Who Captures the Rate Cut? Income, Attention, and the Refinancing Channel of Monetary Policy

Jinho Kim[†]

November 8, 2026

Click here for the latest version

Abstract

I document that lower-income households are far less likely to capture the gains from refinancing when interest rates fall, and I quantify how this gap affects the transmission of monetary policy. When refinancing is financially advantageous, bottom-income-quintile borrowers refinance 43 percent less often than the top quintile. They also take longer to act and respond less to monetary policy shocks, even though refinancing would lower their payments the most relative to income. I explain this gradient with a heterogeneous-agent model in which households choose how closely to monitor their mortgage. Differences in refinancing gains account for about one third of the income gradient, and the rest requires a price of attention that falls with income. A notification policy calibrated to field-experimental evidence raises the five-year consumption response to a rate cut by 12.3 percent, even after households reduce their own monitoring, with the largest proportional gains at the bottom of the income distribution.

JEL codes: E52; G51; G21; D15; D83

Keywords: Mortgage Refinancing; Monetary Policy Transmission; Attention; Household Heterogeneity

[†] Economist, Bank of Korea. Ph.D. in Economics, University of California, Davis. Email: jinho.kim.econ@gmail.com.
I am deeply grateful to James Cloyne for his guidance and support, and to Òscar Jordà, Nico Caramp, and Sanjay Singh for valuable comments and guidance. I used an AI language model to help improve grammar and phrasing in the manuscript. All substantive analyses, model specification, and empirical results were produced by the author.

1 Introduction

Mortgage refinancing is one of the largest discretionary cash-flow decisions a U.S. household can make. Nearly 96 percent of outstanding U.S. residential mortgages carry a fixed rate, so existing borrowers benefit from a decline in mortgage rates only if they actively replace their mortgage. Refinancing can generate substantial and persistent reductions in monthly payments, yet many households fail to refinance even when doing so is financially valuable. A large literature attributes this inaction to information frictions, inattention, fixed costs, and behavioral biases (Keys et al. 2016; Agarwal et al. 2016; Andersen et al. 2020; Zhang 2022), and shows that these frictions weaken the refinancing channel of monetary policy (Beraja et al. 2019; Byrne et al. 2023).

Refinancing behavior differs widely across households, and income is a particularly relevant dimension of this heterogeneity for monetary policy: it is systematically related to both refinancing gains and households' marginal propensities to consume (MPCs). Like other dimensions of refinancing heterogeneity, income differences in refinancing are documented in prior work, which typically treats income as one of several fixed borrower characteristics alongside demographics, race, and financial sophistication (Andersen et al. 2020; Gerardi et al. 2023; Agarwal et al. 2024). This treatment leaves open why the income gradient arises, whether lower-income households gain less from refinancing or face larger frictions in acting on the same gain. The answer matters for the refinancing channel. It determines whether an intervention that lowers these frictions could direct more of the benefit of a rate cut toward households whose refinancing translates most strongly into consumption.

This paper documents how refinancing behavior varies with income, uses a structural model to identify the source of this gradient, and quantifies its implications for the transmission of monetary policy. I document that lower-income borrowers act less often on a given rate incentive, take longer to act, and respond less strongly to an identified monetary policy shock. The same gradient appears on all three margins, and the payment savings it leaves unrealized are largest relative to income for the lowest-income borrowers. I then develop and estimate a structural model in which households choose how intensively to monitor their mortgage. The model allows the refinancing hazard to respond endogenously to the gain from refinancing. Estimated on the observed hazard curve, the

model implies that income differences in those gains are not sufficient to explain the data and that the price of attention itself must also fall with income. I argue that an information intervention could strengthen the pass-through of monetary easing to consumption, most in proportion among lower-income borrowers, who face the higher price of attention and spend more of their relief.

The empirical analysis examines three margins: whether borrowers refinance when doing so is financially advantageous, how quickly they act, and how strongly they respond when monetary policy lowers mortgage rates. I use a loan-month panel constructed from the Fannie Mae and Freddie Mac loan-level performance data and compare refinancing behavior across income groups conditional on the financial incentive to refinance and a rich set of borrower, loan, and local market characteristics. I find that the bottom income quintile refinances about 43 percent less often than the top quintile even when refinancing is financially advantageous. The payment savings that bottom-quintile borrowers leave unrealized amount to about 6.7 percent of their monthly income, the highest share among all income groups.

The gradient appears in the timing of refinancing as well as its frequency. Aligning loans at the month they first become financially advantageous to refinance, I find that the share of loans not yet refinanced falls to 80 percent after about four months for the top income quintile, compared with about ten months for the bottom quintile. This delay is consistent with limited attention, in which lower-income borrowers take longer to notice the opportunity. The same gradient appears when monetary policy lowers rates. Following a policy-induced 1 percentage point decline in the 30-year mortgage rate, identified with high-frequency monetary policy surprises, higher-income borrowers refinance substantially more than lower-income borrowers: the Q5–Q1 differential cumulative response peaks at about 4.3 percentage points after two quarters, a sizable difference relative to the average aggregate response of about 2.7 percentage points at the same horizon.

A natural concern is that lower-income borrowers face tighter credit supply, but the income gradient persists in several tests that hold approval conditions fixed. It holds within credit-score bands, within the lowest payment-to-income (PTI) group, where approval constraints should be least binding, and within origination lender-year cells, where it is nearly unchanged from the baseline. The baseline specification already controls for the current loan-to-value (LTV) ratio,

participation in the Home Affordable Refinance Program (HARP), changes in lending standards, and mortgage denial rates by income quintile, and the gradient remains when reporting-time fixed effects absorb changes in credit conditions over time.

I explain this gradient with a structural model in which households choose how closely to monitor their mortgage. The model embeds refinancing in a heterogeneous-agent consumption-saving problem in which households also decide how much home equity to extract when they refinance. To account for the refinancing inertia and delays documented above, I build on the literature that models limited attention as a key refinancing friction. My central departure is make attention endogenous rather than a fixed parameter. With fixed attention, a model cannot separate how much of the income gradient reflects smaller refinancing gains and how much reflects frictions that differ across income at the same gain. It also holds attention fixed when a policy changes the value of refinancing, so it misstates how households respond to that policy.

Households choose how intensively to monitor their mortgage by trading off the gain from refinancing against a convex monitoring cost, and I allow this cost to vary with income. Attention therefore adjusts with the household's mortgage balance, coupon, and liquidity through their effect on the value of refinancing. This makes attention state dependent and nests existing fixed-attention models. The resulting attention policy also generates the shape of the refinancing hazard across rate gaps without assuming a functional form as in existing approaches.

Estimating the model on refinancing hazards by rate gap and income quintile shows how these endogenous gain and attention-cost channels contribute to the observed income gradient. With a common price of attention, differences in refinancing gains still generate the shape of the hazard across rate gaps at every income quintile, but they produce about one third of the log income gradient among loans with a positive rate gap. Matching the remaining two thirds requires the estimated price of attention to decline with income. This estimated income gradient is consistent with existing evidence in the literature and with independent evidence on financial knowledge and mortgage-search behavior.

Finally, I use the model to quantify the monetary-policy consequences of these frictions and an intervention that directly targets attention. I calibrate a market-wide refinancing notification policy

to the field experiment of Byrne et al. (2023). The intervention increases opportunities to reconsider refinancing, while endogenous attention allows households to reduce their own costly monitoring in response. Despite this crowding out, the intervention raises the five-year discounted cumulative consumption response to a rate cut by 12.3 percent. The proportional increase is largest among households with low baseline attention, including those at the bottom of the income distribution. These results show that an information-based intervention can meaningfully strengthen the pass-through of monetary easing to household consumption even after accounting for households' endogenous adjustment of private attention.

Related Literature This paper contributes to three strands of literature. First, it contributes to the literature on refinancing frictions and heterogeneity in refinancing behavior. Existing work documents substantial failures to refinance and relates them to fixed costs, inattention, financial sophistication, and other borrower characteristics (Keys et al. 2016; Andersen et al. 2020; Gerardi et al. 2023; Agarwal et al. 2024). Models of refinancing capture these frictions through exogenous arrival rates (Berger et al. 2021), through borrower types that differ permanently in their propensity to refinance (Zhang 2022; Berger et al. 2024), or by estimating the probability of acting as an exogenous function of demographics (Andersen et al. 2020). These models are built either to match the aggregate refinancing response or, when they study its distribution, to take differences in refinancing propensity across borrowers as given. I instead derive these differences from a household choice: households choose how closely to monitor their mortgage, and this attention responds to the value of refinancing. The model nevertheless reproduces both the level of attention calibrated in Berger et al. (2021) and the income gradient consistent with Andersen et al. (2020), so that parameters these papers take as exogenous emerge here from household optimization with a microfoundation.

The closest modeling antecedents account for refinancing delay through related but different mechanisms. Berger et al. (2025) study the optimal timing of rate observation for a representative risk-neutral borrower. Maxted et al. (2025) generate refinancing delays through present bias: households act only in low-effort windows that arrive at an exogenous rate common to

all households. In my model this rate is itself a choice: households set their attention within a heterogeneous-agent consumption-saving problem, and its distribution across households is estimated from loan-level refinancing behavior. To my knowledge, this is the first paper to endogenize the speed of refinancing in such a model and to use it to quantify monetary policy transmission. This approach captures how households' own attention, and with it their refinancing propensities, adjusts when rates change, a margin that models with fixed attention hold constant.

Second, the paper contributes to the literature on the refinancing channel of monetary policy. A large body of work documents that refinancing frictions attenuate the cash-flow pass-through of rate cuts (Keys et al. 2016; Agarwal et al. 2016; Di Maggio et al. 2017), and that this attenuation is state-dependent and history-dependent through housing equity, rate gaps, and past interest-rate paths (Beraja et al. 2019; Wong 2021; Berger et al. 2021; Eichenbaum et al. 2022). This work characterizes heterogeneity in the refinancing response along loan and regional states. I add the income dimension. Empirically, I show that both the timing of refinancing and the dynamic response to identified monetary policy shocks—margins that have received less attention in the literature—differ substantially across the income distribution.

I then quantify how this heterogeneity shapes monetary transmission through the joint distribution of attention, mortgage balances, and MPCs, which I summarize along the income distribution. A model with homogeneous friction calibrated to the same average refinancing rate would misallocate refinancing across the income distribution, overpredicting refinancing among lower-income households and underpredicting it among higher-income households. This allocation matters for monetary transmission because the cash-flow relief from refinancing interacts with heterogeneity in MPCs: Jørring and Westphal (2025) provide survey evidence that refinancing propensities and MPCs are correlated across households. In my model this joint distribution arises from household choices, and the counterfactuals show that reducing attention frictions raises transmission with the dollar gains concentrated in the second and third quintiles, where attention is weak but balances are large enough for relief to be sizable. The paper thus adds the distribution of attention across households to the loan and regional states that determine the strength of the refinancing channel.

Third, the paper relates to work on the distributional consequences of refinancing heterogeneity. Differences in refinancing behavior redistribute gains across borrowers through mortgage pricing and differential access to refinancing (Zhang 2022; Berger et al. 2024; Fisher et al. 2024), while a broader literature studies how household heterogeneity shapes the incidence of monetary policy (Coibion et al. 2017; Kaplan et al. 2018; Auclert 2019; Andersen et al. 2023). I document this pattern along the income dimension: forgone refinancing savings are substantially larger relative to income for lower-income borrowers. Despite this gradient, a market-wide notification policy has only small effects on the distributions of consumption and liquid wealth, because the cash flow released by refinancing scales with mortgage balances, which are smaller for lower-income borrowers. My results suggest that attention-based interventions strengthen monetary transmission without materially redistributing consumption.

Roadmap The remainder of the paper is organized as follows. Section 2 describes the loan-level data. Section 3 documents refinancing inequality in the data. Section 4 develops a refinancing model. Section 5 estimates the model and reports the policy experiments. Section 6 concludes.

2 Data

2.1 GSE Loan-Level Data

I use loan-level data from Fannie Mae and Freddie Mac, the two government-sponsored enterprises (GSEs). For each loan I observe the contract terms and borrower characteristics recorded at origination—the original balance, note rate, term, loan purpose, credit score, LTV ratio, PTI ratio, occupancy status, property type, and the state and three-digit ZIP code of the property—together with a monthly record thereafter of the outstanding balance, the current coupon rate, loan age and remaining maturity, delinquency status, and the month the loan terminates. I therefore observe the state of the contract in every month in which a borrower could have chosen to refinance.[1]

[1]The GSE data label this variable as the debt-to-income (DTI) ratio. Because the measure is a flow ratio of monthly debt payments to monthly income, I refer to it as the PTI ratio throughout the paper to distinguish it from stock-based measures of indebtedness.

The analysis focuses on 30-year fixed-rate mortgages—the most common mortgage product in the United States—secured by single-family, owner-occupied properties in the fifty states and the District of Columbia, originated between 2000 and 2025. I restrict the sample in two ways. First, I keep only fixed-rate contracts, because the refinancing option this paper studies is defined relative to a coupon that does not adjust. Second, I keep only primary residences, since investor and second-home borrowers face different underwriting standards and refinancing incentives. By construction, the GSE files cover the conventional conforming segment, so jumbo and government-insured mortgages fall outside the sample and the results apply to conforming mortgages.[2] The resulting panel contains 2,577,037 loans and 102,266,735 loan-month observations.

The GSE files do not report borrower income directly. I impute an income proxy by dividing the principal-and-interest payment implied by the origination terms by the underwriting PTI ratio, and assign loans to quintiles within each origination year. Because the reported ratio is a back-end measure, this procedure understates income levels, so the main analysis uses only the within-year rank. Appendix A.1 and Appendix C.1 detail the construction and show that the estimated gradient survives plausible and adversarial quintile misclassification. When income enters in levels, I rescale the proxy using Survey of Consumer Finances (SCF) estimates of the relationship between front-end and back-end PTI ratios by income quintile and origination year, as described in Appendix A.4.

Identifying refinances. The performance files record that a loan terminated early but not why. A zero balance is consistent with a refinance, a home sale, or an outright payoff, and only the first is the object of this paper. Since no field links a prepaid loan to its replacement, I identify refinances by matching each prepaid loan to the new origination that replaced it in the pooled acquisition records, allowing the replacement to be acquired by either enterprise. Potential replacements must agree with the prepaid loan on the location and type of the property and appear within a short window after the payoff month, and are then scored on the characteristics a refinance should preserve. Appendix A.2 describes the procedure, Appendix A.3 the refinancing outcome

[2]Fannie Mae and Freddie Mac purchase conventional mortgages that satisfy the conforming loan limit and their credit standards. Loans insured by the Federal Housing Administration or the Department of Veterans Affairs are securitized through Ginnie Mae and do not appear in these files.

built from it, and Appendix A.1 the sampling design and the design weights that reconstruct the population risk set.

To evaluate the representativeness of the analysis panel, Appendix A.1 compares the weighted distributions of credit score, original balance, LTV, PTI, and geography with those in the full GSE population of 67,267,355 loans. After weighting, the panel reproduces the population moments closely on every dimension, including the two variables that determine the income ranking below.

2.2 HMDA

The Home Mortgage Disclosure Act (HMDA) requires most mortgage originators to report every application and origination, together with the identity of the reporting institution, the loan amount, the applicant's income, the census tract of the property, and the action taken. HMDA therefore supplies two things absent from the GSE files: which institution originated each loan, and how many applications were denied.

Since there is no common identifier for merging HMDA with the GSE files, I match on loan characteristics. Specifically, after restricting both datasets to fixed-rate, single-family, owner-occupied originations, I match records on origination year, loan purpose, state and metropolitan area, the presence of a co-applicant, and the loan amount within a narrow tolerance, and I require the census tract reported in HMDA to lie within the three-digit ZIP code of the GSE loan.[3] To ensure matching accuracy, I retain only pairs that are unique on both sides, so each linked loan corresponds to exactly one HMDA record and each HMDA record to exactly one loan. This process identifies the originating institution for 244,423 loans from 11,058 lenders across Fannie Mae and Freddie Mac, which I use for the lender-by-year fixed-effects exercise in Appendix C.3. Separately, I use the full HMDA files to construct denial rates by HMDA income quintile. These enter the main specifications as a control for time-varying, income-specific credit availability.

[3]Census tracts that span more than one three-digit ZIP code are excluded from this consistency check.

2.3 Supplementary Datasets

In addition to the loan-level data, I use two household surveys to measure objects that the mortgage files cannot: the level of borrower income behind the reported PTI, and the persistence of a household's position in the income distribution.

SCF. The SCF is a triennial cross-section of U.S. household balance sheets sponsored by the Federal Reserve Board, with detailed measures of income, assets, debts, and debt-service payments, and with oversampling of wealthy households. It reports for each household both a housing PTI and a total debt PTI—the front-end and back-end pair that the GSE files do not separate. I use the eight waves from 2001 to 2022, restricted to owner-occupiers of their primary residence who carry a mortgage on it and whose balance falls below that year's conforming loan limit, with income quintiles formed within this mortgagor sample in each wave. The gap between the two ratios across income quintiles yields the correction applied to the level of imputed GSE income in Appendix A.4.

