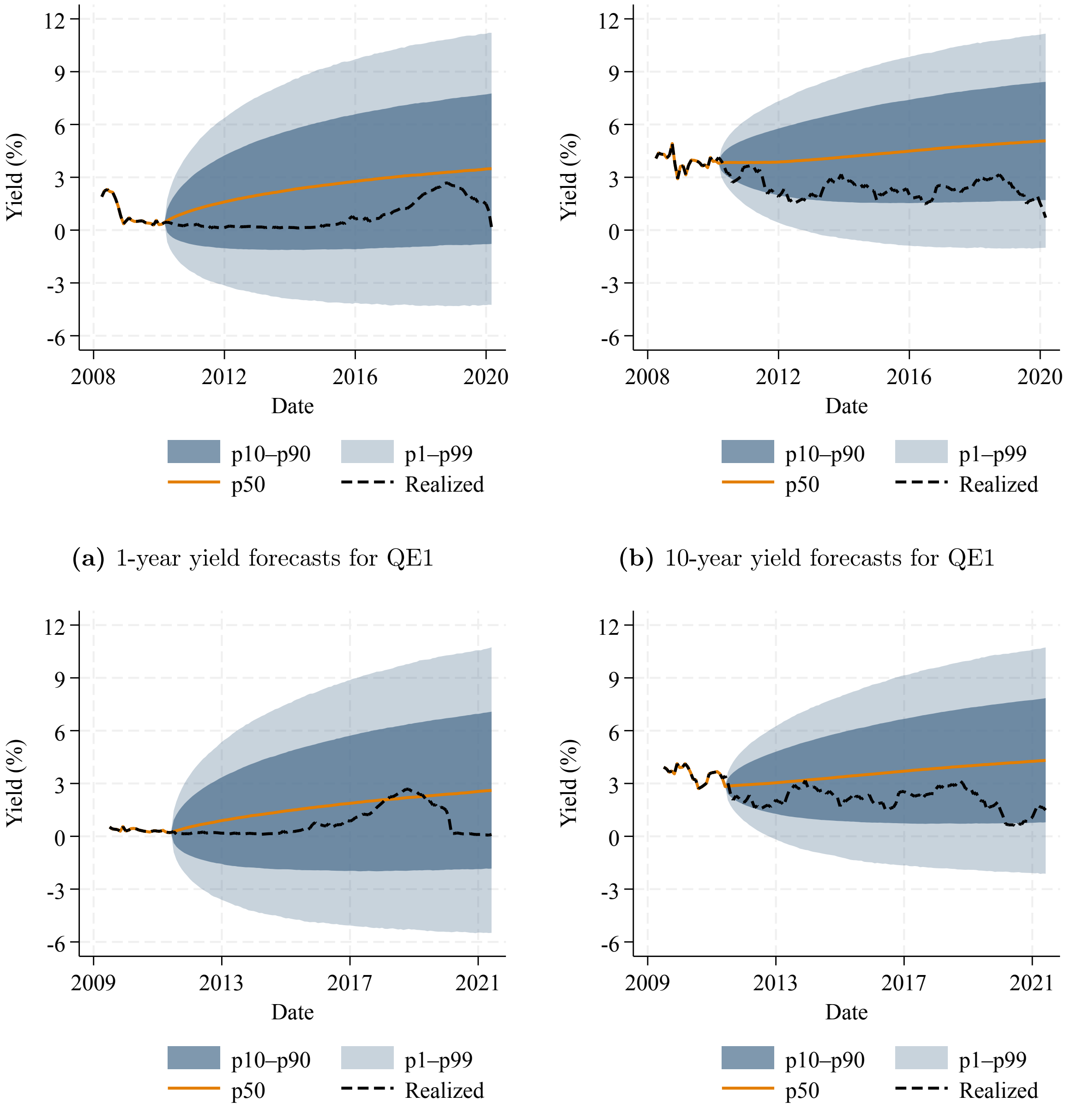
$$\hat{A} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \left(I - B_1\rho B_1^{-1}\right)A_1 \\ A_2 - B_2\rho B_1^{-1}A_1 \end{bmatrix}, \tag{74}$$

$$\hat{\rho} = \begin{bmatrix} \hat{\rho}_{11} & 0 \\ \hat{\rho}_{21} & 0 \end{bmatrix} = \begin{bmatrix} B_1\rho B_1^{-1} & 0 \\ B_2\rho B_1^{-1} & 0 \end{bmatrix}, \tag{75}$$

$$\varepsilon_{t+1} = \begin{bmatrix} \varepsilon_{t+1}^1 \\ \varepsilon_{t+1}^2 \end{bmatrix} = \begin{bmatrix} u_{t+1}^1 \\ B_2 B_1^{-1} u_{t+1}^1 + u_{t+1}^2 \end{bmatrix}. \tag{76}$$

By comparing the number of parameters identified in the restricted VAR and the number of parameters in the term structure model, we conclude that one additional yield in Y_t^2 gives exact identification (Hamilton and Wu, 2012a).

B Relegated Figures



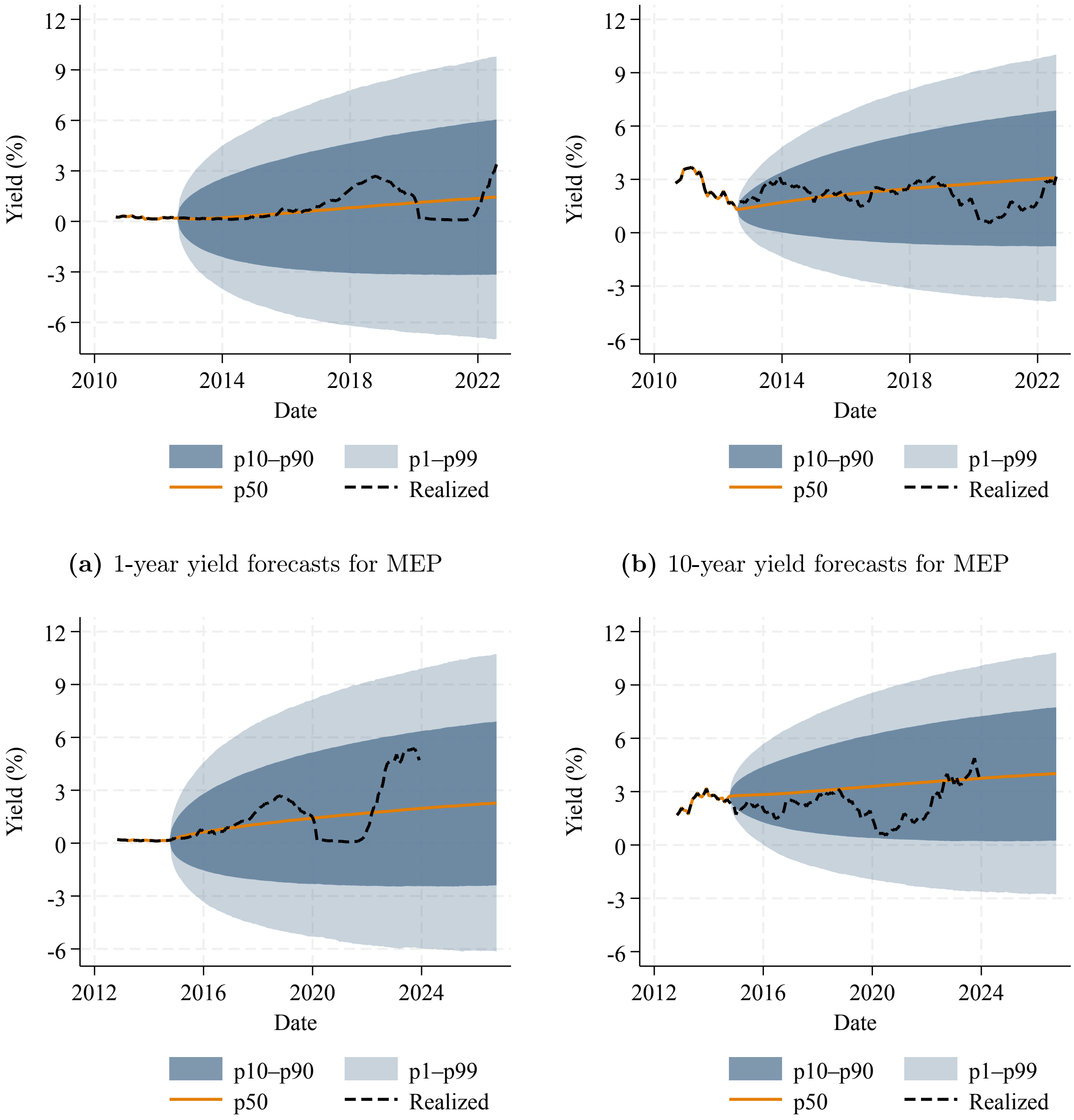
(a) 1-year yield forecasts for QE1

(b) 10-year yield forecasts for QE1

(c) 1-year yield forecasts for QE2

(d) 10-year yield forecasts for QE2

Figure B.1: 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.