PSID. The Panel Study of Income Dynamics (PSID) follows the same households over time and reports annual family income together with housing tenure and whether the household carries a mortgage, so owner-occupier mortgagors can be identified in each wave. The survey has been biennial since 1997, and consecutive waves are therefore exactly two years apart—close to the median time between origination and refinancing in the matched GSE loans. I use the 2005 through 2021 waves to measure how far mortgagor households move across the income distribution over two years, and use those transitions to bound the concern that an income quintile fixed at origination misstates a borrower's position when the refinancing decision is taken. Results are reported in Appendix C.4.

2.4 Summary Statistics

Table 1 describes the main analysis panel by income quintile at origination. Refinancing rates and observation counts are measured at the loan-month level, and borrower and contract characteristics

at the loan level. The table establishes the scale of the panel, previews the unconditional income gradient in refinancing, and shows how borrower and mortgage characteristics vary across income groups. Refinancing is measured using the record-linkage algorithm employed to construct the panel, and all reported moments are weighted unless explicitly labeled otherwise.[4]

Table 1: Summary Statistics by Income Quintile

	Income quintile					
Variable	Q1	Q2	Q3	Q4	Q5	Total
Monthly refinancing rate, weighted (%)	0.909	1.065	1.173	1.270	1.384	1.144
Panel observations (loan-months)	18,655,265	19,822,500	20,306,162	20,793,970	22,688,838	102,266,735
Unique loan IDs	399,635	453,302	494,830	550,002	679,268	2,577,037
Loan share, unweighted (%)	15.51	17.59	19.20	21.34	26.36	100.00
Loan share, weighted (%)	19.63	20.46	20.34	20.01	19.55	100.00
Credit score	730.7	736.8	741.1	746.4	755.6	742.1
	(54.3)	(52.9)	(51.9)	(50.8)	(47.2)	(52.2)
Original LTV (%)	69.0	71.6	71.6	70.5	66.5	69.9
	(20.4)	(17.9)	(17.2)	(16.8)	(17.4)	(18.1)
PTI (%)	39.8	36.6	34.3	31.5	24.5	33.4
	(9.4)	(9.8)	(10.1)	(10.1)	(9.7)	(11.1)
Loan balance at first observation ($)	103,146	157,144	204,625	257,897	324,709	208,611
	(43,601)	(63,443)	(88,681)	(118,691)	(161,527)	(128,456)
Adjusted annual income, SCF-corrected ($)	26,344	47,808	69,909	96,308	162,027	80,129
	(7,623)	(8,342)	(11,333)	(16,012)	(97,576)	(64,278)
Monthly payment ($)	592	906	1,182	1,497	1,960	1,225
	(205)	(289)	(407)	(560)	(853)	(696)
Estimated house value ($)	166,890	236,694	304,823	388,723	527,198	324,073
	(102,616)	(129,130)	(165,366)	(240,893)	(325,028)	(242,287)
Original interest rate (%)	5.03	4.99	4.96	4.92	4.82	4.94
	(1.44)	(1.43)	(1.44)	(1.44)	(1.46)	(1.44)
Remaining maturity (months)	317.6	312.8	309.7	304.7	287.4	306.5
	(74.9)	(77.8)	(79.7)	(82.5)	(89.3)	(81.6)

Notes: The table reports summary statistics for fixed-rate, single-family, owner-occupied GSE mortgages originated from 2000 through 2025. Unless otherwise indicated, entries apply inverse-probability sampling weights that reconstruct the population risk set. Loan-level characteristics are measured at origination or, where indicated, at the first panel observation; standard deviations are in parentheses. The monthly refinancing rate is the share of active loans that refinance in a given month. Income quintiles are assigned from estimated annual income at origination. SCF-corrected income rescales income inferred from PTI using origination-year and income-quintile adjustment factors derived from the SCF. See Appendices A.1 to A.3 for details on panel construction, sampling weights, the matching procedure, and the refinancing outcome.

The panel contains 2,577,037 loans and 102,266,735 loan-months. Because loans with a uniquely matched refinance are retained in full while all others are sampled, raw loan shares rise from 15.5 percent in the bottom quintile to 26.4 percent in the top. The inverse-probability sampling weights restore population shares between 19.6 and 20.5 percent across all five quintiles. The monthly refinancing rate, the share of active loans that refinance in a given month, rises monotonically

[4] Appendix A.1 details the panel construction and sampling weights, while Appendix A.2 and Appendix A.3 describe the matching procedure and refinancing outcome.

from 0.91 percent in the bottom quintile to 1.38 percent in the top, compared with 1.14 percent in the panel as a whole. The Q5–Q1 difference is 0.48 percentage points, or 52 percent of the bottom-quintile rate.

Loan characteristics display distinct patterns across income groups. Mean loan balance at the first panel observation rises from $103,146 in Q1 to $324,709 in Q5, while mean credit score rises from 730.7 to 755.6. Mean original interest rate instead declines from 5.03 percent in Q1 to 4.82 percent in Q5, a difference of 21 basis points, and mean remaining maturity declines from 317.6 months to 287.4 months, a difference of approximately 30 months. Original leverage follows a hump-shaped pattern. Mean LTV is 69.0 percent in Q1, rises to 71.6 percent in Q2 and Q3, declines to 70.5 percent in Q4, and reaches its lowest level, 66.5 percent, in Q5.

The levels of key mortgage characteristics in Table 1 closely align with independently constructed U.S. mortgage datasets used by Keys et al. (2016) and Gerardi et al. (2023).[5] All else equal, the financial gain from refinancing is larger when the outstanding balance is higher, the current coupon is higher, or the mortgage has more remaining months. Credit score and LTV also enter lenders' underwriting and pricing decisions and may therefore affect both refinancing eligibility and the obtainable rate. The regressions that follow control for loan balance, the rate incentive, loan age, credit score, original LTV, and estimated current LTV.

2.5 Identifying In-the-money Loans

Because the paper studies the refinancing channel of monetary policy, the empirical analysis focuses on refinancing opportunities created by lower mortgage rates. A borrower may refinance for other reasons, such as extracting home equity, even when a lower rate is unavailable. Such refinancings are not informative about whether the borrower acts on the financial incentive created by a decline in mortgage rates. I therefore identify the loan-months in which refinancing offers a rate-based financial gain and compare take-up across income groups within those opportunities.

For each loan-month, I define the rate gap as $g_{it} \equiv m_i^* - \tilde{m}_{it}$, where m_i^* is the borrower's current coupon and $\tilde{m}_{it}$ is the rate the borrower could obtain on a new mortgage. I estimate

[5] Keys et al. (2016) report an average unpaid balance of $205,218, an origination credit score of 737, and an origination LTV of 70.7 percent. Gerardi et al. (2023) report corresponding means of $212,000, 744, and 72.6 percent.

$\tilde{m}_{it}$ as the median origination rate among newly originated loans within a granular borrower–loan characteristic cell defined by origination quarter, state, credit-score, LTV ratio, and income quintile, following Defusco and Mondragon (2020). This approach uses rates obtained by observationally similar borrowers in the same local market to construct the refinancing interest rate available to each borrower.

I classify a loan as in the money when its rate gap exceeds a refinancing threshold $\bar{g}_{it}$:

$$\text{In-the-money} \quad \text{if} \quad g_{it} = m_i^* - \tilde{m}_{it} > \bar{g}_{it}. \tag{1}$$

In the baseline specification, I set $\bar{g}_{it} = 0$, so a loan is in the money whenever the current coupon exceeds the estimated obtainable rate. I also use the Agarwal–Driscoll–Laibson (ADL) threshold from Agarwal et al. (2013), $\bar{g}_{it} = \bar{g}_{it}^{ADL}$, which accounts for refinancing costs, expected exogenous prepayment, remaining maturity, and interest-rate volatility. This second definition requires the rate gap to exceed the loan-specific threshold at which refinancing is financially worthwhile.

The ideal measure of $\tilde{m}_{it}$ would be the rate each borrower is offered when the loan comes into the money. Offered rates are not observed in loan-level data: rates are recorded only for loans that are originated, so $\tilde{m}_{it}$ reflects the distribution of rates conditional on approval and take-up. If approved borrowers obtain better terms than the marginal applicant, $\tilde{m}_{it}$ understates the obtainable rate and overstates the rate gap. The cell definition limits this concern: 554,735 of the 663,000 possible cells contain refinance originations, so each borrower is compared with refinancers who match them closely on these characteristics. Also, because the ADL threshold lies above zero, a loan that an overstated gap would classify as in the money under that definition still has a positive true rate gap. I therefore report results under both definitions. Appendix A.5 provides the ADL formula and additional details.

3 Evidence of Refinancing Inequality

3.1 The Income Gradient in Refinancing

Average refinancing probability. To quantify the income gradient in refinancing intensity, I estimate a linear probability model (LPM) of monthly refinancing on loan-months in which the loan is financially advantageous to refinance.[6] Conditioning on in-the-money status focuses the analysis on whether households act on financially profitable opportunities and sets aside refinancing driven by idiosyncratic liquidity needs such as equity extraction.[7]

$$Refi_Indicator_{it} = \alpha + \sum_{q=2}^{5} \beta_q \cdot \mathbf{1}[IncomeQuintile_i = q] + \Gamma' X_{it} + \nu_{z(i)} + \lambda_{v(i)} + \epsilon_{it} \tag{2}$$

The dependent variable equals one if loan i refinances in month t, and β_q measures the refinancing propensity of quintile q relative to the bottom quintile, conditional on being in the money. Cross-quintile differences in refinancing propensity could reflect differences in observable borrower and loan characteristics rather than an independent income effect. I therefore include a rich set of controls X_{it}—the rate gap, credit score, loan age, loan balance, and the estimated LTV, each entered as categorical variables to allow flexible nonlinear effects—together with ZIP-code fixed effects $\nu_{z(i)}$ and origination vintage year-quarter fixed effects $\lambda_{v(i)}$. I additionally control for two time-varying confounders. To absorb the equity channel of Beraja et al. (2019), I construct a monthly loan-level LTV from origination LTV, the current balance, and the Federal Housing Finance Agency (FHFA) three-digit-ZIP house price index, and I control for participation in HARP, which allowed underwater borrowers to refinance from 2009 to 2018. To absorb income-specific credit supply, I control for quarterly changes in the Senior Loan Officer Opinion Survey (SLOOS) credit-tightness index and for mortgage denial rates by HMDA income quintile. The HARP and SLOOS terms are interacted with income-quintile indicators, so their effects may differ across quintiles.

[6] Results are robust to a logit specification. Because income enters through categorical quintile indicators, the LPM imposes no functional-form restriction on the relationship of interest.

[7] For example, households facing idiosyncratic income shocks may refinance to extract home equity regardless of the rate environment, as documented by Chen et al. (2020). Including such refinancings would conflate distinct behavioral margins.

Table 2: Monthly Refinancing Probability: Simple Rate Gap

Dependent	Refinancing Indicator × 100			
Equation	(1)	(2)	(3)	(4)
Income Quintile 2	0.432[a]	0.377[a]	0.374[a]	0.370[a]
	(0.028)	(0.020)	(0.018)	(0.018)
Income Quintile 3	0.746[a]	0.646[a]	0.634[a]	0.540[a]
	(0.046)	(0.031)	(0.027)	(0.024)
Income Quintile 4	1.077[a]	0.942[a]	0.918[a]	0.735[a]
	(0.068)	(0.048)	(0.041)	(0.038)
Income Quintile 5	1.514[a]	1.348[a]	1.300[a]	1.054[a]
	(0.088)	(0.066)	(0.056)	(0.049)
Control variables				
LTV, credit, loan age, rate inc., ZIP		x	x	x
Credit tightness, HARP			x	x
Loan balance, Est. LTV				x
Observations	47,415,612	47,415,612	47,114,678	47,111,510

Notes: Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively. Standard errors are in parentheses. This table reports weighted linear-probability-model coefficients on income-quintile indicators relative to the bottom quintile. The sample is restricted to in-the-money observations under the simple rate-gap definition. The dependent variable is monthly refinancing, scaled by 100, so coefficients are percentage-point differences. Column (1) includes origination-vintage year-quarter fixed effects; columns (2)–(4) also include ZIP-code fixed effects and add the controls indicated in the table. Standard errors are clustered by ZIP code and origination vintage year-quarter. Source: Fannie Mae and Freddie Mac loan-level panel, 2000–2025.

Table 2 reports estimates under the simple rate-gap definition. The unconditional top-to-bottom gap in monthly refinancing probability is 1.51 percentage points and narrows to 1.05 percentage points under the full set of controls. Figure 1 plots the corresponding predicted monthly refinancing probabilities by income quintile as controls are added across specifications. In the full-control specification, the predicted refinancing rate is 1.42 percent for the bottom quintile and 2.48 percent for the top, so bottom-quintile borrowers refinance roughly 43 percent less often than top-quintile borrowers facing an equivalent rate incentive. Predicted refinancing rates increase monotonically with income in every specification, and the gradient remains after the full set of controls is added. The pattern is unchanged under the stricter ADL-adjusted in-the-money criterion of Agarwal et al. (2013), maintaining the same size of the full-control gap. Details are reported in Appendix B.2.

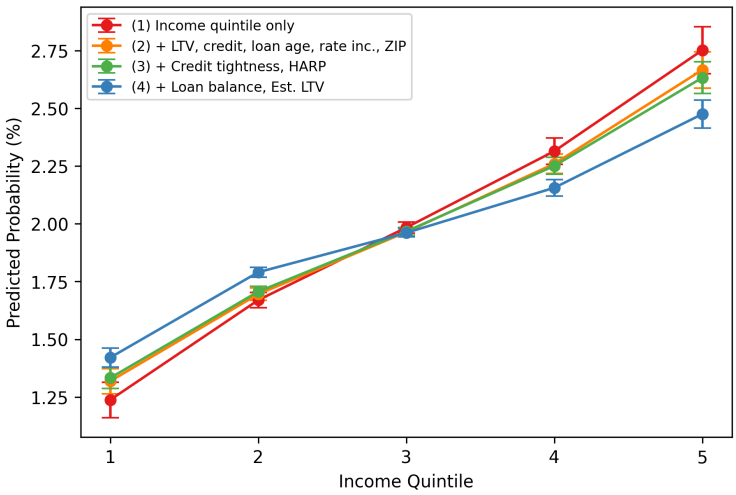


Figure 1: Predicted Refinancing Probability by Income Quintile

Notes: The figure reports predicted monthly refinancing probabilities by income quintile for fixed-rate, single-family, owner-occupied GSE mortgages that are in the money under the simple rate-gap definition. Lines correspond to specifications with successively more controls. Specifications include origination-vintage fixed effects, and all but the first also include ZIP-code fixed effects. Error bars are 95% confidence intervals based on standard errors clustered by ZIP code and origination vintage year-quarter.

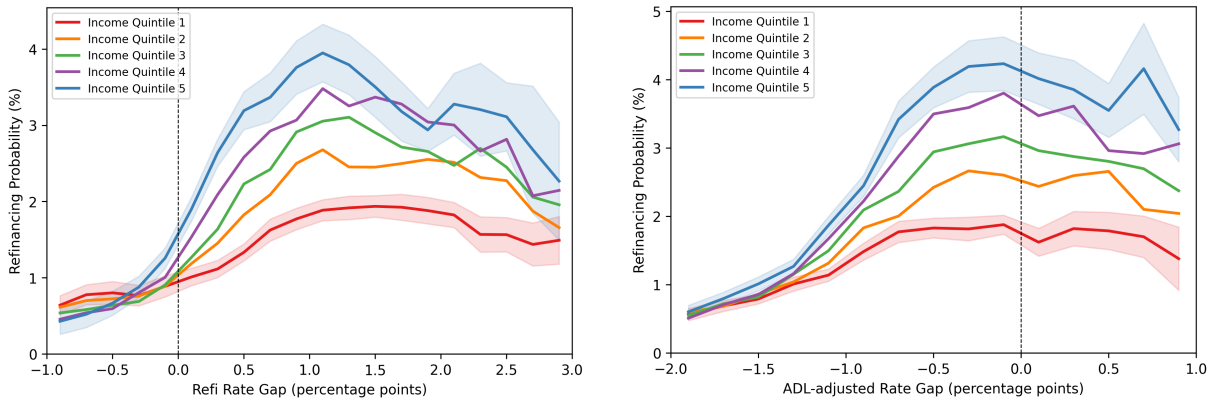
Refinancing probability by rate gap. The estimates above condition on in-the-money status but pool across the full range of rate gaps. I next examine whether the income gradient varies with the strength of the refinancing incentive. Comparing borrowers within rate-gap bins narrows differences in the financial incentive they face, while the pattern across bins shows whether the income gap changes as refinancing becomes more valuable. Further, the shape of the gradient across bins is informative about the mechanism. Differences confined to small positive gaps could reflect uncertainty about whether refinancing is worthwhile after costs, whereas differences that persist at larger gaps could indicate that some borrowers do not reconsider refinancing even when the financial incentive is strong. I therefore estimate

$$\begin{aligned} Refi_Indicator_{it} = \alpha + \sum_{q \in Q} \sum_{k \in K} \beta_q^k \cdot \mathbf{1}[RateGap_{it} \in k] \cdot \mathbf{1}[IncomeQuintile_i = q] \\ + \Gamma' X_{it} + \nu_{z(i)} + \lambda_{v(i)} + \epsilon_{it} \end{aligned} \tag{3}$$

with the same controls and fixed effects as (2), except that the rate incentive is replaced by the rate-gap bins themselves. Here q indexes income quintiles, k indexes rate-gap bins, and one cell is omitted as the reference category. Because the bin-by-quintile cells are fully saturated, the

specification nests the level effects of both the rate gap and income as well as their interaction. Figure 2 reports the fitted hazards. In the out-of-the-money region, refinancing rates are similar across income quintiles. Once loans move into the in-the-money region, the quintiles separate, and the income gap widens as the rate gap increases. The gradient is not confined to borrowers near the refinancing threshold, where uncertainty about whether refinancing is worthwhile after costs may be most relevant. It remains pronounced at larger positive gaps, where the financial incentive to refinance is stronger. This pattern is consistent with the possibility that borrowers differ in whether they become aware of refinancing opportunities even when the potential gains are large.

Panel (b) shows that the same pattern holds under the ADL-adjusted rate gap. Rather than applying a common threshold to all loans, the ADL criterion evaluates the refinancing threshold separately for each loan-month observation using loan-specific features such as the outstanding balance and remaining maturity together with the interest-rate environment. The income gradient therefore persists when the financial incentive to refinance is adjusted at the loan-month level for these differences.



(a) Simple Rate Gap ($\bar{g}_{it} = 0$)

(b) ADL-adjusted Rate Gap ($\bar{g}_{it} = \bar{g}_{it}^{ADL}$)

Figure 2: Refinancing Inequality by Rate Gap

Notes: The figure reports fitted monthly refinancing hazards by rate-gap bin and income quintile from the full-control specification in (3). The sample comprises fixed-rate, single-family, owner-occupied first-lien GSE mortgages in the fifty states and the District of Columbia, 2000–2025, with nonmissing income quintile and rate gap. The specification is weighted and includes ZIP-code and origination-vintage year-quarter fixed effects. Standard errors are clustered by ZIP code and origination vintage year-quarter. Coefficients are estimated on a random 15-million-loan-month subsample and centered on the weighted full-panel hazard at a zero rate gap. Panel (a) uses the simple rate gap with a zero refinancing threshold. Panel (b) uses the ADL-adjusted rate gap. Shaded regions denote 95% confidence intervals.

Robustness. I consider three potential concerns to verify the robustness of the findings, with additional results reported in the appendix. One concern is that the gradient reflects unobserved aggregate conditions rather than differences in behavior across borrowers, since refinancing waves and rate environments vary over time. I re-estimate the full-control specification replacing origination-vintage fixed effects with reporting-time fixed effects at the month, quarter, or year level, which absorb aggregate refinancing cycles, the current rate environment, and any other macroeconomic conditions common to borrowers observed at the same point in time. The gradient remains positive, monotone, and precisely estimated, with a top-quintile coefficient of 0.91 to 1.31 percentage points across the two gap definitions in the specification with monthly fixed effects (Appendix B.3).

I also examine the possibility that the gradient reflects lender-side underwriting or credit supply, not borrower-side behavior. The gradient persists within credit-score bands, a key underwriting input (Figure C.1), and within PTI bins, including the most underwriting-favorable group with PTI below 36 percent, where approval constraints should be least binding (Table C.2). Using the HMDA link to identify each loan's origination lender, the gradient also persists within origination lender-year cells, with a top-quintile coefficient of 1.04 percentage points against 1.05 in the baseline (Appendix C.3). Sorting across lenders and lender vintages therefore does not account for the pattern, although the data do not observe refinance-stage outreach or screening.

One might be worried that the income quintiles, which are fixed at origination, misclassify borrowers at the time of the refinancing decision. PSID transition rates over a two-year horizon, which brackets the median and mean loan age at refinancing, show that moves between the endpoint groups are rare and that about 80 percent of households remain in the same broad income group (Table C.5). Directly perturbing the imputed income ranking with SCF-based correction factors, which reassigns roughly half of loans across quintiles, still preserves a Q5–Q1 gap of 0.77 to 0.98 percentage points, while randomly assigned quintiles eliminate it (Appendix C.1).

3.2 Delays in Refinancing Decisions

The preceding analysis documents how refinancing propensities vary across income groups and rate gaps, but it does not capture when borrowers pull the trigger after a refinancing opportunity first arises. This timing margin is economically important. Each month of delay leaves the borrower making payments on the higher-rate mortgage and therefore adds to forgone savings even if the borrower eventually refinances. Delay can also cause a borrower to miss the refinancing opportunity if market rates rise and the loan moves out of the money before the borrower acts. I therefore examine how the time from first becoming in the money to mortgage exit varies across income groups. To my knowledge, this is the first loan-level evidence on systematic differences across borrower groups in the time taken to act on a refinancing opportunity.

I align all loans at the month in which they first become in the money and compute weighted Kaplan–Meier survival curves by income quintile. In-the-money status is based on the ADL-adjusted rate gap of Section 2.5, which accounts for the declining value of refinancing as a loan approaches maturity.[8]

Figure 3 shows a pronounced income gradient in refinancing speed. Twelve months after first becoming in the money, 22.8 percent of bottom-quintile mortgages have prepaid, compared with 42.5 percent in the top quintile. After twenty-four months, the shares are 36.3 and 59.6 percent. At a fixed threshold, the top quintile reaches a cumulative prepayment rate of 20 percent after roughly four months, while the bottom quintile takes about ten months. Averaged across quintiles, 33.2 percent of in-the-money mortgages prepay within one year and 48.9 percent within two. These magnitudes are broadly comparable to Keys et al. (2016), who find that approximately 60 percent of mortgages identified as in the money prepay over the subsequent two years.[9]

As a robustness check, I ask whether slow refinancers eventually catch up. I collapse the first year after the opportunity arises into a single take-up indicator that treats refinancing in the first

[8]Each loan enters the risk set in the first month in which it is in the money under the ADL-adjusted definition, and the clock runs from that month regardless of whether the loan subsequently moves out of the money. Each loan therefore contributes a single spell. The event is voluntary prepayment, and loans that do not prepay before their last observed month are right-censored.

[9]The difference may partly reflect the sample periods. Keys et al. (2016) study loans active in December 2010 and follow them through December 2012, a period in which mortgage rates fell from already low levels to historical lows. In contrast, my 2000–2025 panel spans a much broader range of interest-rate environments, including the relatively high rates of the early 2000s.

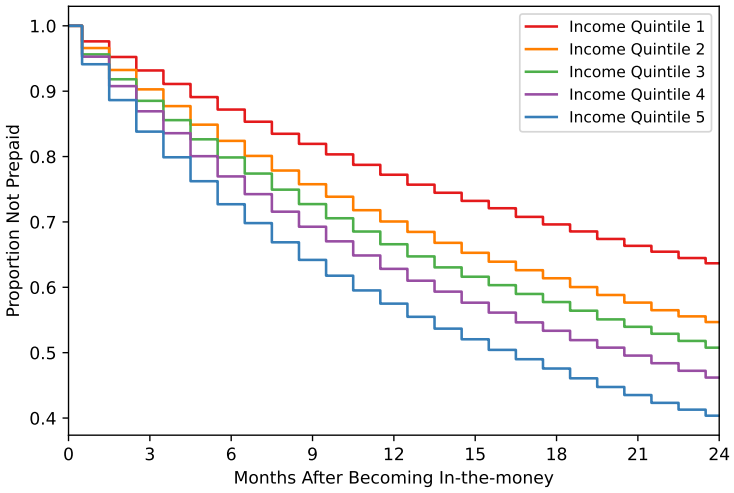


Figure 3: Proportion of Remaining Loans by Months After Becoming In the Money

Notes: The figure reports weighted Kaplan–Meier survival curves by income quintile for fixed-rate, single-family, owner-occupied GSE mortgages. Month zero is the first month in which a loan becomes in the money under the ADL-adjusted rate-gap definition. The event is voluntary prepayment. The vertical axis is the estimated share not yet prepaid, so a lower curve indicates faster mortgage exit. Loans that do not prepay before their final observed month are right-censored.

and twelfth months alike. Delay within the year is therefore ignored, so any gradient arising solely from differences in timing should disappear under this measure. The income gradient instead remains monotone and significant under both in-the-money definitions when I regress twelve-month take-up on income-quintile indicators using the same controls and fixed effects as in the monthly specification. The results therefore indicate that lower-income borrowers are not only slower to refinance but also less likely to act at all, with a substantially larger share still inactive one year after the opportunity arises. Details are reported in Appendix B.4.

3.3 Unrealized Savings from Refinancing

The preceding results establish that lower-income borrowers refinance less often and wait longer after a refinancing opportunity arises. To assess the economic magnitude of these differences, I next measure the payment savings borrowers forgo when they do not refinance. This matters for monetary transmission because refinancing is one way in which lower market rates raise household disposable income, and the size of the missed payment reduction determines how much of that cash-flow benefit a borrower fails to capture.

I quantify these forgone savings using one-to-one matched old-to-new loan pairs, which measure the interest-rate reduction realized at refinancing. I estimate rate reductions by income group with the specification in (2), using the difference between the old and new interest rate as the dependent variable and restricting the sample to in-the-money observations. Applying each group's predicted reduction to the borrower's current balance under a 30-year refinance term yields a counterfactual payment reduction for every in-the-money loan-month, which I express relative to monthly income. Because income inferred from the GSE back-end PTI ratio understates borrower income, I rescale it with SCF-based correction factors by income quintile and origination year (Appendix A.4).

Table 3: Unrealized Savings from Refinancing by Income Quintile

Income Quintile	Avg. Income ($)	Avg. Balance ($)	Est. Rate Reduction (pp)	Mthly. Pymt. Reduction ($)	Reduction / Income (%)
1	24,052	76,692	0.91	130.38	6.71
2	45,175	119,868	0.94	200.20	5.37
3	66,583	156,384	0.96	261.23	4.75
4	91,971	196,869	1.00	337.76	4.44
5	153,473	240,069	1.06	478.92	4.01

Notes: Rate reductions are weighted adjusted predictions from the full-control old-to-new rate-reduction regression estimated among one-to-one matched refinances that are in the money. The regression includes the same full set of controls and fixed effects as column (4) of Tables 2 and B.2. Predictions are applied to in-the-money at-risk loan-months defined using the income-adjusted obtainable-rate gap. Payment savings compare the existing payment at remaining maturity with a counterfactual refinance that resets maturity to 30 years. This reset follows common practice: in Freddie Mac's Quarterly Product Transition Report, 68 to 69 percent of refinancing borrowers in the 2012 Q3-Q4 kept the same product term as the loan they paid off and 27 to 29 percent shortened it. Average income and the reduction relative to income use SCF-corrected borrower income.

Table 3 shows that the bottom and second income quintiles could have saved 6.7 percent and 5.4 percent of corrected monthly income, respectively, through refinancing. This is a significant amount, as it represents additional disposable income occurring in subsequent months, not just a one-time saving. Although the absolute dollar amount of missed gains can be larger for higher-income borrowers, the forgone savings constitute a much larger share of income for lower-income mortgagors.

As a robustness check, I also estimate unrealized savings using an alternative approach: for each in-the-money loan-month, I construct a counterfactual refinancing interest rate by subtracting the income-group-specific predicted rate gap from the borrower's current coupon, and translate

the implied payment reduction into dollars using the loan-specific balance, the existing loan's remaining maturity, and a 30-year counterfactual refinance term. This alternative procedure yields similar estimates, with the bottom quintile facing forgone monthly savings of approximately 6.9 percent of corrected income, corroborating the matched-loan approach (Appendix C.7). These findings suggest that refinancing inequality could materially weaken the pass-through of monetary policy to disposable income. Its implications for aggregate consumption depend on how these forgone savings are distributed across households with different MPCs, a margin I examine in the structural model below.

3.4 Differential Refinancing Response to Monetary Policy Shocks

The evidence above conditions on each borrower's in-the-money status and isolates the friction channel cross-sectionally. A separate question is whether lower-income borrowers also respond less to the same monetary policy shock in real time, when the Federal Reserve cuts rates and refinancing opportunities open for the broad market. Prior work documents the aggregate dynamic refinancing response to monetary policy (Eichenbaum et al. 2022; Beraja et al. 2019; Wong 2021) and cross-sectional heterogeneity in refinancing rates (Andersen et al. 2020; Gerardi et al. 2023; Agarwal et al. 2024), but has not, to my knowledge, estimated how the dynamic refinancing response to identified monetary policy shocks varies across borrowers. This subsection estimates that gradient.

To estimate this response, I aggregate the GSE loan-level panel into a quarterly regional panel. For each three-digit ZIP code, quarter, and income quintile, I compute the stock of outstanding fixed-rate, single-family, owner-occupied mortgages and the number of rate refinances originated in that quarter.[10] The dependent variable is the cumulative refinancing rate over the h quarters following a shock,

$$\mathcal{R}^{h}_{zqt} = \sum_{k=0}^{h} 100 \times \frac{x_{zq,t+k}}{y_{zq,t+k}}, \tag{4}$$

[10] I exclude HARP refinances to eliminate potential bias driven by the program and start the estimation sample in 2004 so that the outstanding stock can build up.

where x_{zqt} is the number of refinances and y_{zqt} the outstanding stock in region z, income quintile q, and quarter t, comparable to the construction in Eichenbaum et al. (2022).

I estimate the income gradient with a local-projection instrumental-variable design in the spirit of Jordà (2005), stacking the five income-quintile cells within each ZIP-quarter. For each horizon h, I estimate

$$\mathcal{R}^h_{zqt} = \alpha^h_{zq} + \lambda^h_t + \beta_h \left(\Delta m_t \times \tilde{q}\right) + \sum_{\ell=1}^{6} \beta_{\ell h} \left(\Delta m_{t-\ell} \times \tilde{q}\right) + \sum_{\ell=1}^{6} \Gamma^h_\ell X_{zq,t-\ell} + \epsilon^h_{zqt}. \tag{5}$$

Here Δm_t is the change in the 30-year mortgage rate, normalized so that a positive value is a decline, and $\tilde{q} = (q-1)/4$ is a normalized income rank, so β_h measures the Q5–Q1 differential refinancing response to a 1 percentage point policy-induced decline in the mortgage rate. ZIP-by-income fixed effects absorb level differences across cells, and quarter fixed effects absorb all aggregate dynamics, so identification comes from whether higher-income cells respond more strongly within the same quarter. The rate-change interactions are instrumented with the high-frequency surprises of Swanson (2021) interacted with income rank. Controls include six lags of refinancing rates, local house-price growth and unemployment, aggregate output growth, and macroeconomic expectations from the Survey of Professional Forecasters (SPF) following Nakamura and Steinsson (2018). Regressions are weighted by the outstanding stock, and standard errors are two-way clustered by region and quarter.

Figure 4 reports the estimated differential. The gap peaks at approximately 4.3 percentage points after two quarters, which is large relative to the average aggregate refinancing response of about 2.7 percentage points at the same horizon. The aggregate refinancing response to monetary easing is therefore concentrated among higher-income borrowers. The gradient is also consistent across two empirical strategies with different identifying variation, the borrower-level rate-gap design above and the monetary-shock interaction design here, which weakens explanations specific to either approach. First-stage estimates, the average aggregate response, and robustness across alternative response measures and income-group definitions are reported in Appendix D.

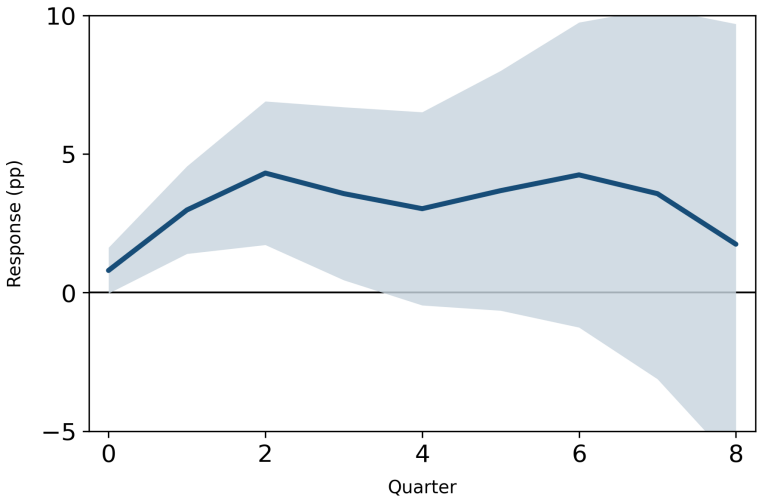


Figure 4: Q5–Q1 Differential Refinancing Response to a Monetary Policy Easing

Notes: The figure reports local-projection instrumental-variable estimates of the Q5–Q1 differential cumulative refinancing response to monetary easing. The sample is a quarterly regional panel of refinancing rates by income quintile. The specification includes ZIP-by-income-quintile and quarter fixed effects, six lags of refinancing and local macroeconomic controls, and is weighted by the outstanding mortgage stock. High-frequency monetary-policy surprises instrument mortgage-rate changes interacted with income rank. The shock is normalized to a 1 percentage point decline in the 30-year mortgage rate. Shaded regions are 95% confidence intervals based on two-way clustering by region and quarter.

4 A Model of Endogenous Refinancing Attention

I develop a structural model of household mortgage refinancing to measure how the gaps documented above change the transmission of monetary policy. The model has two goals. First, it must reproduce the refinancing hazards across income groups and rate gaps, including their low levels and substantial delays even when refinancing is financially valuable. Second, it should map refinancing into household consumption through a consumption-saving problem in which households also choose liquid saving and, when they refinance, optimal cash-out. These two pieces allow the model to trace heterogeneous refinancing behavior into the consumption response to a rate cut.