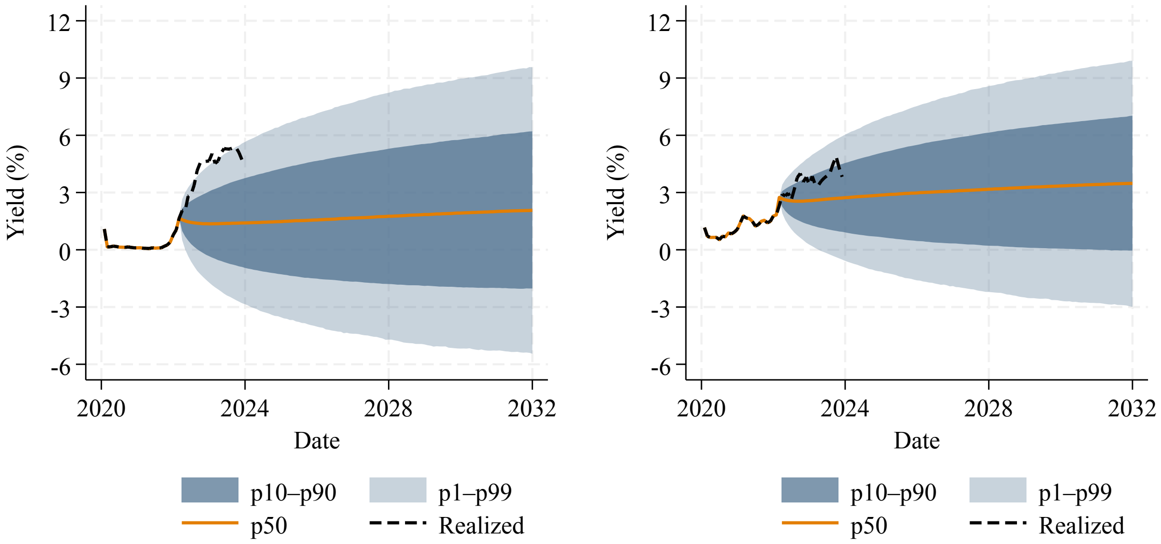
(a) 1-year yield forecasts for MEP

(b) 10-year yield forecasts for MEP

(c) 1-year yield forecasts for QE3

(d) 10-year yield forecasts for QE3

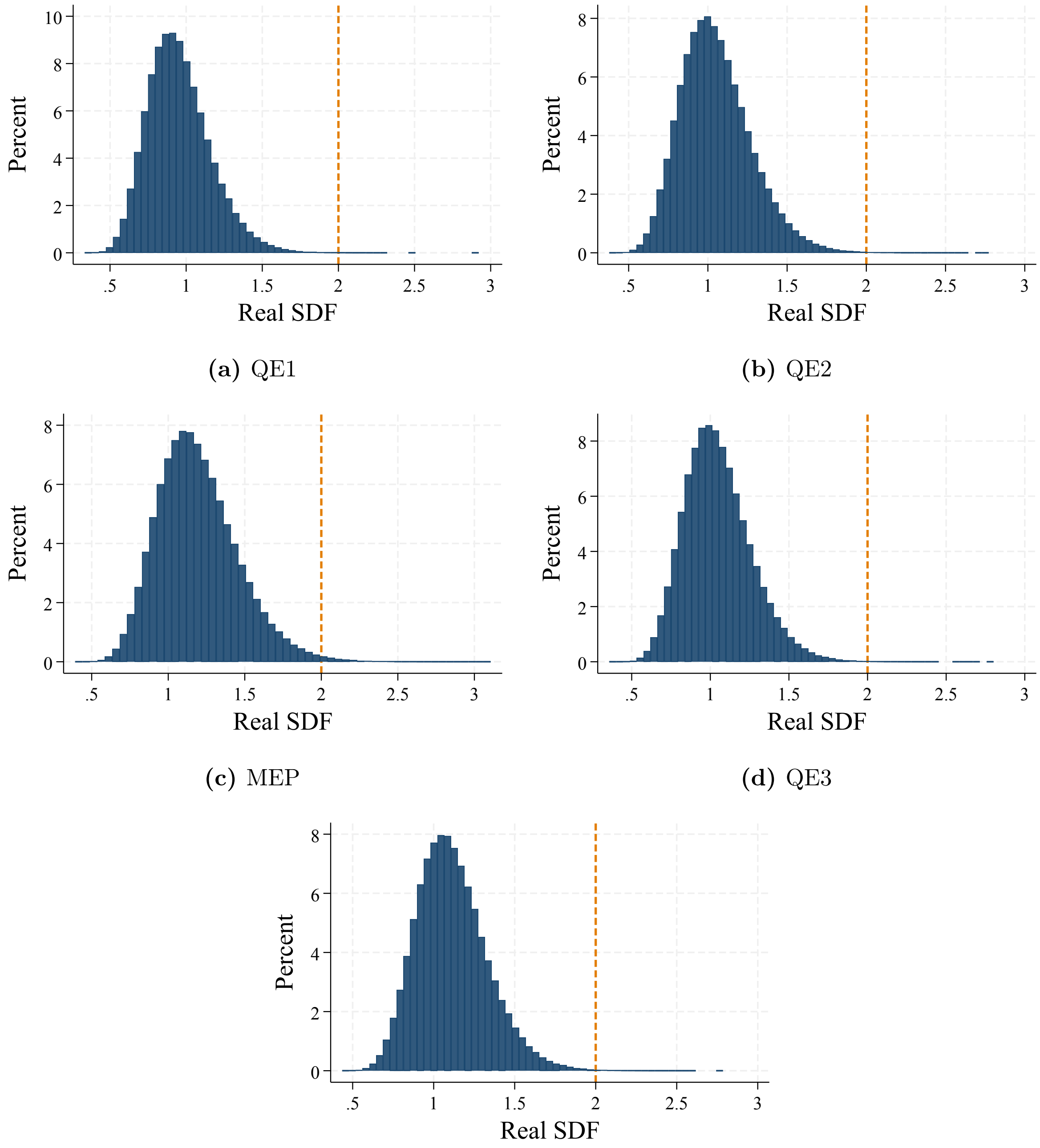
Figure B.2: Term structure forecasts for MEP and QE3. The figure plots the distribution of 10-year-ahead forecasts for the 1-year and 10-year yields (1st, 10th, 50th, 90th, and 99th percentiles across 1,000,000 simulated paths under physical measure) and compares them with realized yields.



(a) 1-year yield forecasts for QE4

(b) 10-year yield forecasts for QE4

Figure B.3: 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 B.4: Envelope condition 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 (45), separately for each QE program. The horizontal line indicates the envelope bound $\lambda_{0,T} \leq \bar{\lambda}$ with $\bar{\lambda} = 2$ at the baseline horizon $T = 10$. This tail-regularity restriction trims only extreme draws: in the program-level Monte Carlo used for our forecasts, approximately 99.5% of simulated paths satisfy the bound.

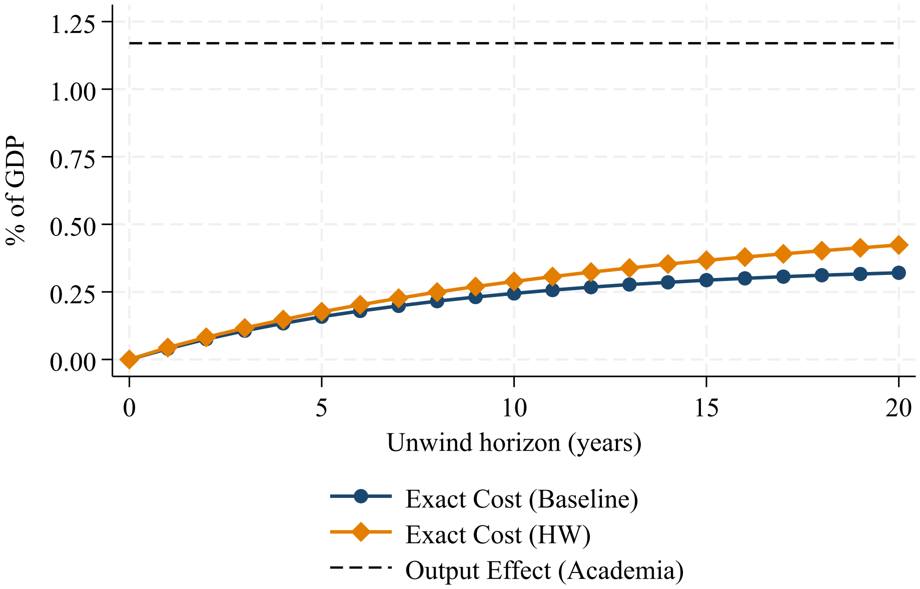


Figure B.5: Aggregate fiscal-efficiency costs: Estimation samples. The figure plots the aggregate exact pricing-kernel cost estimate as a function of the unwind horizon T, expressed as a percent of real GDP at the end of the purchase phase (Y_0). The baseline series uses our preferred term-structure specification and the alternative term structure model parameters from Hamilton and Wu (2012a) specification (HW). The dashed horizontal line reports the output-effect benchmark from Fabo, Jančoková, Kempf, and Pástor (2021). Estimates are computed from 1,000,000 term-structure simulations under $\mathbb{Q}$-measure.

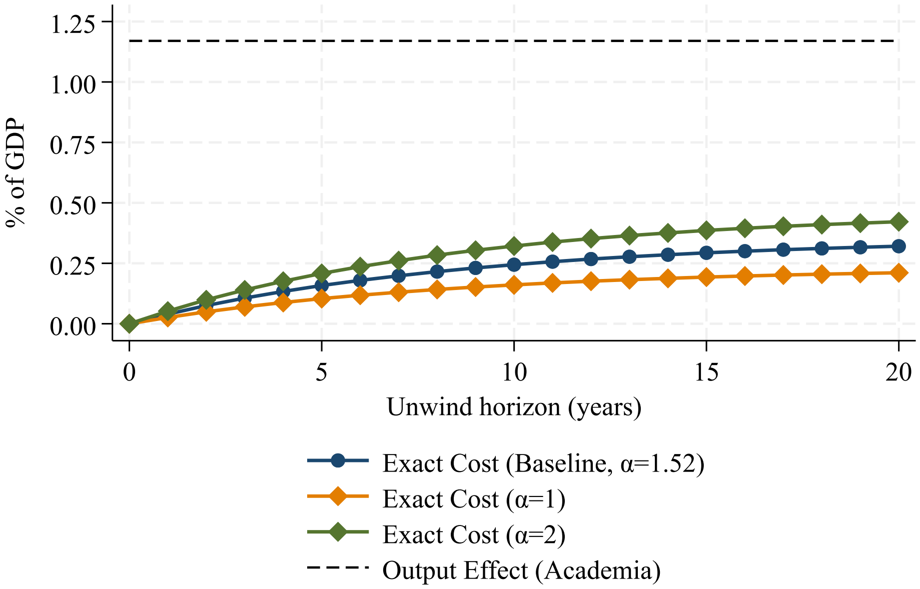


Figure B.6: Aggregate fiscal-efficiency costs: Tax distortion parameters. The figure plots the aggregate exact pricing-kernel cost estimate as a function of the unwind horizon T, expressed as a percent of real GDP at the end of the purchase phase (Y_0). The baseline series uses the calibration $\alpha = 1.52$. The alternative series rescale the baseline estimate proportionally to $\alpha \in \{1, 2\}$. The dashed horizontal line reports the output-effect benchmark from Fabo, Jančoková, Kempf, and Pástor (2021). Estimates are computed from 1,000,000 term-structure simulation under $\mathbb{Q}$-measure.