The evidence above points to a borrower-side friction that remains binding even when refinancing gains are large. Hazards remain far below one even deep in the money, and borrowers who do refinance often act with substantial delay. A fixed refinancing cost alone cannot generate either pattern: it implies a threshold rule under which borrowers refinance as soon as the gap exceeds the threshold. Limited attention generates both, because borrowers act only when they consider

refinancing. The literature typically models it with exogenous consideration rates (Andersen et al. 2020; Berger et al. 2021; Zhang 2022; Berger et al. 2024), which captures this inertia but holds the consideration process fixed as the household's mortgage and financial state evolves. In the simplest case of a constant rate common to all borrowers (Berger et al., 2021), the in-the-money hazard is flat across both income and rate gaps, which is inconsistent with the evidence in Section 3.[11]

I instead make monitoring a household choice. Borrowers choose how intensively to monitor by weighing the value of a refinancing opportunity against a convex attention cost that may vary by income state. The resulting attention policy varies with the household's state through the value of refinancing, so a change in mortgage balance, coupon, liquidity, or market rates can change the refinancing hazard even for the same household. This distinction matters for counterfactuals. When interest rates or refinancing policies change, the model re-optimizes attention, whereas a fixed-attention model holds the consideration rate unchanged. I also show that the exogenous-attention specification is nested as a limiting case.

4.1 Model Setup

The household side is a continuous-time incomplete-markets model with heterogeneous households and idiosyncratic income risk, in the tradition of Aiyagari (1994). I use the model in partial equilibrium, taking the aggregate short-rate process as exogenous and controlled by the monetary authority. Households are heterogeneous with respect to liquid wealth, outstanding mortgage balances, coupon rates, and income. Each household holds a long-term fixed-rate mortgage and can choose whether to refinance, including the option to extract home equity through cash-out refinancing. Refinancing combines a fixed cost with the endogenous monitoring choice. A household's individual state is $x \equiv (W, F, m^*, s)$, its liquid wealth, outstanding mortgage balance, coupon rate, and income state s with income Y_s. The short rate r is the only aggregate state, and I write $V(r, x)$ for the household's value function.

The market mortgage rate is given by a pricing schedule $m(r)$ constructed from risk-neutral lender pricing under an exogenous benchmark prepayment hazard. The schedule maps the

[11] The constant-rate restriction in Berger et al. (2021) is a deliberate simplification for studying the aggregate refinancing channel (see footnote 17).

exogenous short rate into the mortgage rate faced by households. I hold this mapping fixed when solving the household problem, so changes in endogenous refinancing attention do not feed back into mortgage rates.

Housing collateral enters as a permanent household type. Each household belongs to one of three collateral types k, and its house, worth H_k, backs a mortgage balance of at most $\bar{F}_k \equiv \phi^{LTV} H_k$. The three house values are the median estimated house values of the first, third, and fifth income quintiles in the loan panel, and the initial joint distribution of collateral types and income is taken from the panel, so higher-income households more often hold the larger types. Collateral is fixed from then on, while the household's income state continues to evolve. Fixing it abstracts from the long-run link between income and house values. However, the focus of this paper is not on how macroeconomic conditions drive house price appreciation and the resulting changes in home equity, which may shape the transmission of monetary policy. Instead, I isolate the consumption effects of heterogeneous refinancing behavior over a medium-term horizon of five years, holding house prices fixed. This simplification allows me to focus on how differences in refinancing access and behavior shape the pass-through of rate changes to household consumption.

4.2 Uncertainty Environment

The model features both idiosyncratic and aggregate uncertainty. Households face idiosyncratic labor income risk, following a first-order autoregressive, or AR(1), process:

$$d \ln Y_{it} = -\eta_Y \left(\ln Y_{it} - \overline{\ln Y} \right) dt + \sigma_Y \, dZ_{it}^Y \tag{6}$$

Each household's labor income Y_i evolves stochastically according to a mean-reverting process with mean-reversion speed η_Y and volatility σ_Y. The innovation is driven by a standard Brownian motion Z_{it}^Y that is independent across households and orthogonal to all aggregate shocks. Throughout I work with a finite-state approximation to (6): $s \in \{1, \ldots, N_s\}$ indexes the income state, Y_s is the corresponding income level, and $\Theta_{ss'}$ is the transition rate from state s to state s'.

The five-state chain used below has persistence and dispersion chosen so that it reproduces the targeted mean, standard deviation, and half-life of income.

The aggregate short-term interest rate r_t follows a continuous-time Cox–Ingersoll–Ross (CIR) process:

$$dr_t = -\eta_r(r_t - \bar{r})\,dt + \sigma_r\sqrt{r_t}\,dZ_t^r. \tag{7}$$

It is characterized by mean-reversion speed η_r, volatility σ_r, and long-run mean $\bar{r}$. The process is driven by a common Brownian motion Z_t^r, which captures aggregate shocks affecting all households simultaneously. Changes in r_t propagate through to mortgage rates and rate gaps, thereby linking aggregate monetary conditions to household-level refinancing dynamics.

4.3 Mortgage-Rate Schedule

Each mortgage is modeled as a fixed-rate loan contract characterized by its coupon rate m^* and outstanding balance F. Instead of specifying a detailed amortization schedule, the model assumes that a constant fraction α of the loan is repaid each year. This simplification greatly reduces computational complexity without affecting the key results. When a household refinances, the coupon is reset to the prevailing market mortgage rate $m(r_t)$.

I construct the mortgage-rate schedule from risk-neutral lender pricing under a benchmark prepayment hazard that does not depend on household characteristics. Lenders fund at the short rate, which follows the CIR process above. A newly issued loan amortizes at rate α, prepays for exogenous reasons such as moving at rate ν_m, and prepays through refinancing at a benchmark rate $\bar{\chi}$ whenever the market rate has fallen below its coupon, so the benchmark hazard is $\bar{\vartheta}(r_t; m^*) = \nu_m + \bar{\chi}\,\mathbf{1}\{m(r_t) < m^*\}$. Let τ denote the prepayment time generated by this hazard. The value to the lender of a unit balance with coupon m^* when the short rate is r is

$$P(r, m^*; \bar{\vartheta}) = \mathbb{E}\left[\int_0^\tau e^{-\int_0^t r_u\,du} e^{-\alpha t}\,(m^* + \alpha)\,dt + e^{-\int_0^\tau r_u\,du} e^{-\alpha\tau}\,\middle|\, r_0 = r\right], \tag{8}$$

where the first term is interest and scheduled amortization and the second is the remaining balance repaid at par when the loan prepays. The market mortgage rate $m(r)$ is the coupon at which a lender breaks even at issuance, $P(r, m(r); \bar{\vartheta}) = 1$. Mortgage pricing is therfore abastract from the feedback from heterogeneous refinancing behavior to mortgage rates.[12]

4.4 Refinancing Frictions and the Attention Choice

Refinancing in the model runs through two margins. A household must first notice that its coupon sits above the market rate, and must then carry the transaction through. I call the first *attention* and the second *execution*, following Andersen et al. (2020), who separate time-dependent from state-dependent inaction. I model attention through the household's monitoring choice and execution through a hassle cost that is common across households.

Attention. Refinancing requires households to keep track of their current mortgage terms and prevailing mortgage rates. Doing so requires attention, so households choose how closely to monitor by trading off the value of refinancing opportunities against the cost of monitoring. A household that monitors at intensity χ receives opportunities to reconsider its mortgage as a Poisson process with arrival rate χ, and pays a convex flow cost

$$q(\chi; r, x) = V_W(r, x)\, \theta_s \, \frac{\chi^{1+1/\varepsilon}}{1 + 1/\varepsilon}. \tag{9}$$

I treat the time and effort spent on monitoring like money spent today, rather than as a psychological cost common to all households. In (9), θ_s is the dollar price of attention, and the factor $V_W(r, x)$ values this dollar cost at the household's marginal utility of wealth, as it would any other expense. Time spent tracking rates and loan terms could be spent working or economizing on other expenses, so its opportunity cost is measured in consumption, and the same dollar cost weighs more on a household short of cash. A utility cost common to all households would ignore this difference

[12] Endogenous refinancing behavior can feed back into mortgage pricing. As emphasized by Berger et al. (2024), households with higher refinancing propensities are overrepresented among those who prepay, so the prepayment hazard relevant for pricing can differ from its ex-ante population distribution. I abstract from this selection-pricing feedback and hold the benchmark hazard fixed. The schedule therefore does not respond to the level of attention either, and the policy counterfactuals in Section 5 hold $m(r)$ fixed.

and imply that the household with the highest marginal utility of wealth, the one shortest of cash, monitors most for the same dollar gain. That prediction runs against the evidence that lower-income and lower-wealth households respond less to refinancing opportunities (Andersen et al., 2020). Measuring gains and costs in the same units also means that $V_W(r, x)$ cancels, so attention responds to the dollar value of a refinancing opportunity, the amount the household would pay for it (Proposition 1). Under this assumption, liquid wealth affects attention only through the value of cash today relative to savings later, a channel I develop in (27).

The price of attention may differ across households, and factors that affect it, such as financial literacy and the opportunity cost of time, may vary systematically with income. I therefore allow the price of attention to vary with the household's income level,

$$\theta_s = \theta\left(\frac{Y_s}{\bar{Y}}\right)^{\omega}, \tag{10}$$

where $\bar{Y}$ is a reference income.

Execution. Even after noticing a refinancing opportunity, households may hesitate to act because completing a refinance requires time and effort. They must assemble income and asset documentation, arrange an appraisal, compare offers across lenders, and bear the risk that an application is declined. I capture these nonpecuniary barriers with a hassle cost of Φ dollars of consumption, common across households and paid once at the moment of refinancing. Its utility cost is therefore $\Phi V_W(r, x)$.

On a consideration opportunity, the household chooses cash-out $\Delta F^{co} \geq 0$ and pays the monetary closing cost κ. The LTV and PTI limits restrict the choice to a feasible set $\mathcal{D}(r, x)$, and the value of the best available refinance is

$$\widetilde{V}^{refi}(r, x) = \max_{\Delta F^{co} \in \mathcal{D}(r,x)} V\left(r, (W + \Delta F^{co} - \kappa,\ F + \Delta F^{co},\ m(r),\ s)\right). \tag{11}$$

The household refinances only when the dollar gain from refinancing covers the hassle cost:

$$\Delta^{\$}(r,x) = \frac{\widetilde{V}^{refi}(r,x) - V(r,x)}{V_W(r,x)} \;\geq\; \Phi, \qquad G^{\$}(r,x) = D(r,x)\left[\Delta^{\$}(r,x) - \Phi\right]_+, \tag{12}$$

where $D(r,x) = \mathbf{1}\{\mathcal{D}(r,x) \neq \varnothing,\ \Delta^{\$}(r,x) \geq \Phi\}$ and $G^{\$}$ is the surplus an opportunity delivers, net of the hassle cost.

Attention enters the household's problem additively, so the choice of χ separates from consumption and cash-out into the static subproblem

$$\max_{\chi \in [0,\bar{\chi}]} \left\{ \chi\, V_W(r,x)\, G^{\$}(r,x) - q(\chi; r, x) \right\}. \tag{13}$$

Proposition 1 (Closed-form attention rule)**.** *The optimal monitoring intensity is*

$$\chi^*(r,x) = \left(G^{\$}(r,x)/\theta_s\right)^{\varepsilon}, \tag{14}$$

and the attention term in the household's problem equals the smooth convex function $\Psi_s(G^{\$}) = (G^{\$})^{1+\varepsilon} / \left[(1+\varepsilon)\theta_s^{\varepsilon}\right]$*, with* $\Psi_s' = \chi^*$*. In the interior,* $\partial \ln \chi^* / \partial \ln G^{\$} = \varepsilon$ *and* $\partial \ln \chi^* / \partial \ln \theta_s = -\varepsilon$*.*

Proof. Substituting (9) into (13) and dividing by $V_W(r,x) > 0$ leaves $\max_\chi \{\chi G^{\$} - \theta_s \chi^{1+1/\varepsilon}/(1+1/\varepsilon)\}$. The first-order condition $\theta_s \chi^{1/\varepsilon} = G^{\$}$ gives $\chi^* = (G^{\$}/\theta_s)^{\varepsilon}$, unique by strict concavity; the envelope theorem gives $\Psi_s' = \chi^*$, and the elasticities follow by logarithmic differentiation.[13] □

Model hazard and limiting cases. The monthly refinancing hazard implied by the model is

$$h(r,x) = 1 - \exp\left[-\left(\nu_m + \chi^*(r,x)\, D(r,x)\right)/12\right], \tag{15}$$

so both its shape in the rate gap and its level across income quintiles are outputs of the attention policy. The two frictions carry four parameters in all, $(\theta, \varepsilon, \Phi, \omega)$. Section 4.7 derives how the hazard varies with the rate gap and income, and Section 5 estimates the parameters against the hazard surface.

[13] I cap the intensity at $\bar{\chi} = 12$ per year, at most monthly reconsideration. The cap binds on a negligible mass of households at the estimated parameters and is imposed as an economic ceiling and for numerical stability.

The elasticity ε also indexes a family of specifications. As $\varepsilon \to 0$ the monitoring intensity $\chi^* = (G^{\$}/\theta_s)^{\varepsilon}$ approaches one for every household with a positive surplus, so attention stops responding both to the gain and to income, and the hazard becomes constant among in-the-money borrowers. That is the structure of the Calvo benchmark in Berger et al. (2021), where refinancing opportunities arrive at a fixed rate whenever the borrower is in the money. Appendix H develops the argument and shows that the menu-cost specification is nested as well.

4.5 Household Problem

A household chooses a consumption path and a monitoring intensity at every instant, and, when a consideration opportunity arrives, whether to refinance and how much equity to extract. Preferences are constant relative risk aversion with risk-aversion parameter γ, and future utility is discounted at rate δ. The monitoring cost enters the flow payoff, so attention is paid for continuously while refinancing itself happens at isolated dates.

$$V(x_t, r_t) = \sup_{\{C_t, \chi_t, D_t, \Delta F_t^{co}\}} \mathbb{E}_t \left[\int_0^\infty e^{-\delta t} \Big(\frac{C_t^{1-\gamma}}{1-\gamma} - q(\chi_t; r_t, x_t) \Big) dt \right],$$

$$\text{s.t.} \quad dW_t = \big(Y_{s_t} - C_t + r_t W_t \underbrace{- (m_t^* + \alpha) F_t}_{\text{mortgage payment}} \big) dt + \underbrace{dN_t^A D_t}_{\text{refi decision}} \Big(\underbrace{\Delta F_t^{co}}_{\text{cash-out}} - \underbrace{\kappa}_{\text{closing cost}} \Big),$$

$$dF_t = \underbrace{- \alpha F_t \, dt}_{\text{amortization}} + \underbrace{(F_t' - F_t) dN_t^A D_t}_{\text{refi jump}},$$

$$dm_t^* = \underbrace{(m(r_t) - m_t^*) \big(dN_t^m + dN_t^A D_t \big)}_{\text{coupon reset at refinancing (moving + refi)}}.$$

The value function is the household's maximal expected lifetime utility given its current state (r_t, x_t). $D_t \in \{0, 1\}$ records whether an arriving opportunity is executed.

The arrival of consideration opportunities is the one place where this problem departs from a standard refinancing model. The counting process N_t^A has intensity χ_t, the household's own monitoring choice, and the household pays $q(\chi_t; r_t, x_t)$ in (9) to sustain it. The hassle cost Φ is paid

in consumption-equivalent units at the moment of refinancing and enters the value function. An independent Poisson process N_t^m with intensity ν_m captures relocation and other forced coupon resets.

4.6 Dynamic Programming and Distributional Dynamics

Hamilton–Jacobi–Bellman equation. Let $V(r, x)$ denote the household's value function, where $x = (W, F, m^*, s)$. The household problem has the HJB representation

$$\begin{aligned}\delta V(r,x) = \sup_{\substack{C\geq 0\\ \chi\in[0,\bar{\chi}]}} \Big\{ & u(C) - q(\chi; r, x) + \mathcal{L}_r V(r,x) + \mathcal{L}_s V(r,x) \\ & + b_W(r,x;C) V_W(r,x) + b_F(x) V_F(r,x) \\ & + \chi\left[\widetilde{V}^{refi}(r,x) - \Phi V_W(r,x) - V(r,x)\right]_+ \\ & + \nu_m\left[V^m(r,x) - V(r,x)\right]\Big\},\end{aligned} \tag{16}$$

where the liquid-wealth and mortgage-balance drifts are

$$b_W(r,x;C) = Y_s - C + rW - m^*F - \alpha F, \qquad b_F(x) = -\alpha F, \tag{17}$$

and

$$V^m(r,x) = V(r,(W,F,m(r),s)) \tag{18}$$

is the value following an exogenous coupon reset.

The refinancing term highlights the endogenous-attention mechanism. Consideration opportunities arrive at the household's chosen intensity χ, while the hassle cost Φ is incurred only when refinancing is undertaken. Dividing the refinancing surplus in (16) by V_W yields the dollar surplus $G^{\$}(r,x)$ in (12), so the optimal monitoring intensity is given by (14). Because this intensity depends on the value function, the attention policy and the household problem are solved jointly.