C Details of Cost Estimation

We estimate the expected cost of QE portfolios for each QE program according to two different formulas. First, we estimate the cost of QE portfolio directly according to

$$\Delta^{\mathrm{QE}} L_0 = \frac{\alpha}{2}\,\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T \left(r_T^{\mathrm{QE}}\right)^2\right] - \alpha\,\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T\, \theta_T^{\mathrm{nQE}}\, r_T^{\mathrm{QE}}\right] \tag{77}$$

Secondly, we estimate the upper bound for the normalized QE portfolio cost according to

$$\overline{\Delta^{\mathrm{QE}} L_0} \equiv \frac{\alpha\bar{\lambda}}{2}\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[r_T^{\mathrm{QE}}\right]\mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]} + \alpha\bar{\lambda}\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[\theta_T^{\mathrm{nQE}}\right]\mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]}. \tag{78}$$

Our target is to estimate several key components of the costs conditional on information up to time 0 for each QE program, that is, the end of the corresponding purchase phase. In Section 3.4, we generate 1,000,000 factor and nominal pricing kernel paths as well as the corresponding yield curves and QE portfolios returns under $\mathbb{Q}$-measure. We make use of the latent variables we identify to also generate 1,000,000 paths of macro variables we are interested in so that we can estimate the conditional covariance between them and QE portfolio returns. Specifically, we assume the following data-generating process for real GDP growth, inflation, and tax rates, $X_t \in \{\log Y_t/Y_{t-1}, \log P_t/P_{-t1}, \log \theta_t\}$:

$$X_t = \alpha + \sum_i \beta^i F_t^i + u_t, \tag{79}$$

where u_t is possibly serially correlated, and F_t^i is the i-th factor we identify. We specify an AR(1) process for the error: $u_t = \rho u_{t-1} + \eta_t$, where $\mathbb{E}_{t-1}[\eta_t] = 0$. For each QE program, we estimate the model parameters using the tax rate, real GDP growth, inflation and latent factors from November 1971 to the end of QE program purchase phase. We then calculate the corresponding 1,000,000 paths for each macro variable based on the 1,000,000 factor paths generated. The costs and upper bounds are calculated using the realizations of all variables in the 1,000,000 paths.

D Output Effect Measurements

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.

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, which we report in Appendix Table D.1. In this sample, the 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 et al. (2012)	1	LSAP2	DSGE	0.07	0.02
Dahlahaus et al. (2018)	1	LSAP1	VAR	0.86	0.11
Dahlahaus et al. (2018)	1	LSAP2	VAR	0.34	0.11
Dahlahaus et al. (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 et al. (2015)	1	LSAP1	DSGE	0.28	0.03
Engen et al. (2015)	1	LSAP2	DSGE	0.11	0.03
Engen et al. (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 et al. (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 et al. (2018)	1	LSAP2	VAR	0.54	0.17
Popescu (2015)	0	LSAP2	VAR	0.04	0.01
Weale and Wieladek (2016b)	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 D.1: 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 in percentage of real GDP at the end of the 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 D.2 presents the average cumulative effect of each QE program on the level of GDP that results from the steps mentioned above.

	QE1	QE2	MEP	QE3	QE4	All	10pp of GDP
Output Effect (All)	0.36	0.28	0.71	0.85	1.18	3.38	0.63
Output Effect (Academia)	0.18	0.05	0.21	0.44	0.34	1.21	0.23
Output Effect (DSGE)	0.18	0.07	0.26	0.52	0.43	1.45	0.27
Output Effect (VAR)	0.44	0.40	0.94	1.02	1.55	4.34	0.81

Table D.2: 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 in percentage of in percentage of real GDP at the end of the 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 to a central bank.

Output Effects from Recent Studies This appendix presents the back-of-the-envelope computations translating the findings from recent studies published after the Fabo, Jančoková, Kempf, and Pástor (2021) collection exercise into output effects comparable in scale and units to those in the earlier literature. 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 percent 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 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,[33] 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 conservative assumption of a 3-year average duration for agency purchases—reflecting the high-interest-rate environment and faster prepayments over their sample period—this corresponds to approximately 0.7% of GDP per 1% of GDP in 10-year duration equivalent purchases.[34]

[33] See series DPCERE1Q156NBEA from FRED.

[34] 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 adjustment.

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.[35] 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 calibrate 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.

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

Online Appendix

OA.1 Proofs

OA.1.1 Proof of Proposition 1

The government's problem in period 0 is given by

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

subject to the budget constraint

$$G_0 = \tau_0 + B_0^S p_0^S + B_0^L p_0^L. \tag{2}$$

From the solution of the period-1 problem, we know

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s) B_0^L}{1 + p_1^S(s)} \quad \text{and} \quad B_1^S(s) = \frac{B_0^S - B_0^L}{1 + p_1^S(s)}. \tag{3}$$

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

$$\mathbb{E}_0 \left[-\tau_0 p_0^S + \frac{\Lambda_1(s)}{\Lambda_0} \tau_1(s) \frac{1}{1 + p_1^S(s)} + \frac{\Lambda_2(s)}{\Lambda_0} \tau_2(s) \frac{1}{1 + p_1^S(s)} \right] = 0, \tag{4}$$

$$\mathbb{E}_0 \left[-\tau_0 p_0^L + \frac{\Lambda_1(s)}{\Lambda_0} \tau_1(s) \frac{p_1^S(s)}{1 + p_1^S(s)} + \frac{\Lambda_2(s)}{\Lambda_0} \tau_2(s) \frac{p_1^S(s)}{1 + p_1^S(s)} \right] = 0. \tag{5}$$

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

$$\tau_0 = \tau_1(s) = \tau_2(s) = \frac{1}{1 + p_0^S + p_0^L} G_0, \tag{6}$$

$$B_0^S = B_0^L = \frac{1}{1 + p_0^S + p_0^L} G_0. \tag{7}$$

OA.1.2 Proof of Lemma 2

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

$$p_0^S = \beta \mathbb{E}_0 \left[\left(\frac{c_1^b(s)}{c_0^b} \right)^{-\gamma} \right]. \tag{8}$$

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

$$c_0^b = y_0 - \frac{1-\phi}{\theta} G_0 + \frac{1-\phi}{\theta} \tau_0, \tag{9}$$

$$c_1^b(s) = y_1(s) + \frac{1-\phi}{\theta} \tau_1(s). \tag{10}$$

Then,

$$p_0^S = \beta \mathbb{E}_0 \left[\left(\frac{\theta y_1(s) + (1-\phi)\tau_1(s)}{\theta y_0 - (1-\phi)G_0 + (1-\phi)\tau_0} \right)^{-\gamma} \right]. \tag{11}$$

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,

$$\tau_1(s) = \frac{B_0^S + p_1^S(s) B_0^L}{1 + p_1^S(s)} = D \left(\frac{-R_1^c(s)}{1 + p_1^S(s)} (S - S^\star) + \frac{1}{p_0^S + p_0^L} \right). \tag{12}$$

Then, we can rewrite equation (11) as

$$y_0 = \frac{1-\phi}{\theta} D + \left(\beta \mathbb{E}_0 \left[\left(a_1(s) + \frac{1-\phi}{\theta} D \left[\frac{-R_1^c(s)}{1 + p_1^S(s)} (S - S^\star) + \frac{1}{p_0^S + p_0^L} \right] \right)^{-\gamma} \right] \right)^{-\frac{1}{\gamma}}. \tag{13}$$

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

OA.1.3 Proof of Lemma 3

If the ZLB is binding,

$$y_0 = \frac{1-\phi}{\theta} D + \left(\beta \mathbb{E}_0 \left[\left(a_1(s) + \frac{1-\phi}{\theta} D \left[\frac{-R_1^c(s)}{1 + p_1^S(s)} (S - S^\star) + \frac{1}{p_0^S + p_0^L} \right] \right)^{-\gamma} \right] \right)^{-\frac{1}{\gamma}}. \tag{14}$$

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

$$\frac{\partial y_0}{\partial D} = \frac{1-\phi}{\theta} + \frac{1-\phi}{\theta}\beta\mathbb{E}_0\left[\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)} \frac{\frac{S}{p_0^S} + \frac{1-S}{p_0^L}p_1^S(s)}{1+p_1^S(s)}\right] \tag{15}$$

and

$$\frac{\partial y_0}{\partial S} = \frac{1-\phi}{\theta}\beta\mathbb{E}_0\left[\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)} \frac{-R_1^c(s)}{1+p_1^S(s)}\right] D, \tag{16}$$

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

OA.1.4 Proof of Proposition 2

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

$$\frac{\alpha}{1-\theta}D^2 m\left(S - \frac{p_0^S}{p_0^S + p_0^L}\right) = \frac{\partial y_0}{\partial S}. \tag{17}$$

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

$$S = \frac{p_0^S}{p_0^S + p_0^L} + \frac{1-\theta}{\alpha m D}\frac{1}{D}\frac{\partial y_0}{\partial S}, \tag{18}$$

and the optimal size of QE is

$$Q = D(S - S^\star) = \frac{1-\theta}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S}. \tag{19}$$

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:

$$\frac{\alpha}{1-\theta}D^2 m(S - S^\star) - (1-\mu)\frac{\partial y_0}{\partial S} = 0, \tag{20}$$

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

$$Q = (1-\mu)\frac{1-\theta}{\alpha m}\frac{1}{D}\frac{\partial y_0}{\partial S}. \tag{21}$$

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:

$$\left(a_0 - \frac{1-\phi}{\theta}D\right)^{-\gamma} - \beta\mathbb{E}_0\left[\left(a_1(s) + \frac{1-\phi}{\theta}\frac{-R_1^c(s)}{1+p_1^S(s)}Q + \frac{1-\phi}{\theta}\frac{D}{p_0^S+p_0^L}\right)^{-\gamma}\right] = 0. \quad (22)$$

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

OA.1.5 Proof of Proposition 3

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

$$L_0 = \mathbb{E}_0\left[\sum_{t=0}^{\infty}\Lambda_{0,t}Y_t\xi\left(\theta_t\right)\right]. \quad (23)$$

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:

$$\Delta^{\text{QE}}L_0 = \mathbb{E}_0\left[\sum_{t=0}^{\infty}\Lambda_{0,t}Y_t\left(\xi\left(\theta_t^{\text{QE}}\right) - \xi\left(\theta_t^{\text{nQE}}\right)\right)\right]. \quad (24)$$

We define $\Delta^{\text{QE}}\theta_t = \theta_t^{\text{QE}} - \theta_t^{\text{nQE}}$ where $\Delta^{\text{QE}}\tau_t$ is the contribution of QE on the tax rate. We expand the difference in the deadweight losses $\xi\left(\theta_t^{\text{QE}}\right) - \xi\left(\theta_t^{\text{nQE}}\right)$ to second order around the counterfactual tax rate θ_t^{nQE}, which is exact for $\xi\left(\theta_t\right) = \frac{\alpha}{2}(\theta_t)^2$:

$$\Delta^{\text{QE}}L_0 = \mathbb{E}_0\left[\sum_{t=0}^{\infty}\Lambda_{0,t}Y_t\left(\xi'\left(\theta_t^{\text{nQE}}\right)\Delta^{\text{QE}}\theta_t + \frac{1}{2}\xi''\left(\theta_t^{\text{nQE}}\right)\left(\Delta^{\text{QE}}\theta_t\right)^2\right)\right]. \quad (25)$$

We take $\xi(\theta_t) = \frac{\alpha}{2}(\theta_t)^2$, and $\xi'(\theta_t^{\text{nQE}}) = \alpha\theta_t^{\text{nQE}}$, $\xi''(\theta_t^{\text{nQE}}) = \alpha$. Then

$$\Delta^{\text{QE}}L_0 = \mathbb{E}_0\left[\sum_{t=0}^{\infty}\Lambda_{0,t}Y_t\left(\alpha\theta_t^{\text{nQE}}\Delta^{\text{QE}}\theta_t + \frac{\alpha}{2}\left(\Delta^{\text{QE}}\theta_t\right)^2\right)\right]. \quad (26)$$

Since we assume $\Delta^{\text{QE}}\tau_T = -R_T^{\text{QE}}$ and $\Delta^{\text{QE}}\tau_t = 0$ for any $t \neq T$, the expression for $\Delta^{\text{QE}}L_0$ can be rewritten as

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

where $r_T^{\text{QE}} = R_T^{\text{QE}}/P_TY_T$. We can write an equivalent expression for equation (27) using the risk-neutral $\mathbb{Q}$-measure, as specified in Appendix A.1.2:

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

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.

OA.1.6 Proof of Proposition 4

To derive the upper bound, first notice that

$$\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_T\left(\theta_T^{\text{nQE}}\,r_T^{\text{QE}}\right)\right] = \text{Cov}_0^{\mathbb{Q}}\left[\theta_t^{\text{nQE}}, \lambda_{0,T}Y_Tr_T^{\text{QE}}\right] \leq \sqrt{\text{Var}_0^{\mathbb{Q}}\left[\theta_t^{\text{nQE}}\right]\text{Var}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_Tr_T^{\text{QE}}\right]} \tag{29}$$

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

$$\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_T\left(r_T^{\text{QE}}\right)^2\right] \leq \sqrt{\text{Var}_0^{\mathbb{Q}}\left[r_T^{\text{QE}}\right]\text{Var}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_Tr_T^{\text{QE}}\right]}. \tag{30}$$

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

$$\text{Var}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_Tr_T^{\text{QE}}\right] = \mathbb{E}_0^{\mathbb{Q}}\left[\left(\lambda_{0,T}Y_Tr_T^{\text{QE}}\right)^2\right] - \left(\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T}Y_Tr_T^{\text{QE}}\right]\right)^2 \leq (\bar{\lambda})^2\mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_Tr_T^{\text{QE}}\right)^2\right]. \tag{31}$$

Therefore,

$$
\begin{aligned}
\Delta^{\mathrm{QE}} L_0 = -\alpha\, \mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T \left(\theta_T^{\mathrm{nQE}}\, r_T^{\mathrm{QE}}\right)\right] + \frac{\alpha}{2}\, \mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T \left(r_T^{\mathrm{QE}}\right)^2\right] \\
= - \alpha \mathrm{Corr}_0\left[\theta_t^{\mathrm{nQE}}, \lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right] \sqrt{\mathrm{Var}_0\left[\theta_t^{\mathrm{nQE}}\right] \mathrm{Var}_0\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right]} \\
+ \frac{\alpha}{2} \mathrm{Corr}_0\left[r_T^{\mathrm{QE}}, \lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right] \sqrt{\mathrm{Var}_0\left[r_T^{\mathrm{QE}}\right] \mathrm{Var}_0\left[\lambda_{0,T} Y_T r_T^{\mathrm{QE}}\right]} \\
\leq \alpha\bar{\lambda} \sqrt{\mathrm{Var}_0\left[\theta_t^{\mathrm{nQE}}\right] \mathbb{E}_0\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]} + \frac{\alpha\bar{\lambda}}{2} \sqrt{\mathrm{Var}_0\left[r_T^{\mathrm{QE}}\right] \mathbb{E}_0\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]}. \qquad (32)
\end{aligned}
$$

OA.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 condition coincides with equation (34) in the main text.

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. $\partial a_1(s)/\partial s > 0$.[36] 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

$$
\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s) B_0^L}{1 + p_1^S(s)} = D\left[-\frac{R_1^c(s)}{1 + p_1^S(s)}(S - S^\star) + \frac{1}{p_0^S + p_0^L}\right], \qquad (33)
$$

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

$$
c_0^b = y_0 - \frac{1-\phi}{\theta} D, \qquad (34)
$$

$$
c_1^b(s) = a_1(s) + \frac{1-\phi}{\theta}\tau_1(s), \qquad (35)
$$

$$
c_2^b(s) = a_2 + \frac{1-\phi}{\theta}\tau_2(s). \qquad (36)
$$

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

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

the period-0 Euler equation as in Lemma 3 yields

$$\frac{\partial y_0}{\partial S} = \frac{1-\phi}{\theta} D\, \mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)} \frac{-R_1^c(s)}{1+p_1^S(s)}\right]. \tag{37}$$

We now give a simple sufficient condition for the expectation in (37) to be strictly positive.

Condition 1 (Small maturity deviation). *Assume*

$$\frac{D(S-S^\star)}{a_2} \le \frac{4\theta}{\gamma(1-\phi)\left(\frac{1}{p_0^S}+\frac{1}{p_0^L}\right)}. \tag{38}$$

Lemma OA.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 $Cov(f(X), g(X)) \ge 0$, with strict inequality whenever X is non-degenerate and both $f(X)$ and $g(X)$ are non-constant.*[37]

Proposition OA.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

$$\frac{\partial \tau}{\partial p}(s) = \frac{B_0^L - B_0^S}{(1+p(s))^2} = -\frac{D(S-S^\star)}{(1+p(s))^2}\left(\frac{1}{p_0^S}+\frac{1}{p_0^L}\right). \tag{39}$$

Define

$$H(s) \equiv \frac{1-\phi}{\theta}\frac{D(S-S^\star)}{(1+p(s))^2}\left(\frac{1}{p_0^S}+\frac{1}{p_0^L}\right) \ge 0. \tag{40}$$

Then (39)–(36) imply

$$\frac{\partial c_1^b}{\partial s}(s) = a_1'(s) - H(s)\frac{\partial p}{\partial s}(s), \qquad \frac{\partial c_2^b}{\partial s}(s) = -H(s)\frac{\partial p}{\partial s}(s). \tag{41}$$

[37] See, for example, Theorem 2.14 in Boucheron et al. (2013).

Next, the date-1 Euler equation implies

$$p(s) = \beta \left(\frac{c_2^b(s)}{c_1^b(s)} \right)^{-\gamma}. \tag{42}$$

Taking logs in (42), differentiating with respect to s, substituting (41), and solving for $p'(s) \equiv \partial p / \partial s(s)$ yields

$$\frac{\partial p}{\partial s}(s) = \frac{\gamma\, p(s)}{c_1^b(s)} \left[1 + \gamma H(s) p(s) \frac{c_2^b(s) - c_1^b(s)}{c_1^b(s) c_2^b(s)} \right]^{-1} a_1'(s). \tag{43}$$

Combining (41) and (43) gives

$$\frac{\partial c_1^b}{\partial s}(s) = \left(1 - \gamma H(s) \frac{p(s)}{c_2^b(s)} \right) \left[1 + \gamma H(s) p(s) \frac{c_2^b(s) - c_1^b(s)}{c_1^b(s) c_2^b(s)} \right]^{-1} a_1'(s). \tag{44}$$

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, $(p-1)^2 \geq 0$), (40) implies

$$\gamma H(s) \frac{p(s)}{c_2^b(s)} < \gamma\, \frac{1 - \phi}{\theta} \frac{D(S - S^\star)}{a_2} \left(\frac{1}{p_0^S} + \frac{1}{p_0^L} \right) \frac{1}{4}. \tag{45}$$

Under condition 1, the right-hand side of (45) is weakly smaller than one, hence

$$1 - \gamma H(s) \frac{p(s)}{c_2^b(s)} > 0 \qquad \forall s. \tag{46}$$

Moreover, the denominator in (43) and (44) 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 (46) and $c_1^b(s) > 0$. Since $a_1'(s) > 0$, (43) and (44) 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 (37). Define

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

Then $X(s)Y(s)$ equals the integrand in (37). 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,

$$\mathbb{E}_0[X(s)Y(s)] = \mathrm{Cov}_0(X(s), Y(s))\,. \tag{47}$$

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

$$\frac{\partial X(s)}{\partial s} = -\gamma\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)} \frac{a_1'(s)}{c_0^b}\left[1 + \gamma H(s)p_1^S(s)\frac{c_2^b(s) - c_1^b(s)}{c_1^b(s)c_2^b(s)}\right]^{-1}\left[\frac{1}{p_0^S} + \gamma H(s)\frac{p_1^S(s)}{c_2^b(s)}R_1^c(s)\right]. \tag{48}$$

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 (46),

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

All remaining factors in (48) 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 OA.1, $\mathrm{Cov}_0(X(s), Y(s)) > 0$. Combining (37) and (47) yields $\partial y_0/\partial S > 0$. ■

OA.3 Model Extensions

OA.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 $\theta\tau_t^b + (1-\theta)\tau_t^h = \tau_t$, and $\theta\tau_t^b = \psi\tau_t$.