Distribution dynamics. Given the household policies, let $\mu_t(r,x)$ denote the cross-sectional distribution of households over aggregate and individual states. The intensity at which households

actually refinance is

$$\lambda^{refi}(r,x) = \chi^*(r,x)D(r,x), \tag{19}$$

where $D(r,x)$ is the execution decision defined in (12). The distribution evolves according to the Kolmogorov forward equation

$$\begin{aligned}\partial_t \mu_t = & -\partial_W\big(b_W\mu_t\big) - \partial_F\big(b_F\mu_t\big) + \mathcal{L}_r^*\mu_t + \mathcal{L}_s^*\mu_t \\ & - \big(\nu_m + \lambda^{refi}\big)\mu_t + \nu_m\mathcal{J}_m^*\mu_t + \mathcal{J}_{refi}^*\Big(\lambda^{refi}\mu_t\Big).\end{aligned} \tag{20}$$

Here $\mathcal{L}_r^*$ and $\mathcal{L}_s^*$ are the adjoints of the interest-rate and income generators. The operators $\mathcal{J}_m^*$ and $\mathcal{J}_{refi}^*$ map mass from pre-jump states into the states reached after an exogenous coupon reset and refinancing, respectively. The refinancing mapping incorporates the household's optimal cash-out choice.

Equation (20) links endogenous attention to aggregate dynamics: changes in refinancing gains affect χ^*, which changes the rate at which households move across mortgage states and, through their consumption and saving policies, the evolution of the wealth distribution. I use these distributional dynamics to compute the monetary-policy and refinancing-policy counterfactuals below. Details of the discretization and numerical solution are provided in Appendix F.

4.7 Anatomy of the Refinancing Hazard

This subsection shows how the model generates the shape of the refinancing hazard in the rate gap $g \equiv m^* - m(r)$ and its level across income. Both follow from the attention rule of Proposition 1, $\chi^* = (G^\$/\theta_s)^\varepsilon$, once the surplus $G^\$$ is known. A benchmark gives $G^\$$ in closed form. It holds the rate gap fixed between refinancing events and values the mortgage liability separately from wealth. Appendix G sets out the benchmark and proves the results below.

The surplus is positive only above a breakeven gap. The effective fixed cost of refinancing is $\kappa_{eff} \equiv \kappa + \Phi$, and the gap at which the present value of the interest saving just covers this cost is

$$\underline{g}(F) = \frac{\tilde{\rho}\,\kappa_{eff}}{F}, \qquad \tilde{\rho} \equiv \delta + \nu_m + \alpha, \tag{21}$$

where $\tilde{\rho}$ discounts the saving for time preference, relocation, and amortization.

Proposition 2 (Refinancing hazard in closed form)**.** *Let $\varepsilon = 1$. The optimal monitoring intensity is*

$$\chi^*(g, F) = \begin{cases} 0, & g < \underline{g}(F), \\ \sqrt{\tilde{\rho}^2 + \dfrac{2F}{\theta_s}\big(g - \underline{g}(F)\big)} - \tilde{\rho}, & g \geq \underline{g}(F). \end{cases} \tag{22}$$

Attention is zero below $\underline{g}$, rises from zero at $\underline{g}$ with slope $F/(\theta_s\tilde{\rho})$, and is concave in the gap above it. For general ε,

$$\chi^* \sim \left[\frac{(1+\varepsilon)\,F\big(g - \underline{g}(F)\big)}{\theta_s}\right]^{\varepsilon/(1+\varepsilon)} \tag{23}$$

far in the money, where the numerator is the raw gain from refinancing.

For a single loan, Proposition 2 gives an inaction region followed by a concave rise, with no inflection point. Balances differ within an income group, and because $\underline{g} \propto 1/F$, larger loans turn on at smaller gaps. Averaging across loans adds a convex segment where loans are still crossing their breakevens, which gives the group hazard the S-shape in Figure 2 (Appendix G.2). Andersen et al. (2020) obtain the same profile by assuming that the refinancing probability is a logistic function of the incentive, with location, slope, and height as free parameters. Here κ_{eff}, θ_s, ε, and the balance distribution pin down all three.

Attention across income. The central fact in Section 3 is the income gradient in refinancing. With endogenous attention, income moves attention through three channels in (14), two through the surplus $G^{\$}$ and one through the price θ_s.

Write the raw gain as $X \equiv F\big(g - \underline{g}(F)\big) = gF - \tilde{\rho}\,\kappa_{eff}$. Near the breakeven $G^{\$} \approx X/\tilde{\rho}$, and far in the money (23) applies, so

$$\ln \chi^* \approx a \ln X - a \ln \theta_s + \text{const}, \qquad a = \varepsilon \text{ near the breakeven}, \quad a = \frac{\varepsilon}{1+\varepsilon} \text{ far in the money.} \tag{24}$$

The exponent is smaller far in the money because attention is already high there, so the next opportunity arrives soon and refinancing now adds less over waiting for it. Differentiating (24) across income states with $\theta_s = \theta(Y_s/\bar{Y})^{\omega}$ gives

$$\frac{d \ln \chi^*}{d \ln Y} = a \Big[\underbrace{\Big(1 + \frac{\tilde{\rho}\kappa_{eff}}{X}\Big) \frac{d \ln F}{d \ln Y}}_{\text{balance}} - \underbrace{\omega}_{\text{cost}} \Big]. \tag{25}$$

The first channel is the balance term, and it raises attention among higher-income households. A refinancing lowers the payment on every dollar of F, so the present value of the rate saving is proportional to F, while the fixed costs κ and Φ are not. A larger loan therefore has a lower breakeven $\underline{g}(F)$ and delivers a larger surplus at every gap above it. Because the fixed costs are netted out of the gain, a one percent larger balance raises the raw gain by $1 + \tilde{\rho}\kappa_{eff}/X$ percent. This factor is largest near the breakeven, where the fixed costs absorb a large share of the saving, and there the elasticity of attention to F exceeds ε. Higher-income households carry larger balances, both in the loan panel and in the model, where higher-income households more often hold the larger collateral types, so the term is positive.

The second channel is the cost term, and when the price of attention falls with income it also raises attention among higher-income households. Income enters θ_s through (10), and the parameter ω governs how the price varies with income and is estimated. At a given surplus the income elasticity of attention is $-\varepsilon\omega$, as in Proposition 1, and at a given raw gain it is $-a\omega$. With $\omega < 0$, higher-income households monitor more intensively at a given gain. With $\omega = 0$, the channel is shut down and all households face the same price. Section 5.4 tests whether the data support an income-dependent monitoring cost.

The third channel works through liquid wealth W, and its sign depends on whether the household gains more from cash-out or from payment relief. The marginal value of wealth cancels from (14), so W moves attention through the surplus $G^{\$} = D\,[\Delta^{\$} - \Phi]_+$ by changing how the household values cash-out and payment relief. To first order,

$$\Delta^{\$} \approx (1 - p_F)\,\Delta F^{co} + p_m\, g - \kappa, \qquad p_F \equiv -\frac{V_F}{V_W}, \quad p_m \equiv -\frac{V_{m^*}}{V_W}, \tag{26}$$

where p_F is the liquid wealth the household would give up to cut its mortgage by one dollar, p_m is the dollar value of a one-point lower coupon, and cash-out ΔF^{co} is bounded by the capacity $\bar{b}$ permitted by the collateral and PTI limits.

A household with low liquid wealth values future dollars less, which raises the value of cash-out and lowers the value of payment relief. Whether scarce liquidity raises or lowers the gain from refinancing therefore depends on whether the cash the household can draw, $\bar{b}$, exceeds the present value of the saving, $gF/\tilde{\rho}$. Formally, such a household has $p_F < 1$, so cash-out raises $\Delta^{\$}$, and a lower p_m, so payment relief is worth less. Both prices value future mortgage dollars, so they fall approximately in proportion from their market values, $p_F = 1$ and $p_m = F/\tilde{\rho}$. This gives $p_m \approx p_F F/\tilde{\rho}$, with p_F rising toward one as W grows. Substituting into (26) with $\Delta F^{co} = \bar{b}$ for a household at its borrowing limit gives

$$\Delta^{\$} \approx \bar{b} - \kappa + p_F\Big(\frac{gF}{\tilde{\rho}} - \bar{b}\Big), \tag{27}$$

so where $G^{\$} > 0$, $d\chi^*/dW$ has the sign of $gF/\tilde{\rho} - \bar{b}$.

Cash-out dominates when $\bar{b}$ is large or when the balance or the rate gap is small, and scarce liquidity then raises attention. When payment relief exceeds $\bar{b}$, the loss in relief outweighs the gain from cash-out, and scarce liquidity lowers attention. Lower-income households hold less liquid wealth in the model, so the first case flattens the income gradient and the second case steepens it.

Predictions of existing models. These channels also locate existing refinancing models. The Calvo benchmark of Berger et al. (2021), the $\varepsilon \to 0$ limit discussed in Section 4.4, shuts all three: consideration arrives at a constant rate, so the in-the-money hazard is flat across both income and rate gaps. Its hybrid variant adds a menu-cost trigger, which restores a threshold in the rate gap but leaves the hazard flat across income at a common balance, because the arrival rate still does not respond to the gain. Models with permanent borrower types (Zhang 2022; Berger et al. 2024) assign each household a fixed propensity to refinance, so differences across borrowers are given rather than derived and do not move with the gap, the balance, or a policy. The model of Eichenbaum et al. (2022) works mainly through the liquidity channel: refinancing is driven by equity extraction

for consumption smoothing, so low-income households with little liquid wealth refinance more and the income gradient takes the opposite sign. In the present model the sign and size of the gradient are not imposed. They depend on how the balance, cost, and liquidity channels combine, which Section 5 estimates.

5 Quantitative Analysis

5.1 Calibration

Refinancing-friction parameters. Table 4 reports the calibrated and estimated parameter values. Four parameters govern refinancing frictions: the price of attention θ, the elasticity of attention to the gain ε, the hassle cost Φ, and the income elasticity of the price of attention ω. I estimate them by simulated method of moments (SMM), targeting monthly refinancing hazards by rate-gap bin and income quintile. The empirical hazards are estimated with controls for the loan balance and the estimated LTV ratio. I therefore compute the model hazard for each income group at a common distribution of balances and collateral across income groups, holding each group's own distribution of liquid wealth, and match this conditional hazard to the data. The estimator minimizes

$$L(\theta, \varepsilon, \Phi, \omega) = \sum_{s} \sum_{b} w_{sb} \left[h^{c}_{sb}(\theta, \varepsilon, \Phi, \omega) - \hat{h}_{sb} \right]^2, \tag{28}$$

where h^{c}_{sb} is the conditional model hazard and $\hat{h}_{sb}$ the empirical hazard in income quintile s and rate-gap bin b, and w_{sb} is the share of loan-month observations in bin b within quintile s.

Different features of the hazard surface identify different parameters, as shown in Appendix G. The breakeven gap identifies the effective fixed cost of refinancing, κ_{eff}, estimated at \$3,901. Because mortgage balances vary within an income quintile, the location of the hazard reflects the average breakeven $\mathbb{E}_s[\underline{g}(F)]$, while its slope also reflects the dispersion of breakevens. Both are implied by the balance distribution, which the estimation holds common across income groups. Far in the money, the level of the hazard depends on θ and ε jointly, and its curvature across

rate gaps within an income quintile identifies the attention elasticity, $\hat{\varepsilon} = 0.326$. At a common balance and collateral composition, relative hazard levels across income quintiles identify the income elasticity of the price of attention, $\hat{\omega} = -1.084$.

Implied attention and external benchmarks. The model-generated attention intensities are consistent with the literature. Among in-the-money households, the model-implied average χ^* is 0.252 per year, within the 0.228–0.270 range reported by Berger et al. (2021).[14]

The model's implication that attention rises with income is consistent with evidence on refinancing behavior and mortgage search. Andersen et al. (2020) specify time-dependent and state-dependent inaction as exogenous functions of demographics, and estimate that the probability of being awake to refinancing opportunities rises with income, education, and wealth. In my model, the average attention intensity among in-the-money households rises monotonically from 0.130 per year in the bottom income quintile to 0.404 in the top. Evidence on search behavior at origination in the United States points the same way. A mortgage-search index constructed from the National Survey of Mortgage Originations (NSMO) rises monotonically across income quintiles (Figure J.1 (b)).[15]

Monitoring costs across income. The parameter ω is the income elasticity of the price of attention $\theta_s = \theta\,(Y_s/\bar{Y})^{\omega}$ in (10). The estimate $\hat{\omega} = -1.084$ implies that higher-income households face a lower cost of sustaining a given level of attention. An elasticity close to -1 means that the price of attention falls roughly in proportion to income, so doubling income halves the cost of a unit of attention.

I interpret ω as the net of two opposing forces. The opportunity cost of time rises with income together with wages, implying $\omega > 0$. Consistent with this, the psychological cost of executing a refinancing in Andersen et al. (2020) increases with income, which the authors interpret as the

[14] For each income state, χ^* is averaged over in-the-money household states weighted by the stationary distribution. In the hybrid Calvo-menu-cost specification of Berger et al. (2021), the total arrival rate combines a costless component of 0.125 and a costly component of 0.145. They are exercised only when the value of refinancing exceeds the cost, the effective in-the-money hazard in the hybrid model lies between these two values.

[15] The NSMO index measures search at origination, not monitoring of refinancing opportunities. Using it as evidence for refinancing therefore requires the assumption that differences in search behavior at origination persist into refinancing decisions.

value of time. Financial literacy, experience, and information networks also rise with income, reducing the time and cognitive burden required to sustain a given level of attention and implying $\omega < 0$. Survey evidence points to a steep income gradient in this second force. In the 2022 SCF, the share of mortgagor households answering all three financial-literacy questions correctly more than doubles from the bottom to the top income quintile (Figure J.1 (a)), and low-literacy households are less likely to refinance (Cota and Šterc, 2024). The estimate $\hat{\omega} < 0$ indicates that the second force dominates on the monitoring margin.

Other model parameters. The remaining parameters follow the quantitative macro-finance literature. The coefficient of relative risk aversion is set to $\gamma = 2$, a standard value in macroeconomic models. The time-preference parameter δ and the parameters governing the interest-rate process, $(\eta_r, \sigma_r, \bar{r})$, follow Berger et al. (2021), implying an annual mean short rate of 3.5 percent and an interest-rate volatility of 6 percent. Mortgage balances are not set to a single value. The simulated distribution is initialized from the joint distribution of balances, rate gaps, and income in the loan panel. The stochastic income process follows Flodén and Lindé (2001), with a log-income persistence corresponding to a half-life of 7.3 years and an annualized volatility of 21 percent.

Two underwriting limits cap how much a household may borrow. The LTV ratio $\phi^{LTV} = 0.8$ caps the balance it may carry against its home, and the PTI ratio $\phi^{PTI} = 0.43$ caps the mortgage payment it may commit relative to income. Following Greenwald (2018), I apply the borrowing limits to transactions that increase mortgage principal. A pure rate refinance leaves the balance unchanged and lowers the required payment, whereas a cash-out refinance increases principal and is therefore subject to both constraints. Because house prices are fixed in my model, the LTV constraint limits the amount of equity that can be extracted and leaves eligibility for a pure rate refinance unaffected.

Further details on the cross-sectional distribution of the baseline model are provided in Figure E.1.

Table 4: Model Parameter Values

Panel A: Exogenous Processes and Preferences			
Parameter	**Value**	**Description**	**Source**
$\ln 2/\eta_r$	5.3 years	Half-life of interest-rate shock	Berger et al. (2021)
$\bar{r}$	3.5% (p.a.)	Unconditional mean short rate	Berger et al. (2021)
σ_r	6% (p.a.)	Volatility of short rate	Berger et al. (2021)
$\ln 2/\eta_Y$	7.3 years	Half-life of (log) income shock	Flodén and Lindé (2001)
$\mathbb{E}[Y_t]$	\$58,000	Unconditional mean income	Flodén and Lindé (2001)
σ_Y	21% (p.a.)	Log-income volatility	Flodén and Lindé (2001)
γ	2	Relative risk aversion	Macro literature standard
δ	14% (p.a.)	Rate of time preference	Berger et al. (2021)
α	3% (p.a.)	Amortization rate of principal	Berger et al. (2021)
ν_m	4.1% (p.a.)	Arrival rate of exogenous coupon resets	Berger et al. (2021)

Panel B: Refinancing Frictions (estimated jointly by SMM)			
Parameter	**Estimate**	**Description**	**Identified by**
$\hat{\theta}$	\$190,887	Price of attention	Height of the hazard far in the money
$\hat{\varepsilon}$	0.326 (0.065)	Elasticity of attention to the dollar gain	Curvature above the breakeven
$\hat{\Phi}$	\$901	Hassle cost	Breakeven gap
$\hat{\omega}$	−1.084 (0.218)	Income elasticity of the price of attention	Income gradient of the hazard
$\kappa_{eff} = \kappa + \hat{\Phi}$	\$3,901	Effective fixed cost of refinancing	Derived

Panel C: Contract and Institutional Parameters (set externally)			
Parameter	**Value**	**Description**	**Source**
κ	\$3,000	Monetary closing cost of refinancing	Eichenbaum et al. (2022)
ϕ^{LTV}	0.80	Maximum LTV ratio	See text
ϕ^{PTI}	0.43	Maximum PTI ratio	See text
$\bar{\chi}$	12 (p.a.)	Upper bound on monitoring intensity	Monthly time grid
$\bar{F}_k = \phi^{LTV} H_k$	{\$104,270, \$203,790, \$370,150}	Maximum balance by collateral type	GSE loan panel
Ω_k	{0.267, 0.492, 0.241}	Population share by collateral type	GSE loan panel

Notes: Estimates are from the baseline model. Panel B reports the four refinancing-friction parameters estimated jointly by simulated method of moments. The targets are the fitted monthly refinancing hazards by rate-gap bin and income quintile from the full-control specification of (3). The model hazards are computed conditional on a common balance and collateral composition across income quintiles. Standard errors are in parentheses. They are omitted for θ, because far in the money a change in θ can be offset by a change in ε with little change in the hazard, and for Φ, which the breakeven identifies only jointly with κ. The PTI constraint is omitted for rate refinancing.