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

$$\begin{aligned} G_0 &= \tau_0 + p_0^S B_0^S + p_0^L B_0^L && (49)\\ 0 &= \tau_1(s) - B_0^S + p_1^S(s) B_1^S(s) && (50)\\ 0 &= \tau_2(s) - B_1^S(s) - B_0^L. && (51) \end{aligned}$$

For households, we have

$$\begin{aligned} (1-\theta)c_0^h &= (1-\theta)y_0 - (1-\theta)\tau_0^h - (1-\theta)\xi(\tau_0^h) && (52)\\ (1-\theta)c_1^h(s) &= (1-\theta)y_1(s) - (1-\theta)\tau_1^h(s) - (1-\theta)\xi(\tau_1^h(s)) && (53)\\ (1-\theta)c_2^h(s) &= (1-\theta)y_2 - (1-\theta)\tau_2^h(s) - (1-\theta)\xi(\tau_2^h(s)). && (54) \end{aligned}$$

For domestic bondholders,

$$\begin{aligned} \theta c_0^b &= \theta y_0 - (1-\phi)(p_0^S B_0^S + p_0^L B_0^L) - \theta\tau_0^b - \theta\xi(\tau_0^b) && (55)\\ \theta c_1^b(s) &= \theta y_1(s) - (1-\phi)(-B_0^S + p_1^S(s)B_1^S(s)) - \theta\tau_1^b(s) - \theta\xi(\tau_1^b(s)) && (56)\\ \theta c_2^b(s) &= \theta y_2 - (1-\phi)(-B_1^S(s) - B_0^L) - \theta\tau_2^b(s) - \theta\xi(\tau_2^b(s)), && (57) \end{aligned}$$

and for foreign bondholders

$$\begin{aligned} c_0^f &= -\phi(p_0^S B_0^S + p_0^L B_0^L) && (58)\\ c_1^f(s) &= -\phi(-B_0^S + p_1^S(s)B_1^S(s)) && (59)\\ c_2^f(s) &= -\phi(-B_1^S(s) - B_0^L). && (60) \end{aligned}$$

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

$$\begin{aligned} c_0^b &= y_0 - \frac{1-\phi}{\theta}G_0 + \frac{1-\phi-\psi}{\theta}\tau_0 - \xi\left(\frac{\psi}{\theta}\tau_0\right) && (61)\\ c_1^b(s) &= y_1(s) + \frac{1-\phi-\psi}{\theta}\tau_1(s) - \xi\left(\frac{\psi}{\theta}\tau_1(s)\right) && (62)\\ c_2^b(s) &= y_2 + \frac{1-\phi-\psi}{\theta}\tau_2(s) - \xi\left(\frac{\psi}{\theta}\tau_2(s)\right). && (63) \end{aligned}$$

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

$$\theta\xi(\tau_t^b) + (1-\theta)\xi(\tau_t^h) = \theta\xi\left(\frac{\psi}{\theta}\tau_t\right) + (1-\theta)\xi\left(\frac{1-\psi}{1-\theta}\tau_t\right). \qquad (64)$$

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

$$\theta\xi(\tau_t^b) + (1-\theta)\xi(\tau_t^h) = \frac{\alpha}{2}\left[\frac{\psi^2}{\theta} + \frac{(1-\psi)^2}{1-\theta}\right](\tau_t)^2. \tag{65}$$

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

$$p_0^S = \beta\mathbb{E}_0\left[\left(\frac{c_1^b(s)}{c_0^b}\right)^{-\gamma}\right] = \mathbb{E}_0\left[\frac{\Lambda_1(s)}{\Lambda_0}\right] \tag{66}$$

$$p_0^L = \beta^2\mathbb{E}_0\left[\left(\frac{c_2^b(s)}{c_0^b}\right)^{-\gamma}\right] = \mathbb{E}_0\left[\frac{\Lambda_2(s)}{\Lambda_0}\right] \tag{67}$$

$$p_1^S(s) = \beta\left(\frac{c_2^b(s)}{c_1^b(s)}\right)^{-\gamma} = \frac{\Lambda_2(s)}{\Lambda_1(s)}. \tag{68}$$

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:

$$\max_{\{\tau_1(s),\tau_2(s),B_1^S(s)\}}\left\{\left(a_1(s) + \frac{\Lambda_2(s)}{\Lambda_1(s)}a_2\right) - \frac{\alpha}{2}\left[\frac{\psi^2}{\theta} + \frac{(1-\psi)^2}{1-\theta}\right]\left((\tau_1(s))^2 + \frac{\Lambda_2(s)}{\Lambda_1(s)}(\tau_2(s))^2\right)\right\}, \tag{69}$$

subject to the budget constraints:

$$0 = \tau_1(s) + B_1^S(s)p_1^S(s) - B_0^S \tag{70}$$

$$0 = \tau_2(s) - B_1^S(s) - B_0^L. \tag{71}$$

First-order condition is

$$\tau_1(s)(-p_1^S(s)) + \frac{\Lambda_2(s)}{\Lambda_1(s)}\tau_2(s) = 0. \tag{72}$$

Then the period-1 problem solution is

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s)B_0^L}{1 + p_1^S(s)}. \tag{73}$$

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

$$\tau_1(s) = \tau_2(s) = D \frac{S/p_0^S + p_1^S(s)(1-S)/p_0^L}{1 + p_1^S(s)} \tag{74}$$

$$\tau_0 = G_0 - (p_0^S B_0^S + p_0^L B_0^L) = G_0 - D. \tag{75}$$

The government's period-0 objective is

$$\mathbb{E}_0 \left[y_0 + \frac{\Lambda_1(s)}{\Lambda_0} a_1(s) + \frac{\Lambda_2(s)}{\Lambda_0} a_2 \right] \tag{76}$$

$$-\frac{\alpha}{2}(\frac{\psi^2}{\theta} + \frac{(1-\psi)^2}{1-\theta}) \left\{ \mathbb{E}_0 \left[(\tau_0)^2 + \frac{\Lambda_1(s)}{\Lambda_0}(\tau_1(s))^2 + \frac{\Lambda_2(s)}{\Lambda_0}(\tau_2(s))^2 \right] \right\}. \tag{77}$$

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:

$$y_0 - \frac{\alpha}{2} \left(\frac{\psi^2}{\theta} + \frac{(1-\psi)^2}{1-\theta} \right) \left\{ (G_0 - D)^2 + D^2 \left[m \left(S - \frac{p_0^S}{p_0^S + p_0^L} \right)^2 + \frac{1}{p_0^S + p_0^L} \right] \right\}, \tag{78}$$

where

$$m = \mathbb{E}_0 \left[\left(\frac{R_1^c(s)}{1 + p_1^S(s)} \right)^2 \left(\frac{\Lambda_1(s)}{\Lambda_0} + \frac{\Lambda_2(s)}{\Lambda_0} \right) \right]. \tag{79}$$

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

$$\frac{\partial y_0}{\partial S} - \alpha \left(\frac{\psi^2}{\theta} + \frac{(1-\psi)^2}{1-\theta} \right) D^2 m (S - S^\star) = 0. \tag{80}$$

The new multiplier in marginal cost $\frac{\psi^2}{\theta} + \frac{(1-\psi)^2}{1-\theta}$ 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,

$$p_0^S = \beta \mathbb{E}_0 \left[\left(\frac{a_1(s) + \frac{1-\phi-\psi}{\theta}\tau_1(s) - \frac{\alpha}{2}(\frac{\psi}{\theta}\tau_1(s))^2}{y_0 - \frac{1-\phi-\psi}{\theta}(G_0 - \tau_0) - \frac{\alpha}{2}(\frac{\psi}{\theta}\tau_0)^2} \right)^{-\gamma} \right], \tag{81}$$

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:

$$\frac{\partial y_0}{\partial S} = \frac{\beta}{p_0^S}\mathbb{E}_0\left[\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)}\left[\frac{1-\phi-\psi}{\theta} - \alpha\left(\frac{\psi}{\theta}\tau_1(s)\right)\frac{\psi}{\theta}\right]\frac{\partial \tau_1(s)}{\partial S}\right] \tag{82}$$

$$= \frac{D}{p_0^S}\frac{1-\phi-\psi}{\theta}\mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)}\frac{-R_1^c(s)}{1+p_1^S(s)}\right] \tag{83}$$

$$- \frac{D}{p_0^S}\frac{\psi}{\theta}\mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)}\frac{-R_1^c(s)}{1+p_1^S(s)}\alpha\left(\frac{\psi}{\theta}\tau_1(s)\right)\right]. \tag{84}$$

Here $\frac{1-\phi-\psi}{\theta}$ 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.

OA.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 (20). 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:

$$\min_{\tau_1(s),\tau_2(s)} \tau_1(s)^2 + \frac{\Lambda_2^g(s)}{\Lambda_1^g(s)}\tau_2(s)^2 \quad \text{s.t.} \quad \tau_1(s) + p_1^S(s)\tau_2(s) = B_0^S + p_1^S(s)B_0^L. \tag{85}$$

Define the wedge

$$\chi(s) \equiv \frac{\Lambda_2^g(s)/\Lambda_1^g(s)}{\Lambda_2(s)/\Lambda_1(s)} = \frac{\Lambda_2^g(s)/\Lambda_1^g(s)}{p_1^S(s)}. \tag{86}$$

Solving (85) yields

$$\tau_2(s) = \frac{B_0^S + p_1^S(s) B_0^L}{p_1^S(s) + \chi(s)}, \qquad \tau_1(s) = \chi(s)\, \tau_2(s). \tag{87}$$

Substituting (87) into the continuation tax-loss term gives the exact simplification

$$\frac{\Lambda_1^g(s)}{\Lambda_0^g}\tau_1(s)^2 + \frac{\Lambda_2^g(s)}{\Lambda_0^g}\tau_2(s)^2 = \frac{\Lambda_1^g(s)}{\Lambda_0^g} \cdot \frac{\chi(s)}{p_1^S(s) + \chi(s)} \left(B_0^S + p_1^S(s) B_0^L\right)^2. \tag{88}$$

When $\Lambda_t^g \equiv \Lambda_t$, we have $\chi(s) \equiv 1$ and (87) 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

$$B_0^S + p_1^S(s) B_0^L = D\left(\frac{1 + p_1^S(s)}{p_0^S + p_0^L} - (S - S^\star) R_1^c(s)\right). \tag{89}$$

Plugging (88)–(89) into the period-0 objective (and using $\tau_0 = G_0 - D$) yields the following analogue of Lemma 1. Define

$$\begin{aligned}
A_0 &\equiv \mathbb{E}_0\left[\frac{\Lambda_1^g(s)}{\Lambda_0^g}\, \frac{\chi(s)}{p_1^S(s) + \chi(s)} \left(\frac{1 + p_1^S(s)}{p_0^S + p_0^L}\right)^2\right], \\
A_1 &\equiv \mathbb{E}_0\left[\frac{\Lambda_1^g(s)}{\Lambda_0^g}\, \frac{\chi(s)}{p_1^S(s) + \chi(s)} \left(\frac{1 + p_1^S(s)}{p_0^S + p_0^L}\right) R_1^c(s)\right], \\
A_2 &\equiv \mathbb{E}_0\left[\frac{\Lambda_1^g(s)}{\Lambda_0^g}\, \frac{\chi(s)}{p_1^S(s) + \chi(s)} \left(R_1^c(s)\right)^2\right].
\end{aligned} \tag{90}$$

Then the period-0 problem can be written as

$$\max_{S,D}\left\{y_0 - \frac{\alpha}{2(1-\theta)}\left((G_0 - D)^2 + D^2\left[A_2\,(S - S^g)^2 + \left(A_0 - \frac{A_1^2}{A_2}\right)\right]\right)\right\}, \tag{91}$$

where

$$S^g \equiv S^\star + \frac{A_1}{A_2}. \tag{92}$$

When $\Lambda_t^g \equiv \Lambda_t$, $\chi(s) \equiv 1$ and (90) satisfies $A_1 = 0$, $A_2 = m$, and $A_0 = 1/(p_0^S + p_0^L)$, so (91) 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 (91) 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 (87) and (89), taxes in period 1 satisfy

$$\tau_1(s) = \frac{\chi(s)}{p_1^S(s) + \chi(s)} D \left(R_1^c(s)(S^\star - S) + \frac{1 + p_1^S(s)}{p_0^S + p_0^L} \right). \tag{93}$$

Differentiating the ZLB output expression as in Lemma 3 yields

$$\frac{\partial y_0}{\partial S} = \frac{1-\phi}{\theta} D \, \mathbb{E}_0 \left[\beta \left(\frac{c_1^b(s)}{c_0^b} \right)^{-(1+\gamma)} \frac{\chi(s)}{p_1^S(s) + \chi(s)} \left(-R_1^c(s) \right) \right], \tag{94}$$

which coincides with (32) when $\chi(s) \equiv 1$. Taking the first-order condition with respect to S in (91) gives

$$\frac{\alpha}{1-\theta} D^2 A_2 \left(S - S^g \right) = \frac{\partial y_0}{\partial S}. \tag{95}$$

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 $\Lambda_t^g \equiv \Lambda_t$.

OA.3.3 Price-strategic Government

In the baseline model, the government chooses (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 $\omega \in [0, 1]$ of the induced price changes. We recover Proposition 1 and Proposition 2 as the special case $\omega = 0$.

Equilibrium price schedules. Assume that, in equilibrium, the date-0 bond prices and the date-1 short price satisfy

$$p_0^S = p_0^S(B_0^S, B_0^L), \qquad p_0^L = p_0^L(B_0^S, B_0^L), \qquad p_1^S(s) = p_1^S\big(s; B_0^S, B_0^L\big), \tag{96}$$

with all functions differentiable. The parameter ω captures how much of the derivatives of these schedules the government internalizes when forming first-order conditions. Thus,

$\omega = 0$ corresponds to price taking, while $\omega = 1$ 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, the period-1 government problem is unchanged and taxes are smoothed across periods 1 and 2:

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s) B_0^L}{1 + p_1^S(s)}. \tag{97}$$

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

$$\tau_0 = G_0 - p_0^S B_0^S - p_0^L B_0^L. \tag{98}$$

Under (96), issuance affects τ_0 through equilibrium proceeds. For $j \in \{S, L\}$, define the ω-internalized marginal-proceeds price

$$\tilde{p}_{0,j} \equiv p_{0,j} + \omega \Big(B_0^S \frac{\partial p_0^S}{\partial B_{0,j}} + B_0^L \frac{\partial p_0^L}{\partial B_{0,j}} \Big), \tag{99}$$

so that the government perceives

$$\left(\frac{\partial \tau_0}{\partial B_{0,j}} \right)_\omega = -\tilde{p}_{0,j}. \tag{100}$$

In addition, if $p_1^S(s)$ depends on issuance, differentiating (97) yields the exact identity

$$\frac{\partial \tau_1(s)}{\partial B_{0,j}} = w_j(p_1^S(s)) + \frac{B_0^L - B_0^S}{(1 + p_1^S(s))^2} \frac{\partial p_1^S(s)}{\partial B_{0,j}}, \tag{101}$$

and the government perceives the second term in (101) scaled by ω.

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

$$0 = \mathbb{E}_0 \left[-\tau_0\, \tilde{p}_{0,j} + W(s)\, \tau_1(s)\, w_j(p_1^S(s)) + \omega\, W(s)\, \tau_1(s)\, \frac{B_0^L - B_0^S}{(1 + p_1^S(s))^2} \frac{\partial p_1^S(s)}{\partial B_{0,j}} \right], \qquad j \in \{S, L\}. \tag{102}$$

When $\omega = 0$, (102) reduces to the price-taking conditions used in the proof of Proposition 1. For $\omega > 0$, the optimal maturity reflects two additional incentives: (i) issuance is evaluated at marginal proceeds (99), and (ii) the government internalizes how issuance shifts the continuation short price schedule $p_1^S(s)$ in (101).

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

$$\frac{\partial y_0}{\partial S} = \frac{\alpha}{2(1-\theta)} D^2 \left[2m(S - S^\star) + \omega \left\{ (S - S^\star)^2 \frac{\partial m}{\partial S} - 2m(S - S^\star) \frac{\partial S^\star}{\partial S} + \frac{\partial}{\partial S} \left(\frac{1}{p_0^S + p_0^L} \right) \right\} \right], \tag{103}$$

where the derivatives are taken with respect to S holding D fixed. When $\omega = 0$, the bracketed ω-term drops out and (103) collapses to the main-text optimality condition (31), yielding Proposition 2. For $\omega > 0$, the optimal deviation from tax smoothing (and hence the optimal QE size) is generally characterized implicitly by (103).

OA.3.4 Zero-interest Liabilities

In the main draft, 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.[38] We assume that M is outstanding in periods 0 and 1 and is redeemed in period 2. The government budget constraints become

$$\begin{aligned} G_0 &= \tau_0 + B_0^S p_0^S + B_0^L p_0^L + M, && (104)\\ 0 &= \tau_1(s) + B_1^S(s) p_1^S(s) - B_0^S, && (105)\\ 0 &= \tau_2(s) - B_1^S(s) - B_0^L - M. && (106) \end{aligned}$$

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.

[38] 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.

Proposition OA.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*

$$\widetilde{Q}^\star = \min\left\{\frac{1-\theta}{\alpha m}\frac{1}{\widetilde{D}}\frac{\partial y_0}{\partial\widetilde{S}}, \overline{Q}\right\}, \tag{107}$$

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

$$\mathbb{E}_0\left\{y_0 - \frac{\alpha}{2(1-\theta)}\left[(G_0 - p_0^S B_0^S - p_0^L B_0^L - M)^2 + \frac{\Lambda_1(s)}{\Lambda_0}\frac{(B_0^S + p_1^S(s)(B_0^L + M))^2}{1+p_1^S(s)}\right]\right\}, \tag{108}$$

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

$$y_0 - \frac{\alpha}{2(1-\theta)}\left[(\tilde{G}_0 - \tilde{D})^2 + \tilde{D}^2\left(m(\tilde{S} - \frac{p_0^S}{p_0^S + p_0^L})^2 + \frac{1}{p_0^S + p_0^L}\right)\right], \tag{109}$$

where

$$m = \mathbb{E}_0\left[\frac{\Lambda_1(s)}{\Lambda_0(1+p_1^S(s))}\left(\frac{1}{p_0^S} - \frac{p_1^S(s)}{p_0^L}\right)^2\right] = \mathbb{E}_0\left[\frac{\Lambda_1(s)}{\Lambda_0(1+p_1^S(s))}\left(-R_1^c(s)\right)^2\right]. \tag{110}$$

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

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

$$\frac{\partial y_0}{\partial M} = \frac{1-\phi}{\theta}\,\mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(1+\gamma)}\frac{-R_1^m(s)}{1+p_1^S(s)}\right] \leq 0, \tag{111}$$

where $R_1^m(s) \equiv \frac{p_1^S(s)}{p_0^L} - p_1^S(s) \geq 0$ is the time-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

$$y_0 = \frac{1-\phi}{\theta}D + \left(\beta\mathbb{E}_0\left[\left(a_1(s) + \frac{1-\phi}{\theta}\frac{B_0^S + p_1^S(s)\left[\frac{D-p_0^S B_0^S}{p_0^L} + M\left(1-\frac{1}{p_0^L}\right)\right]}{1+p_1^S(s)}\right)^{-\gamma}\right]\right)^{-\frac{1}{\gamma}} \tag{112}$$

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

$$\frac{\partial y_0}{\partial M} = \frac{1-\phi}{\theta}\mathbb{E}_0\left[\beta\left(\frac{c_1^b(s)}{c_0^b}\right)^{-(\gamma+1)}\frac{p_1^S(s)}{1+p_1^S(s)}\left(1-\frac{1}{p_0^L}\right)\right] < 0. \tag{113}$$

□

Two features deliver the sign restriction in Appendix OA.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.[39]

OA.3.5 Mortgage-Backed Securities

This section extends the three-period model in Section 2 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.[40] 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.

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 corresponding 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

[39] 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.