5.2 Targeted Hazard Moments

I first evaluate the model against the moments used to estimate the refinancing-friction parameters. The targeted moments are monthly refinancing hazards by interest-rate-gap bins and income quintile. These moments are central to the quantitative exercise because they jointly identify the two dimensions of heterogeneity emphasized in the model. Variation across rate-gap bins pins down the breakeven gap and the curvature of the hazard, and so the hassle cost Φ and the elasticity

ε. Variation across income quintiles identifies the income elasticity of the price of attention, ω, given the gains that the model already generates.

The empirical counterpart is the refinancing hazard by rate-gap bin and income quintile estimated from the loan-level panel, using the simple rate gap of Figure 2(a). The model counterpart is constructed by initializing the simulated mortgage distribution with the income-specific distribution of outstanding coupons observed in the loan panel in January 2005.[16] Holding this initial distribution fixed, I feed in the realized path of market mortgage rates and compute monthly refinancing hazards within the same rate-gap bins and income groups used in the empirical analysis.

Figure 5 compares the model-implied hazards with their empirical counterparts. The baseline model reproduces the main patterns in the data. Refinancing hazards rise with the rate gap, but they remain far below one even for borrowers who are deep in the money. This feature is informative about the attention friction. Profitable refinancing opportunities are not acted upon mechanically, even when the monetary gains are large. The model also captures the income gradient in refinancing hazards, with substantially higher refinancing propensities among borrowers in the fourth and fifth income quintiles than among those in the lower part of the distribution.

The targeted moments capture both margins that matter for monetary transmission, responsiveness to rate gaps and its heterogeneity across income, and provide the basis for the counterfactual exercises below.

5.3 Untargeted Model Validation

Average refinancing hazards. As a check on the aggregate income gradient, I next examine the average monthly refinancing probability among *in-the-money* borrowers by income quintile. This moment is not directly targeted in the calibration, which relies primarily on refinancing hazards across rate-gap bins. It is nevertheless informative because it summarizes whether the calibrated model reproduces the aggregate income gradient in refinancing behavior that motivates the paper.

Figure 6 Panel (a) shows that refinancing propensity in the data increases monotonically with income, producing a clear upward-sloping income gradient across quintiles, and that the baseline

[16] The empirical sample begins in 2000, but the model simulation starts from the mid-sample to allow for a sufficient burn-in period.

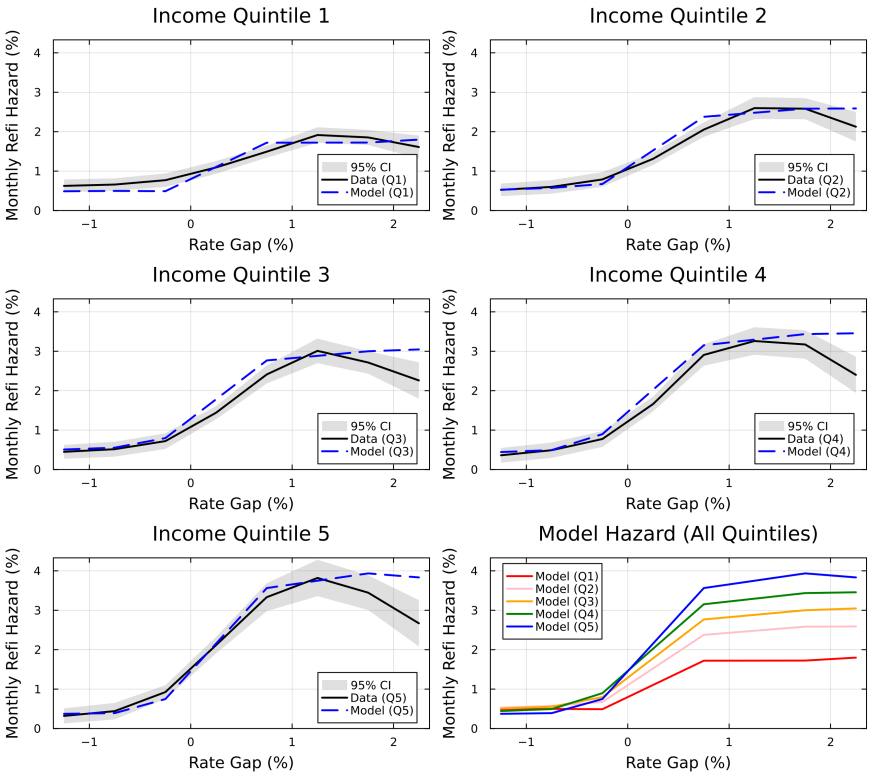


Figure 5: Hazard Rate by Gap Bin: Model vs. Data

Notes: The figure compares monthly refinancing hazards by income quintile and rate-gap bin in the data with those generated by the baseline model. These data hazards are targeted moments in the calibration of the income-specific refinancing-friction parameters. The model simulation is initialized with the January 2005 income-specific distribution of outstanding coupons and uses the realized path of mortgage rates. The data counterpart uses the Fannie Mae and Freddie Mac loan-level panel, 2000–2025, and the same rate-gap bins as Figure 2. Shaded regions are 95% confidence intervals around the empirical hazards.

model reproduces this pattern. Panel (b) compares the data with two benchmark refinancing models. The Calvo benchmark of Berger et al. (2021) is the $\varepsilon \to 0$ limit of the baseline model, in which consideration arrives at a constant rate that responds neither to the dollar gain nor to income. Appendix H derives the limit analytically. The benchmark generates an essentially flat profile, showing that matching the average level of refinancing leaves the cross-section without an income gradient. Its hybrid variant with a menu-cost trigger produces a similarly flat income profile.[17]

[17] Berger et al. (2021) emphasize that the pure-Calvo restriction is chosen for tractability. Because refinancing opportunities arrive independently of household states, the mortgage-pricing problem can be separated from households' consumption-saving decisions, allowing the equilibrium mortgage rate to be written as a function of the short rate, $m = m(r)$. They show that this parsimonious specification delivers time-series implications close to their richer hybrid model and use it to study the path dependence of monetary policy. My model instead allows refinancing hazards to

By contrast, the model of Eichenbaum et al. (2022) generates an income gradient of the opposite sign. Refinancing there is driven by the consumption-smoothing motive. Low-income households hold little liquid wealth and refinance to extract equity, while high-income households sit on buffer assets and have less reason to act. Under their baseline calibration, the annual refinancing probability among in-the-money mortgage holders falls from 5.5 percent in the bottom income quintile to 2.5 percent in the top. Because the benchmarks use different data and sample periods, only the cross-sectional shapes are comparable. Section 5.4 decomposes how much of the gradient survives when the income slope of the price of attention is switched off.

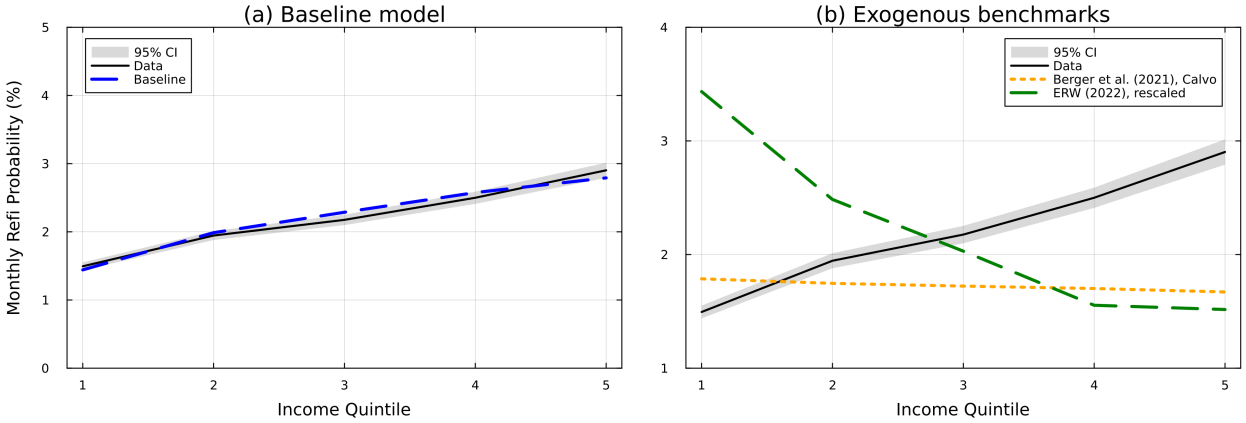


Figure 6: Average Refinancing Hazard by Income Quintile: Model Fit and Exogenous Benchmarks

Notes: Average monthly refinancing hazards among in-the-money borrowers by income quintile. Panel (a) compares the data with the baseline model. Panel (b) compares the data with the Calvo benchmark of Berger et al. (2021) and Eichenbaum et al. (2022). The Eichenbaum et al. (2022) series is the annual refinancing probability among in-the-money working-age mortgage holders, converted to a monthly rate and rescaled to match the sample mean, so benchmark comparisons concern only the cross-sectional shape. Shading denotes 95% confidence intervals.

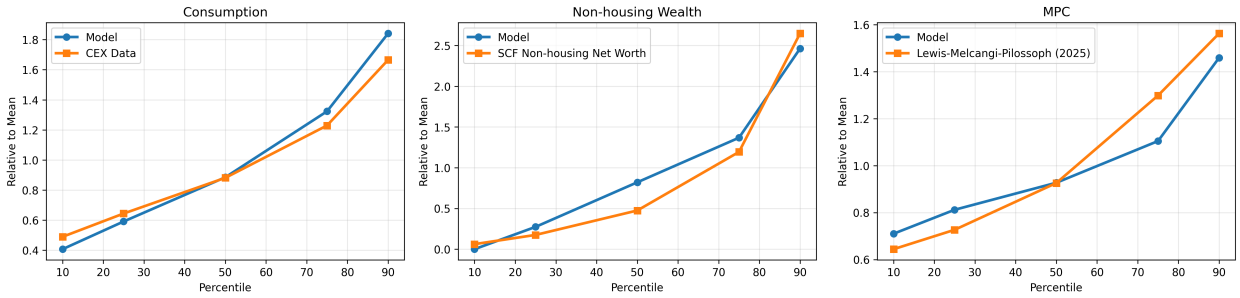
Distributional moments. I next examine whether the model delivers plausible cross-sectional heterogeneity in consumption, non-housing wealth, and MPCs. These moments are not targeted in the calibration, but they are important for interpreting the monetary-transmission results because refinancing affects aggregate consumption through household cash flows and MPC heterogeneity.

Figure 7 compares the Lorenz curves for consumption and non-housing wealth with their empirical counterparts from the Consumer Expenditure Survey (CEX; 1988–2018) and the 2001

vary endogenously across household states, but does not feed the resulting prepayment behavior back into mortgage pricing.

SCF, and contrasts the model-implied MPC profile with the estimates in Lewis et al. (2026). All objects are computed at percentiles of the cross-sectional distribution and normalized by their means. The model reproduces the broad distributional patterns in the data—realistic inequality in non-housing wealth and consumption—and generates an MPC profile that is close to the empirical benchmark.

This validation matters because the quantitative transmission results depend on where refinancing gains accrue in the liquidity distribution. The fact that the model generates plausible non-housing wealth inequality and MPC heterogeneity reduces the concern that the consumption impulse response functions (IRFs) and counterfactual results are driven by an unrealistic distribution of liquidity or MPCs.



(a) Consumption (b) Non-housing Wealth (c) MPC

Figure 7: Lorenz Curve and MPC: Model vs. Data

Notes: Panels (a) and (b) compare model-implied distributions of consumption and non-housing wealth with their empirical counterparts. Panel (c) compares the model-implied MPC profile with external estimates. The model's financial asset W is mapped to non-housing net worth in the SCF. All series are normalized by their means. These moments are not targeted in the calibration.

5.4 Model Mechanisms

A common monitoring technology. I next ask how much of the income gradient can be reproduced without income variation in monitoring costs. I impose a common monitoring cost, $\omega = 0$, and re-estimate all remaining parameters against the same moments. This restriction gives every income state the same price of attention, $\theta_s = \theta$. Households then differ in refinancing gains mainly through the liquidity channel in (27), because the targeted hazards hold the balance and collateral composition common across income groups, as in (28), which shuts the balance channel

in (25). Scarce liquid wealth works in both directions. It makes cash-out more valuable, but it also leaves many low-income households unable to pay the closing cost in cash or to roll it into the new loan within the PTI limit, so they cannot refinance even when the rate gap is large. When the second effect dominates, low-income households refinance less often even at a common price of attention.

The common-cost model reproduces the shape of the hazard in the rate gap. Figure 8 shows that the attention mechanism alone, without income differences in monitoring costs, generates this shape—the inaction region below the breakeven, the liftoff, and the concave rise—at every quintile.

However, the common-cost model fails to explain refinancing at the two ends of the income distribution, which indicates that monitoring costs differ across income levels. Figure 9 shows that it overpredicts refinancing in the bottom quintile and underpredicts it in the fourth and fifth quintiles, whereas the data rise monotonically across income. Allowing monitoring costs to vary across households, as in the baseline model, reproduces the gradient in the data. Quantitatively, even with the remaining parameters re-estimated, the log difference in in-the-money hazards between the top and bottom quintiles is 0.22 in the common-cost model, against 0.66 in the data, or about one third of the gradient. Differences in refinancing gains alone therefore cannot make low-income households as inattentive, or high-income households as attentive, as the data require.

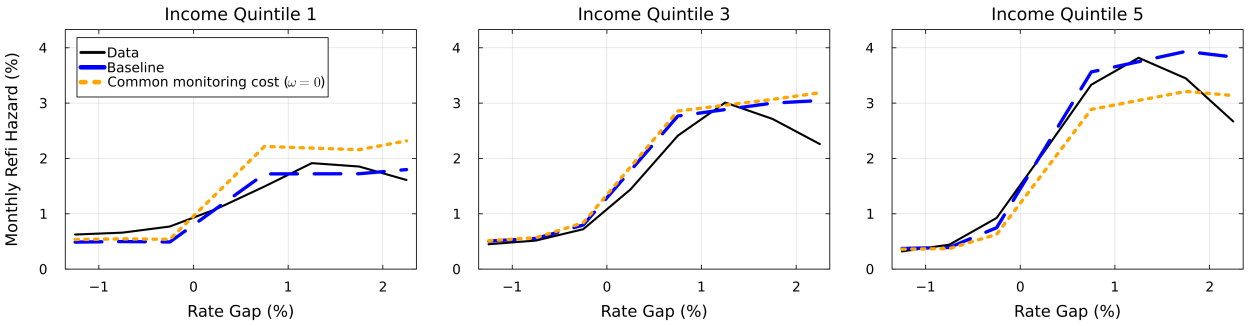


Figure 8: Hazard Rate by Gap Bin: Common Monitoring Technology

Notes: The figure compares the data and baseline-model monthly refinancing hazards by rate-gap bin, shown for the first, third, and fifth income quintiles, with a counterfactual that sets $\omega = 0$, so that the price of attention is common across income states, with the remaining parameters re-estimated against the same moments. Households still differ in the dollar gain from refinancing. The exercise is a model-based accounting comparison, conditional on the model structure.