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

$B_0^{M,g}$ in period 2. The government's budget constraints become

$$G_0 + p_0^M B_0^{M,g} = \tau_0 + p_0^S B_0^S + p_0^L B_0^L, \tag{114}$$

$$0 = \tau_1(s) + p_1^S(s) B_1^S(s) - B_0^S, \tag{115}$$

$$0 = \tau_2(s) + B_0^{M,g} - B_1^S(s) - B_0^L. \tag{116}$$

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

$$c_0^h = y_0 - \frac{\tau_0}{1-\theta} - \frac{\alpha}{2}\left(\frac{\tau_0}{1-\theta}\right)^2, \tag{117}$$

$$c_1^h(s) = y_1(s) - \frac{\tau_1(s)}{1-\theta} - \frac{\alpha}{2}\left(\frac{\tau_1(s)}{1-\theta}\right)^2, \tag{118}$$

$$c_2^h(s) = y_2 - \frac{\tau_2(s)}{1-\theta} - \frac{\alpha}{2}\left(\frac{\tau_2(s)}{1-\theta}\right)^2 - \frac{B^M}{1-\theta}. \tag{119}$$

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 2. In particular, the date-0 MBS price satisfies

$$p_0^M = \mathbb{E}_0\left[\frac{\Lambda_2(s)}{\Lambda_0}\right], \tag{120}$$

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

$$p_0^M = p_0^L. \tag{121}$$

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 (121) into (114)–(116) 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

$$\widetilde{B}_0^L \equiv B_0^L - B_0^{M,g}. \tag{122}$$

Using (122), the government's budget constraints can be rewritten as

$$G_0 = \tau_0 + p_0^S B_0^S + p_0^L \widetilde{B}_0^L, \tag{123}$$

$$0 = \tau_1(s) + p_1^S(s) B_1^S(s) - B_0^S, \tag{124}$$

$$0 = \tau_2(s) - B_1^S(s) - \widetilde{B}_0^L. \tag{125}$$

These constraints are identical to the baseline constraints in Section 2, with B_0^L replaced by $\widetilde{B}_0^L$.

Proposition OA.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:

$$\tau_1(s) = \tau_2(s) = \frac{B_0^S + p_1^S(s)\widetilde{B}_0^L}{1 + p_1^S(s)}, \tag{126}$$

and the associated short-bond issuance rule is

$$B_1^S(s) = \frac{B_0^S - \widetilde{B}_0^L}{1 + p_1^S(s)}. \tag{127}$$

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

$$\widetilde{D} \equiv p_0^S B_0^S + p_0^L \widetilde{B}_0^L, \qquad \widetilde{S} \equiv \frac{p_0^S B_0^S}{\widetilde{D}}, \tag{128}$$

and keep $S^\star \equiv p_0^S/(p_0^S + p_0^L)$ as in Section 2.

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

$$\Delta B_0 \equiv B_0^S - \widetilde{B}_0^L = (B_0^S - B_0^L) + B_0^{M,g}. \tag{129}$$

Equation (129) 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

$$\widetilde{Q} \equiv (\widetilde{S} - S^\star)\widetilde{D}. \tag{130}$$

Using (128), one obtains

$$\widetilde{Q} = \frac{p_0^S p_0^L}{p_0^S + p_0^L} \Delta B_0. \tag{131}$$

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 (128), 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 OA.5 (Optimal QE with MBS). *With $\widetilde{D}$ and $\widetilde{S}$ defined in (128), 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*

$$\widetilde{Q}^\star = \min\left\{\frac{1-\theta}{\alpha m}\frac{1}{\widetilde{D}}\frac{\partial y_0}{\partial \widetilde{S}},\ Q\right\}, \tag{132}$$

where 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.

OA.4 QE episodes in the U.S.

Figure OA.4.1 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.[41] 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. 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.[42] 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 2020–March 2022) involved cumulative purchases of about \$4.6 trillion in long-term Treasury securities and agency MBS.

OA.5 Accounting Decomposition of the Fiscal-cost

Our quantitative exercise in Equation (39)–Proposition 4 computes the expected efficiency cost of financing QE-induced fiscal gains or losses through distortionary taxation.

[41] 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.

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

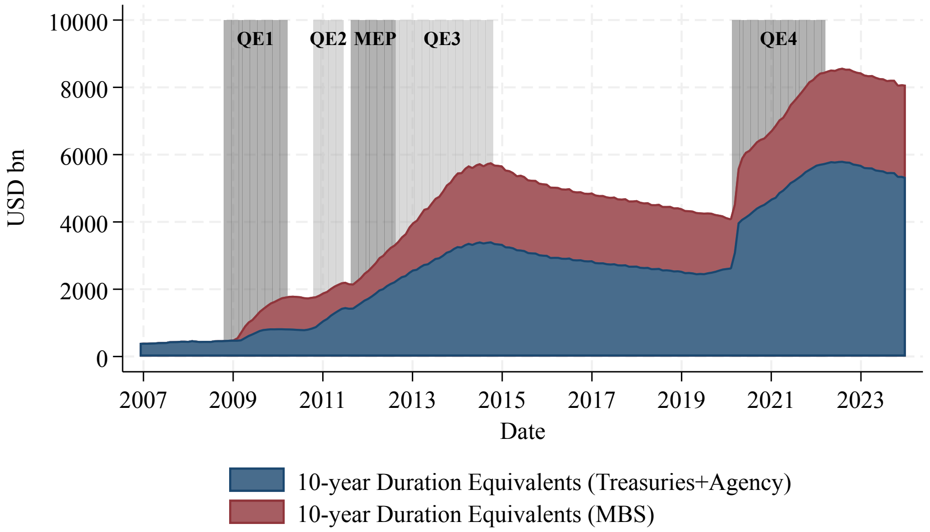


Figure OA.4.1: Size of the Federal Reserve's bond portfolio. This figure plots the evolution of the size of the Federal Reserve's bond portfolio measured in 10-year duration equivalents. The blue area depicts the 10-year duration equivalent values of Treasury and federal agency debt holdings, and the red area the 10-year duration equivalent values of MBS holdings. Shaded areas denote the periods of each specific asset purchase program. Section 3.4 details how we conservatively map MBS holdings into duration equivalents given limited information on prepayment and principal-paydown timing in SOMA releases.

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 (39) 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 (42), 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 (45), 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

$$\Delta^{\mathrm{QE}} x_t \equiv x_t^{\mathrm{QE}} - x_t^{\mathrm{nQE}}. \tag{133}$$

From nominal tax revenues to tax rates Deadweight losses in Equation (39) depend on the tax take relative to the contemporaneous tax base, $\tau_t/(P_tY_t)$. 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:

$$\Delta^{\mathrm{QE}}\left(\frac{\tau_T}{P_TY_T}\right) = \frac{1}{P_T^{\mathrm{QE}}Y_T^{\mathrm{QE}}}\left(\Delta^{\mathrm{QE}}\tau_T - \frac{\tau_T^{\mathrm{nQE}}}{P_T^{\mathrm{nQE}}Y_T^{\mathrm{nQE}}}\,\Delta^{\mathrm{QE}}(P_TY_T)\right). \tag{134}$$

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 (134).

A consolidated flow identity at the unwind date Fix an unwind date T as in Equation (42). 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 $n \in \{1, \ldots, N\}$ 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

$$I_T^{\mathrm{TSY},s}(n) \equiv B_T^{\mathrm{TSY},s}(n) - B_{T-1}^{\mathrm{TSY},s}(n+1), \qquad n = 1, \ldots, N, \tag{135}$$

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

$$\tau_T^s = P_T^sG_T^s + B_{T-1}^{\mathrm{TSY},s}(1) - \sum_{n=1}^{N} p_T^s(n)\, I_T^{\mathrm{TSY},s}(n) - R_T^s. \tag{136}$$

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).[43]

[43] 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.

Lemma OA.2 (Terminal tax decomposition). *Under (136), the QE–no-QE change in nominal tax revenues at date T admits the exact decomposition*

$$\begin{aligned}\Delta^{\mathrm{QE}}\tau_T = \Delta^{\mathrm{QE}}(P_T G_T) + \Delta^{\mathrm{QE}} B_{T-1}^{\mathrm{TSY}}(1) - \sum_{n=1}^{N} \Delta^{\mathrm{QE}} p_T(n)\, I_T^{\mathrm{TSY,nQE}}(n) \\ - \sum_{n=1}^{N} p_T^{\mathrm{QE}}(n)\, \Delta^{\mathrm{QE}} I_T^{\mathrm{TSY}}(n) - \Delta^{\mathrm{QE}} R_T, \end{aligned} \tag{137}$$

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 (137) simplifies to

$$\begin{aligned}\Delta^{\mathrm{QE}}\tau_T = \Delta^{\mathrm{QE}}(P_T G_T) + \Delta^{\mathrm{QE}} B_{T-1}^{\mathrm{TSY}}(1) - \sum_{n=1}^{N} \Delta^{\mathrm{QE}} p_T(n)\, I_T^{\mathrm{TSY,nQE}}(n) \\ + \sum_{n=1}^{N-1} p_T^{\mathrm{QE}}(n)\, \Delta^{\mathrm{QE}} B_{T-1}^{\mathrm{TSY}}(n+1) - \Delta^{\mathrm{QE}} R_T. \end{aligned} \tag{138}$$

Proof. Subtract (136) under $s = \mathrm{nQE}$ from (136) 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 (137). If $\Delta^{\mathrm{QE}} B_T^{\mathrm{TSY}}(n) = 0$ for all n, then (135) implies $\Delta^{\mathrm{QE}} I_T^{\mathrm{TSY}}(n) = -\Delta^{\mathrm{QE}} B_{T-1}^{\mathrm{TSY}}(n+1)$ for $n \leq N-1$ and $\Delta^{\mathrm{QE}} I_T^{\mathrm{TSY}}(N) = 0$. Substituting into (137) yields (138). □

Interpreting the terms The decomposition in Lemma OA.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 (136) 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 $-\sum_n \Delta^{\mathrm{QE}} p_T(n)\, I_T^{\mathrm{TSY,nQE}}(n)$ 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 $-\sum_n p_T^{\mathrm{QE}}(n)\,\Delta^{\mathrm{QE}} I_T^{\mathrm{TSY}}(n)$ captures active changes in issuance quantities and maturity composition. Under terminal equalization, this term reduces to the portfolio-offset term in (138), which depends only on the legacy maturity vector inherited at $T-1$.

(iv) Central-bank transfer channel. The term $-\Delta^{\mathrm{QE}} R_T$ 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 (42).

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 $\Delta^{\mathrm{QE}}(P_T G_T) = 0$ and $\Delta^{\mathrm{QE}} I_T^{\mathrm{TSY}}(n) = 0$), and we evaluate both scenarios along the same realized path of bond prices (so $\Delta^{\mathrm{QE}} p_T(n) = 0$). Under these restrictions, (137) reduces to $\Delta^{\mathrm{QE}}\tau_T = -\Delta^{\mathrm{QE}} R_T$. With the consolidated-government interpretation in Equation (43), $\Delta^{\mathrm{QE}} R_T = R_T^{\mathrm{QE}}$ and therefore $\Delta^{\mathrm{QE}}\tau_T = -R_T^{\mathrm{QE}}$. The remaining terms in Lemma OA.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 (45).