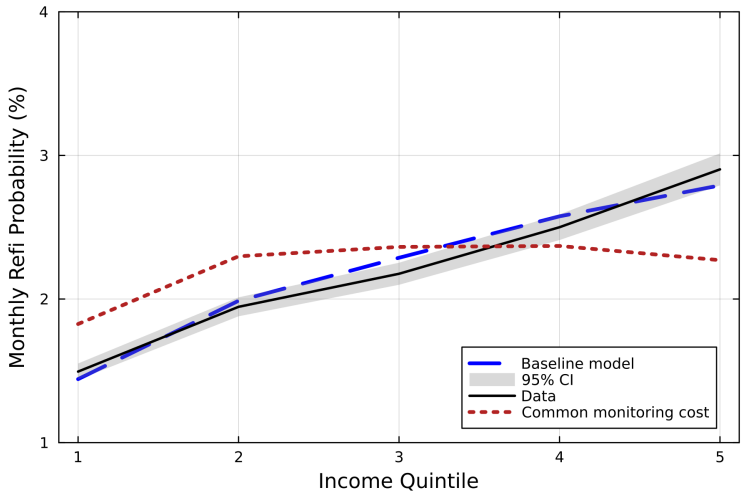
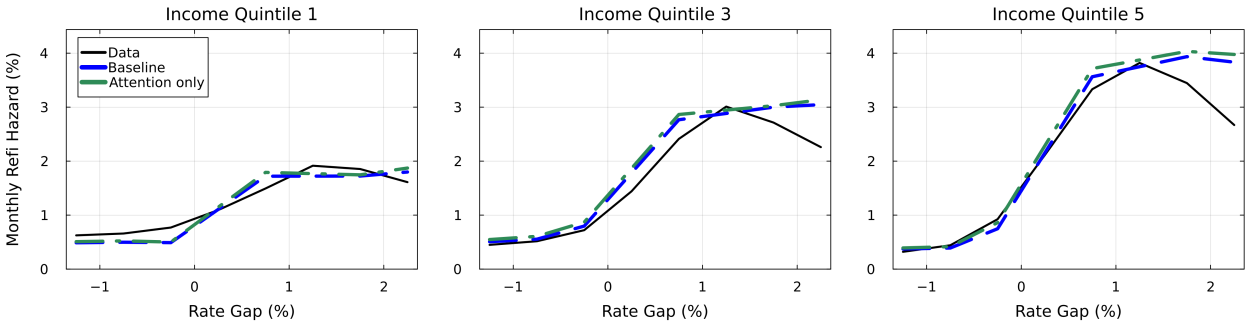


Figure 9: Average Hazard Rate: Common Monitoring Technology

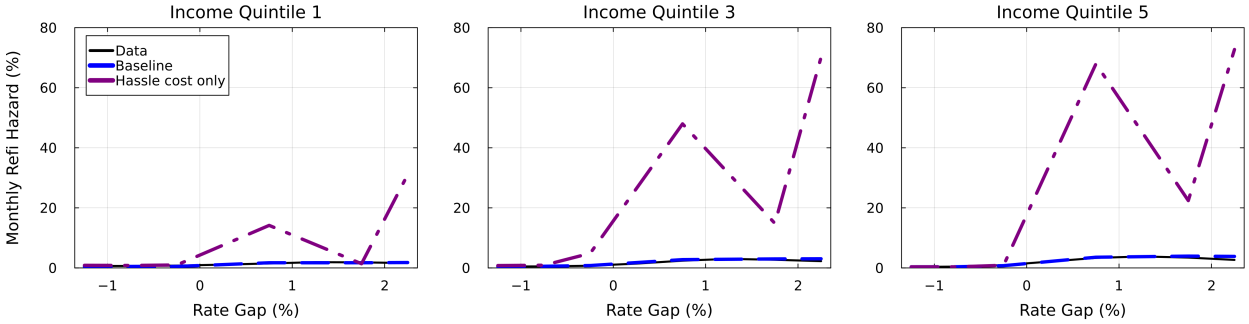
Notes: The figure compares average monthly refinancing hazards among in-the-money borrowers by income quintile in the data, the baseline model, and a counterfactual that sets $\omega = 0$ so that the price of attention is common across income states, with the remaining parameters re-estimated against the same moments. Hazards are averaged over in-the-money gap bins using the empirical distribution of loans across bins.

Friction channels. I next separate the roles of endogenous attention and the hassle cost by removing each friction in turn. Figure 10 (a) retains the endogenous attention choice but switches off the hassle cost. The resulting hazards remain close to the baseline over most rate-gap bins, indicating that endogenous attention accounts for much of the observed sluggishness in refinancing. The hassle cost plays a more localized role by generating the inaction region at small refinancing gains.

Figure 10 (b) performs the reverse experiment, retaining the hassle cost while holding attention at full consideration. Refinancing then rises far above the data once the gain covers the hassle cost. Without limited attention, the model therefore cannot reproduce the low refinancing hazards observed even among borrowers with substantial financial incentives to refinance. Limited attention is thus essential for matching the level of refinancing, while the hassle cost primarily governs inaction near the refinancing boundary. This division of roles is consistent with Berger et al. (2021), who show that transaction costs alone generate excessive refinancing and that infrequent consideration is needed to reproduce observed inertia. Appendix H shows how both the Calvo and hassle-cost-only specifications arise as boundary cases of the model's attention technology.



(a) Attention-Only Friction



(b) Hassle-Cost-Only Friction

Figure 10: Hazard Rate by Gap Bin: Isolating the Friction Channels

Notes: Both panels compare the data and baseline-model monthly refinancing hazards by rate-gap bin, shown for the first, third, and fifth income quintiles. Panel (a) retains only the endogenous attention channel and switches off the hassle cost. Panel (b) retains only the hassle cost and holds attention at full consideration; with attention saturated, refinancing responds almost mechanically to the rate gap and the implied hazard is far above the data, so the vertical scale differs accordingly.

5.5 Monetary Transmission

I now quantify how unequal refinancing responses shape the consumption effects of monetary policy. A decline in interest rates raises the value of refinancing for borrowers with high outstanding coupons. The frictions defined in Section 4 determine how many borrowers respond, how quickly they respond, and where the resulting cash-flow relief accrues in the income and liquidity distribution. The consumption impulse response therefore summarizes the interaction between the interest-rate stimulus, refinancing frictions, cash-out choices, and household MPC heterogeneity.[18]

IRF construction. I compute the response to a 150 basis point decline in the policy rate—one step of the model's rate grid—which lowers the market mortgage rate by 85 basis points on impact. Both rates then revert along the estimated rate process. The comparison is between two economies that start from the same cross-sectional distribution of liquid wealth, mortgage balances, coupon rates, and income.[19] In the baseline path, the economy evolves without the rate cut. In the shocked path, the policy rate is lowered at time zero and the distribution is simulated forward in the same model environment. Aggregate consumption is computed by integrating the household consumption policy over the evolving distribution. The consumption IRF is measured as the percentage deviation of aggregate consumption in the shocked economy from the baseline economy at each horizon. Details of the IRF construction are provided in Appendix I.

Baseline transmission. I first examine the baseline economy to establish the strength of the refinancing channel before turning to counterfactuals. Normalized to a 100 basis point policy cut, aggregate consumption rises 1.41 percent on impact and averages 0.89 percent above the no-shock path over five years. The response is front-loaded. The rate cut immediately shifts household policy functions and the return on liquid savings, while the subsequent dynamics are governed

[18] The consumption responses in this section are untargeted. They serve as an independent check on the estimated model, not as inputs to its calibration.

[19] The exercise is partial equilibrium. The policy-rate path is taken as given, and I do not model feedback from aggregate consumption to interest rates, house prices, or lender balance sheets.

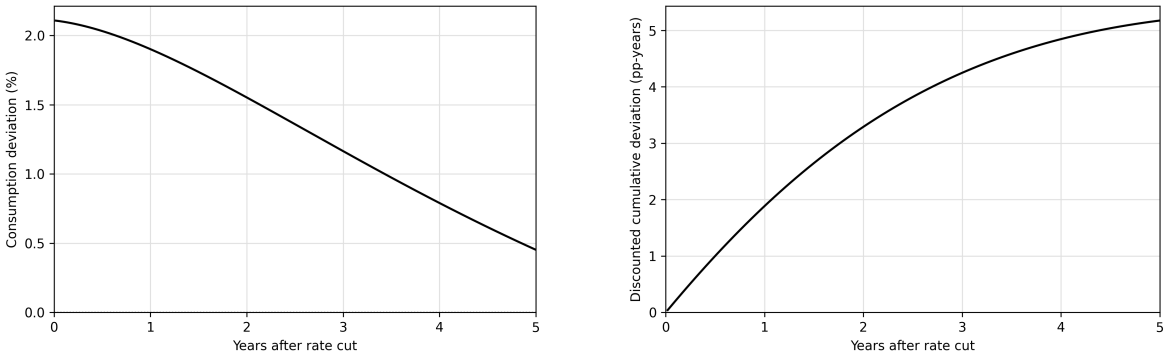
by the stochastic arrival of refinancing opportunities, cash-out decisions, and the evolution of the mortgage distribution.

The impact response falls between the two closest structural benchmarks, both of which report the effect of a short-rate cut on impact. For a 100 basis point decline, Berger et al. (2021) obtain 0.67 percent in their ergodic economy and 0.96 percent after a secular decline in mortgage rates, while Eichenbaum et al. (2022) obtain 2.7 percent for a cut of the same size from a flat rate history. The ordering follows from what each model lets households do. Berger et al. work through the refinancing cash-flow channel alone, whereas households in my model also extract equity when they refinance. Eichenbaum et al. add a life-cycle structure in which cash-out is more frequent, while aggregate income and house prices move with rates. The resulting house-price appreciation further expands home equity and equity extraction. The model here is an intermediate case. Households choose cash-out optimally when they refinance, but the environment is infinite-horizon and abstracts from both life-cycle dynamics and house-price movements.[20]

Reduced-form estimates line up with the model at their own horizons. Foulis et al. (2026) find that a 100 basis point decline in the market mortgage rate raises consumption by 3.0 percent over the window ending six months after deal expiry. The model's six-month average, normalized to the mortgage rate, is 2.44 percent. The difference is consistent with the mechanism they identify. Much of their consumption response operates through additional borrowing induced by rate-driven house-price appreciation, whereas the model here holds house prices fixed and captures equity extraction only against existing collateral values. Cloyne et al. (2020) estimate a four-year cumulative increase of \$535 for U.S. mortgagors after a temporary 25 basis point policy-rate cut, about \$2,140 when scaled fourfold, against \$2,292 in the model, the undiscounted four-year cumulative response to a 100 basis point cut. The two are close, but they are not the same object. Cloyne et al. attribute most of their response to a general-equilibrium income channel that is absent here, so the model reaches their total through the mortgage channel alone.[21]

[20]The five-year discounted cumulative response is $\int_0^5 e^{-\delta t}\, IRF_C(t)\, dt = 5.17$ percentage-point-years at $\delta = 0.14$, or \$2,927 per household; undiscounted, \$3,786. Scaled linearly to a 100 basis point policy cut, the first-year average is 1.35 percent and the five-year cumulative response is \$1,951 discounted and \$2,524 undiscounted.

[21]The two experiments differ in persistence as well as in size. The shock in Cloyne et al. (2020) is 25 basis points on impact and returns to zero after about four to eight quarters, whereas the short rate here is mean-reverting with a



(a) Consumption IRF (b) Cumulative Consumption IRF

Figure 11: Consumption Response to a 150 Basis Point Rate Cut

Notes: The figure reports the model-implied aggregate consumption response to a 150 basis point policy-rate cut—one step of the model's rate grid, which lowers the market mortgage rate by 85 basis points on impact—in the baseline economy. Panel (a) is the percentage deviation of consumption from an otherwise identical no-cut path. Panel (b) cumulates that deviation, discounted at $\delta = 0.14$, in percentage-point-years. The exercise holds the initial distribution and all model primitives fixed and is partial equilibrium, so it measures the refinancing and equity-extraction margins and does not deliver a full monetary-policy multiplier.

Homogeneous-attention benchmark. I next compare the baseline model with a benchmark in which consideration opportunities arrive at constant rates that do not vary across households or respond to their states. I choose homogeneous exogenous attention rates to match the baseline model's average refinancing hazards in the in- and out-of-the-money regions, while holding the remaining parameters, initial distribution, and execution decision rules fixed. This benchmark produces a slightly smaller aggregate consumption response to the rate cut: the five-year average consumption response to a 100 basis point policy cut is 0.84 percent, compared with 0.89 percent in the baseline model.

By contrast, the homogeneous-attention benchmark misstates which households the consumption gains accrue to. Because it holds the arrival of consideration opportunities fixed across households, it cannot generate the income gradient in refinancing that the baseline model is estimated to match, and it therefore assigns refinancing to households who in the data rarely capture it. The consequence is amplified by MPC heterogeneity: the low-income households to which the benchmark over-assigns refinancing are also those that spend the largest share of the relief they

half-life of 5.3 years. A more persistent cut keeps households in the money longer and so raises the four-year cumulative response, which works in the direction of the gap in the text.

receive, so the benchmark overstates their consumption response by more than it overstates their refinancing. Compared to the baseline model, the five-year response of the bottom income quintile is 16 percent larger under homogeneous attention, while the responses of the third through fifth quintiles are 7 to 12 percent smaller. The same pattern appears in refinancing: in the pre-shock distribution, the homogeneous-attention model raises the refinancing rate of the bottom income group and slightly lowers that of the top group. A model with homogeneous attention therefore modestly understates the aggregate consumption response while allocating the consumption gains from a rate cut differently across the income distribution.

5.6 Policy Experiment

Refinancing notifications. I next ask how much of the lost transmission can be recovered by making refinancing opportunities more salient. Byrne et al. (2023) run a field experiment in which eligible mortgage borrowers receive redesigned disclosures and reminder notices, raising the share of attentive borrowers from 24 to 39 percent. The intervention shifts attention without changing the gains from refinancing or the terms of the mortgage, providing an empirical benchmark for the notification counterfactual. Because the intervention is common across borrowers, any heterogeneity in its effects across the income distribution is generated by the model.

I model a notification as an exogenous stream of free consideration opportunities arriving at rate $\lambda_{free} \geq 0$, so a household reconsiders its mortgage at total rate $\chi^*(r,x) + \lambda_{free}$. The policy is costless to the household. The monitoring cost in (9) is paid on the private intensity χ^* alone, and a free arrival triggers a refinance only where the household would have executed anyway. All estimated parameters are held at their baseline values. I choose λ_{free} so that the model reproduces Byrne et al.'s relative increase in the attentive share, defining that share over the refinancing risk set $\{(r,x) : D(r,x) = 1\}$ as

$$\mathcal{A}(\lambda_{free}) = \frac{\sum_x \mu(x)\, D(r,x)\left[1 - e^{-(\chi^*(r,x)+\lambda_{free})T}\right]}{\sum_x \mu(x)\, D(r,x)}, \qquad T = \tfrac{4}{12}, \tag{29}$$

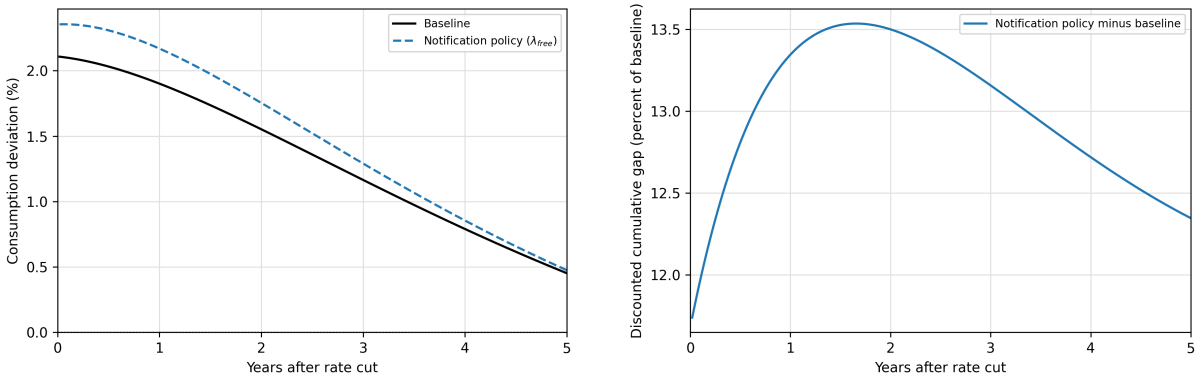
and requiring $\mathcal{A}(\lambda_{free})/\mathcal{A}(0) = 0.39/0.24$. Given differences in mortgage institutions and the broader economic environment between Ireland and the United States, I target the relative increase in attention instead of the level estimated in the Irish experiment.

Although $\chi^* = (G^\$/\theta_s)^\varepsilon$ does not depend directly on λ_{free}, notifications affect private attention through the value function. Free reconsideration raises both V and $\widetilde{V}^{refi}$, but it raises the value of remaining unrefinanced by more— $\partial V/\partial\lambda_{free} > \partial\widetilde{V}^{refi}/\partial\lambda_{free} > 0$— because it is immediately useful while the rate gap is open and not once the coupon has just been reset. The difference $\widetilde{V}^{refi} - V$ therefore narrows, the surplus $G^\$$ in (12) falls, and optimal private monitoring falls with it. The crowding-out is large enough to matter for the calibration. Holding the value function fixed, an arrival rate of 0.167 per year would reproduce the increase in the attentive share, while letting attention respond raises the required rate to 0.263, sixty percent higher.

Consumption effects. Notifications strengthen the consumption response to a rate cut. Over five years, the discounted cumulative response rises by 12.3 percent, from \$2,927 to \$3,289 per household. The amplification is larger at shorter horizons, 13.3 percent at one year and 13.5 percent at two, because free consideration opportunities bring refinancing forward. Panel (b) of Figure 12 traces the amplification over the horizon.

The gains are concentrated among households with weaker private attention. Relative to the baseline response, the increase is 18 to 20 percent in the bottom two income quintiles, compared with about 6 percent in the top two. This pattern is not imposed by the intervention, which provides the same arrival rate of free consideration opportunities to every income state. In dollar terms, however, the largest gains accrue to the middle of the income distribution because lower-income households start from a smaller baseline response. Figure 13 sets the two metrics side by side, with panel (a) showing the levels and panel (b) each quintile's share of the additional response.