OA.6 Results with Hamilton-Wu Parameters

QE Portfolio Returns Due to limited access to yield data at maturity up to 15 years before 1971, the sample we use train our term structure model starts from 1971 in our main analysis. As a comparison, Hamilton and Wu (2012a) (HW) trains the term structure model using yield data from 1952 to 2000, but only at maturity up to 5 years. In this section, we reconduct our analysis using the term structure parameters from HW as a robustness check. For each QE program, we first reconstruct latent factors using yield curve data up to the end of the purchase phase using Hamilton-Wu term structure parameters. We then generate 1,000,000 random samples of future factor paths, and compute corresponding yield curve paths as well as macro variable paths up to 20 years.

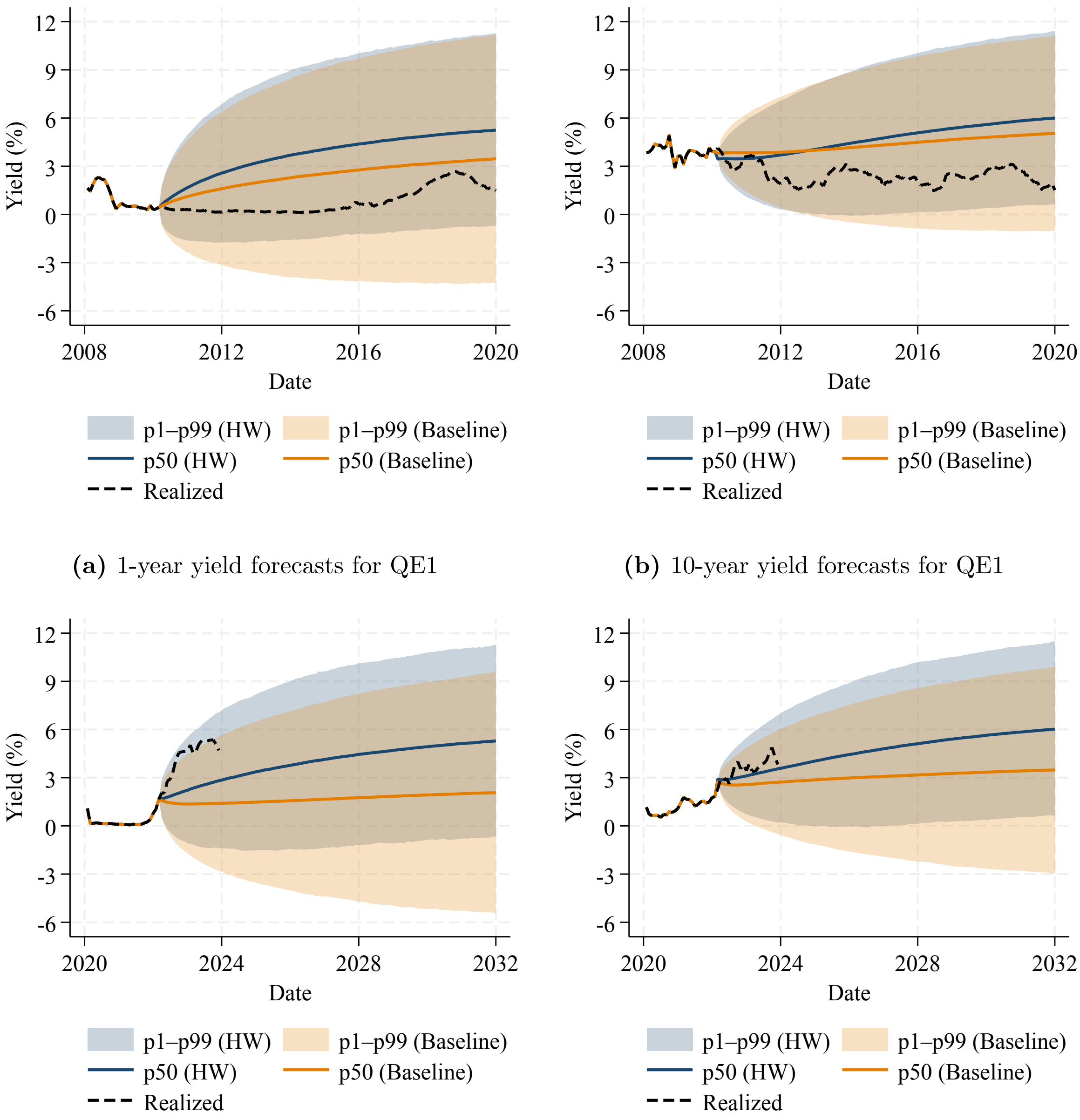
Figure OA.6.1 compares the term structure prediction in the Hamilton-Wu parameter robustness check with the prediction in our main analysis. As can be observed, the portfolio return volatility is larger when Hamilton-Wu parameter is used due to the higher in-sample mean and volatility for interest rates in the Hamilton-Wu sample period.

However, our cost estimates are still relatively robust.

Cost-Benefit Analysis The costs of QE are estimated analogously as in our main analysis. We calculate the expected cost of QE portfolios using equation (45) as well as its upper bound using equation (46) for each QE program. In our baseline parameterization, we use Saez, Slemrod, and Giertz's (2012) estimate of the marginal deadweight losses, $\xi'(\theta_T) = \alpha\theta_T = 0.38$, which yields $\alpha = 1.52$ with a tax rate of 25%. The results are reported in Panel A of Table OA.6.1, in comparison with the output effect estimates reported in Panel B of Table OA.6.1. Compared to our baseline analysis with the term structure data starting from 1971, the estimates of expected costs only increase by a small fraction.

	QE1	QE2	MEP	QE3	QE4	All
Panel A: Estimated Cost (% of GDP)						
Pricing-Kernel Estimate	0.03	0.01	0.07	0.08	0.09	0.28
Upper-Bound Estimate	0.15	0.05	0.24	0.28	0.31	1.03
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 OA.6.1: The Trade-off of QE Programs Estimated with Hamilton-Wu Parameters. This table reports our calculation of the trade-off of QE programs the Fed conducted, in percentage of GDP at the end of the purchase phase, Y_0. The first 5 columns present the trade-off for each QE program separately, and the last column summarizes the trade-off by aggregating across all QE programs. QE portfolios are assumed to be held for $T = 10$ years, and the distribution of losses is calculated using 1,000,000 yield curve forecasts under $\mathbb{Q}$-measure generated with the term structure model trained by Hamilton and Wu (2012a). The expected cost of QE portfolios is calculated using equation (45). The upper-bound cost of the QE portfolio is calculated using Equation (46). Output effects are calculated by averaging the estimates of articles surveyed in Fabo, Jančoková, Kempf, and Pástor (2021). We discount output effects using the conservative approach detailed in Section 3.4.



(a) 1-year yield forecasts for QE1

(b) 10-year yield forecasts for QE1

(c) 1-year yield forecasts for QE4

(d) 10-year yield forecasts for QE4

Figure OA.6.1: Term Structure Forecast with Hamilton-Wu Parameters. This figure compares 10-year forecasts of 1-year and 10-year yield for QE1 and QE4 from term structure models trained by the authors and Hamilton and Wu (2012a). For each QE program, we train our preferred 3-factor term structure model using samples of the yield curve data that include yields with maturity of up to 15 years, starting from November 1971 and ending on the month when QE purchases were finished. Hamilton-Wu trains the term structure model using yield data from 1952 to 2000, but only at maturity up to 5 years. We generate 1,000,000 random paths under physical measure for the yield curve for each term structure model and compute the 1st, 50th, and 99th percentile for 1-year and 10-year yields.

OA.6.1 Details of Cost Estimation

We estimate the expected cost of QE portfolios for each QE program according to two different formulas. First, we estimate the cost of QE portfolio directly according to

$$\Delta^{\mathrm{QE}} L_0 = \frac{\alpha}{2}\,\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T \left(r_T^{\mathrm{QE}}\right)^2\right] - \alpha\,\mathbb{E}_0^{\mathbb{Q}}\left[\lambda_{0,T} Y_T\,\theta_T^{\mathrm{nQE}}\,r_T^{\mathrm{QE}}\right] \tag{139}$$

Secondly, we estimate the upper bound for the normalized QE portfolio cost according to

$$\overline{\Delta^{\mathrm{QE}} L_0} \equiv \frac{\alpha\bar{\lambda}}{2}\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[r_T^{\mathrm{QE}}\right]\mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]} + \alpha\bar{\lambda}\sqrt{\mathrm{Var}_0^{\mathbb{Q}}\left[\theta_T^{\mathrm{nQE}}\right]\mathbb{E}_0^{\mathbb{Q}}\left[\left(Y_T r_T^{\mathrm{QE}}\right)^2\right]}. \tag{140}$$

Our target is to estimate several key components of the costs conditional on information up to time 0 for each QE program, that is, the end of the corresponding purchase phase. In Section 3.4, we generate 1,000,000 factor and nominal pricing kernel paths as well as the corresponding yield curves and QE portfolios returns under $\mathbb{Q}$-measure. We make use of the latent variables we identify to also generate 1,000,000 paths of macro variables we are interest in so that we can estimate the conditional covariance between them and QE portfolio returns. Specifically, we assume the following data-generating process for real GDP growth, inflation, and tax rates, $X_t \in \{\log Y_t/Y_{t-1}, \log P_t/P_{-t1}, \log\theta_t\}$:

$$X_t = c + \sum_i \beta^i F_t^i + u_t, \tag{141}$$

where u_t is possibly serially correlated, and F_t^i is the i-th factor we identify. We specify an AR(1) process for the error: $u_t = \rho u_{t-1} + \eta_t$, where $\mathbb{E}_{t-1}[\eta_t] = 0$. For each QE program, we estimate the model parameters using the tax rate, real GDP growth, inflation and latent factors from November 1971 to the end of QE program purchase phase. We then calculate the corresponding 1,000,000 paths for each macro variable based on the 1,000,000 factor paths generated. The costs and upper bounds are calculated using the realizations of all variables in the 1,000,000 paths.