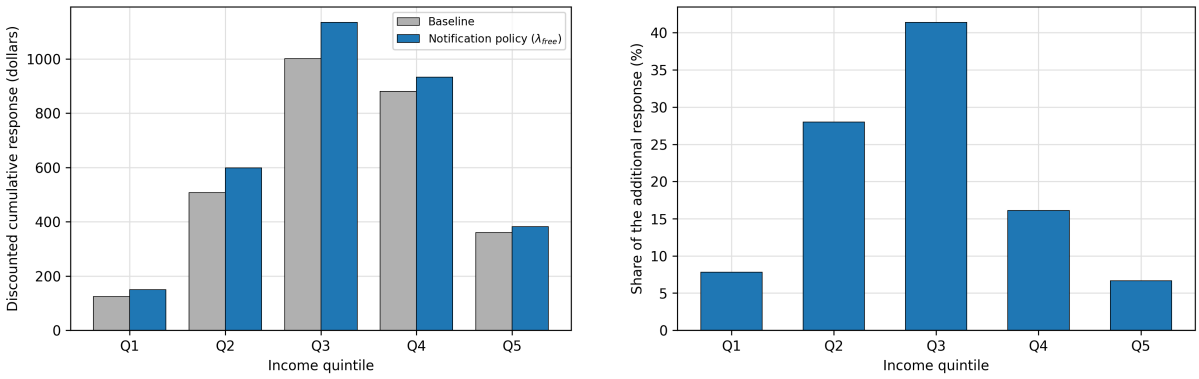
I interpret this experiment as a model-based accounting exercise rather than a welfare evaluation of an actual notification program. It quantifies how much monetary transmission can be recovered by providing costless consideration opportunities, conditional on the estimated friction structure and holding the mortgage-rate schedule and execution costs fixed.



(a) Consumption IRF (b) Cumulative amplification

Figure 12: Consumption Response to a Rate Cut, With and Without Notifications

Notes: The experiment is a 150 basis point cut in the policy rate—one step of the model's rate grid, which lowers the market mortgage rate by 85 basis points on impact. Panel (a) reports the percentage deviation of aggregate consumption from an otherwise identical no-cut path, in the baseline economy (solid) and under the notification counterfactual (dashed). Panel (b) reports the discounted cumulative counterfactual-minus-baseline gap as a percent of the baseline cumulative response. Both are drawn at the converged arrival rate $\hat{\lambda}_{free} = 0.263$ per year, with all estimated parameters held at their baseline values.



(a) Cumulative response by quintile (b) Contribution to the additional response

Figure 13: Distributional Decomposition of the Notification Counterfactual

Notes: Panel (a) reports the five-year discounted cumulative consumption response by income quintile in the baseline economy and under the notification counterfactual, in dollars per household. Panel (b) reports each quintile's share of the discounted counterfactual-minus-baseline increase. Quintiles are defined on income at origination. The rate cut, the initial distribution, and all model primitives are held fixed.

Sensitivity to the MPC profile. The notification gain converts extra refinancing cash flow into spending, so its size could in principle hinge on the model's MPC profile, which is untargeted and validated against Lewis et al. (2026). I test this inside the model, without imposing MPC schedules from outside. Tilting the discount rate by income, $\delta_s = \delta\,(z_s/\bar{z})^{\psi}$ with $\psi = \pm 0.25$, re-solves the full economy and widens or narrows the endogenous MPC gap between the bottom and top quintiles by about a quarter in each direction, while liquidity, balances, and attention remain equilibrium objects. Across these economies the notification gain moves from 12.48 to 12.49 percent of the baseline response, against 12.35 in the baseline, and its incidence is essentially unchanged (Table K.1 in Appendix K). The gain is anchored by the extra cash flow delivered to marginal refinancers—who sit in high-MPC, low-liquidity states in every economy—rather than by the level of the MPC gradient.

5.7 Implications for Inequality

Finally, I examine how reduced refinancing frictions change inequality in the cross-sectional distribution at the end of the 2005–2023 rate path. Table 5 reports the market-wide notification policy of the main experiment above, in which free consideration opportunities arrive at rate $\hat{\lambda}_{free} = 0.263$ per year in every income state.

The notification policy has little effect on overall consumption inequality but modestly compresses the distribution of liquid wealth. The consumption Gini is essentially unchanged, the top-10 share rises by 0.06 percentage points, and the bottom-50 share falls by 0.09. On the liquid-wealth side the Gini falls by 1.7 percent, the top-10 share by 0.26 percentage points, and the bottom-50 share rises by 0.64 points to 15.50 percent.

Table 5: Consumption and Liquid-Wealth Inequality under Reduced Refinancing Frictions

	Consumption				Liquid wealth			
Metric	Base	CF	CF–Base	%Δ	Base	CF	CF–Base	%Δ
Gini	0.30	0.30	+0.001	< 0.5%	0.49	0.49	-0.008	-1.70%
Top 10 share	20.41%	20.46%	+0.06 pp	< 0.5%	30.07%	29.81%	-0.26 pp	-0.86%
Bottom 50 share	28.70%	28.62%	-0.09 pp	< 0.5%	14.86%	15.50%	+0.64 pp	+4.33%

Notes: This table compares inequality in the baseline economy and the notification counterfactual. Both columns use the cross-sectional distribution in December 2023, after the economy follows the realized 2005–2023 rate path from the January 2005 initial distribution under each policy. Shares are read off the Lorenz curve, splitting the grid cell that contains each percentile cutoff in proportion. In the notification counterfactual, free consideration opportunities arrive at rate $\lambda_{free} = 0.263$ per year in every income state, which raises the share of attentive households by roughly 60 percent. Share measures are reported in percent, and CF–Base differences for share measures are reported in percentage points.

6 Conclusion

This paper examines how refinancing frictions are distributed across the income distribution and how that distribution shapes the transmission of monetary policy through household mortgage cash flows. In the Fannie Mae and Freddie Mac loan-level panel covering originations from 2000 to 2025, refinancing inequality is large in intensity, in timing, and in its response to policy. Once refinancing becomes financially advantageous, the bottom income quintile refinances about 43 percent less often than the top quintile, and the share of loans still unrefinanced falls to 80 percent within roughly four months for the top quintile but only after about ten months for the bottom. The unrealized savings left by lower-income households amount to approximately 6.7 percent of monthly income, a recurring shortfall over the remaining life of the loan. Following an identified 100 basis point easing, the cumulative refinancing response of the top quintile exceeds that of the bottom by 4.3 percentage points after two quarters. These patterns survive controls for credit quality, updated home equity, lender-side credit tightness, and origination-lender selection.

These facts anchor the structural model and are not reproduced by the benchmark refinancing models considered in the paper. In the model developed here, households optimally choose how intensively to monitor their mortgage by weighing the gains from refinancing against the price of attention, which generates an income gradient in refinancing endogenously and nests the

refinancing technologies of Berger et al. (2021) as boundary cases. Even with a common price of attention across income groups, differences in refinancing gains generate more monitoring at higher incomes, but they produce only about one third of the income gradient in the data. The common-cost model overpredicts refinancing in the bottom quintile and turns downward between the fourth and fifth quintiles. Allowing the price of attention to decline with income accounts for the remaining two thirds and better explains the patterns at both ends of the distribution.

The quantitative implications for monetary transmission are sizable. Guided by the field experiment of Byrne et al. (2023), I evaluate a market-wide notification policy that provides free opportunities to reconsider the mortgage and calibrate its arrival rate to reproduce the experiment's relative increase in attention. Under this policy, the five-year discounted cumulative consumption response to a rate cut rises by 12.3 percent, from \$2,927 to \$3,289 per household. The proportional increase is largest in the lower-income groups: the bottom quintile gains 20.4 percent relative to its baseline response, compared with 6.0 percent for the top quintile. In dollar terms, the largest increase occurs in the middle of the income distribution, where mortgage balances are larger than at the bottom while MPCs remain substantial.

By providing additional opportunities to reconsider refinancing, the notification policy raises the aggregate consumption response even after accounting for the resulting decline in private monitoring. The increase is larger among households with low baseline attention, shifting the additional refinancing response toward parts of the distribution where the consumption response to cash-flow relief is larger. More broadly, the results show that the distribution of refinancing frictions across households, not only their average level, matters for the transmission of monetary policy.

References

Agarwal, Sumit, John C. Driscoll, and David I. Laibson, "Optimal Mortgage Refinancing: A Closed-Form Solution," *Journal of Money, Credit and Banking*, 2013, *45* (4), 591–622.

_ , Richard J. Rosen, and Vincent Yao, "Why Do Borrowers Make Mortgage Refinancing Mistakes?," *Management Science*, 2016, *62* (12), 3494–3509.

_ , Souphala Chomsisengphet, Hua Kiefer, Leonard C. Kiefer, and Paolina C. Medina, "Refinancing Inequality During the COVID-19 Pandemic," *Journal of Financial and Quantitative Analysis*, August 2024, *59* (5), 2133–2163.

Aiyagari, S. Rao, "Uninsured Idiosyncratic Risk and Aggregate Saving," *The Quarterly Journal of Economics*, 1994, *109* (3), 659–684.

Andersen, Asger Lau, Niels Johannesen, Mia Jørgensen, and José-Luis Peydró, "Monetary Policy and Inequality," *The Journal of Finance*, 2023, *78* (5), 2945–2989.

Andersen, Steffen, John Y. Campbell, Kasper Meisner Nielsen, and Tarun Ramadorai, "Sources of Inaction in Household Finance: Evidence from the Danish Mortgage Market," *American Economic Review*, October 2020, *110* (10), 3184–3230.

Auclert, Adrien, "Monetary Policy and the Redistribution Channel," *American Economic Review*, June 2019, *109* (6), 2333–2367.

Beraja, Martin, Andreas Fuster, Erik Hurst, and Joseph Vavra, "Regional Heterogeneity and the Refinancing Channel of Monetary Policy*," *The Quarterly Journal of Economics*, February 2019, *134* (1), 109–183.

Berger, David, Konstantin Milbradt, Fabrice Tourre, and Joseph Vavra, "Mortgage Prepayment and Path-Dependent Effects of Monetary Policy," *American Economic Review*, September 2021, *111* (9), 2829–2878.

_ , _ , _ , and _ , "Refinancing Frictions, Mortgage Pricing and Redistribution," Working Paper 32022, National Bureau of Economic Research 2024.

_ , _ , _ , **and** _ , "Optimal Mortgage Refinancing with Inattention," *American Economic Review: Insights*, 2025, *7* (4), 497–515.

Byrne, Shane, Kenneth Devine, Michael King, Yvonne McCarthy, and Christopher Palmer, "The Last Mile of Monetary Policy: Inattention, Reminders, and the Refinancing Channel," Working Paper 31043, National Bureau of Economic Research 2023. Forthcoming, Journal of Finance.

Chen, Hui, Michael Michaux, and Nikolai Roussanov, "Houses as ATMs: Mortgage Refinancing and Macroeconomic Uncertainty," *The Journal of Finance*, 2020, *75* (1), 323–375.

Cloyne, James, Clodomiro Ferreira, and Paolo Surico, "Monetary Policy when Households have Debt: New Evidence on the Transmission Mechanism," *The Review of Economic Studies*, 2020, *87* (1), 102–129.

Coibion, Olivier, Yuriy Gorodnichenko, Lorenz Kueng, and John Silvia, "Innocent Bystanders? Monetary policy and inequality," *Journal of Monetary Economics*, June 2017, *88*, 70–89.

Cota, Marta and Ante Šterc, "Financial Skills and Search in the Mortgage Market," 2024. SSRN Working Paper 4787496.

Defusco, Anthony A. and John Mondragon, "No Job, No Money, No Refi: Frictions to Refinancing in a Recession," *The Journal of Finance*, 2020, *75* (5), 2327–2376.

Di Maggio, Marco, Amir Kermani, Benjamin J. Keys, Tomasz Piskorski, Rodney Ramcharan, Amit Seru, and Vincent Yao, "Interest Rate Pass-Through: Mortgage Rates, Household Consumption, and Voluntary Deleveraging," *American Economic Review*, November 2017, *107* (11), 3550–3588.

Eichenbaum, Martin, Sergio Rebelo, and Arlene Wong, "State-Dependent Effects of Monetary Policy: The Refinancing Channel," *American Economic Review*, March 2022, *112* (3), 721–761.

Fisher, Jack, Alessandro Gavazza, Lu Liu, Tarun Ramadorai, and Jagdish Tripathy, "Refinancing cross-subsidies in the mortgage market," *Journal of Financial Economics*, August 2024, *158*, 103876.

Flodén, Martin and Jesper Lindé, "Idiosyncratic Risk in the United States and Sweden: Is There a Role for Government Insurance?," *Review of Economic Dynamics*, 2001, *4* (2), 406–437.

Foulis, Angus, Jonathon Hazell, Atif Mian, and Belinda Tracey, "How Do Interest Rates Affect Consumption? Household Debt and the Role of Asset Prices," Working Paper 34911, National Bureau of Economic Research 2026.

Gerardi, Kristopher, Paul S. Willen, and David Hao Zhang, "Mortgage prepayment, race, and monetary policy," *Journal of Financial Economics*, March 2023, *147* (3), 498–524.

Greenwald, Daniel L., "The Mortgage Credit Channel of Macroeconomic Transmission," MIT Sloan Research Paper 5184-16, MIT Sloan School of Management 2018. Legacy version, February 2018.

Jordà, Òscar, "Estimation and Inference of Impulse Responses by Local Projections," *American Economic Review*, March 2005, *95* (1), 161–182.

Jørring, Adam Tejs and Ryan Westphal, "How Important is Household Heterogeneity for Monetary Policy? Evidence from Mortgage Refinancing," SSRN Working Paper 5150407, Social Science Research Network 2025.

Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante, "Monetary Policy According to HANK," *American Economic Review*, March 2018, *108* (3), 697–743.

Keys, Benjamin J., Devin G. Pope, and Jaren C. Pope, "Failure to refinance," *Journal of Financial Economics*, December 2016, *122* (3), 482–499.

Lewis, Daniel, Davide Melcangi, and Laura Pilossoph, "Latent Heterogeneity in the Marginal Propensity to Consume," *The Review of Economic Studies*, 2026. Advance access, published online 18 February 2026.

Maxted, Peter, David Laibson, and Benjamin Moll, "Present Bias Amplifies the Household Balance-Sheet Channels of Macroeconomic Policy," *The Quarterly Journal of Economics*, February 2025, *140* (1), 691–743.

Nakamura, Emi and Jón Steinsson, "High-Frequency Identification of Monetary Non-Neutrality: The Information Effect*," *The Quarterly Journal of Economics*, August 2018, *133* (3), 1283–1330.

Swanson, Eric T., "Measuring the effects of Federal Reserve forward guidance and asset purchases on financial markets," *Journal of Monetary Economics*, March 2021, *118*, 32–53.

Wong, Arlene, "Refinancing and the Transmission of Monetary Policy to Consumption," Working Paper 2021-57, Princeton University, Department of Economics 2021.

Zhang, David Hao, "Closing Costs, Refinancing, and Inefficiencies in the Mortgage Market," Working Paper, Harvard Business School 2022.

A Data Construction and Measurement

A.1 Data and Construction of the Loan-Month Panel

Universe. The data are the public single-family loan-level files of Fannie Mae and Freddie Mac, which report origination characteristics and a monthly performance record ending in a zero-balance code. Pooling the two publishers, I keep fixed-rate, single-family, owner-occupied first-lien mortgages in the fifty states and the District of Columbia originated between 2000 and 2025. Investor and second-home loans are excluded because they face different underwriting and refinancing incentives.

Income. The files report the underwriting PTI measure but not income. From the monthly principal-and-interest payment M_i implied by the original balance, rate, and term, I impute

$$\widehat{Y}_i = \frac{12M_i}{\mathrm{PTI}_i^{\mathrm{BE}}/100} \tag{30}$$

and assign quintiles within each origination year. Income is therefore predetermined with respect to the refinancing decision, and the specifications use the borrower's rank, not the level. Appendix A.4 shows that the level is understated because the reported ratio is a back-end measure, and corrects it where income enters as a level.

Incentive. The obtainable rate $\tilde{m}_{it}$ is the median note rate on refinance mortgages originated in the same quarter to observationally similar borrowers: the same state, credit-score band, LTV band, and income quintile, following Defusco and Mondragon (2020). The rate gap is $g_{it} = c_i - \tilde{m}_{it}$; the ADL-adjusted gap subtracts the threshold described in Appendix A.5. Both are evaluated at the state of the mortgage in the month before termination.

Collateral. Refinancing requires collateral, and collateral values move with local house prices. For each loan-month I construct an estimate of the current LTV ratio by scaling the property value implied at origination by the growth in the local house price index and dividing the current balance

by the result:

$$\widehat{\mathrm{LTV}}_{it} = 100 \times \frac{B_{it}}{\left(B_{i0}/\mathrm{LTV}_{i0}\right) \times 100 \times \left(H_{z(i),t}/H_{z(i),0}\right)}. \tag{31}$$

The index H is the FHFA's quarterly three-digit-ZIP all-transactions index, interpolated to monthly frequency. I use this index rather than a finer one because the loan data identify location at the three-digit ZIP level, so the index is native to the geography of the panel and no spatial crosswalk is required. From the same construction I derive the home-equity measures used as controls.

Sampling and design weights. The full population contains approximately 3.7 billion loan-month observations, making estimation on the complete sample computationally burdensome. I therefore subsample the data and reweight observations to recover population estimates. Sampling is conducted at the loan level and stratified by a match-quality indicator—whether the loan receives a *one-to-one* refinancing match (Appendix A.2). Loans with a one-to-one match are retained in full, while a uniform one-percent sample is drawn from all remaining loans based on the loan identifier. Complete monthly histories are retained for sampled loans in both strata. Retaining all one-to-one matched loans ensures sufficient observations for estimating the rare narrow refinancing outcome; baseline refinances in the sampled stratum are restored to their population representation through sampling weights.

Table A.1: Construction of the Loan-Month Analysis Sample

Group	Loan-months	Design weight
One-to-one matched refinancing loans	65,813,883	1
Prepaid, match not one-to-one	19,781,147	100
Prepaid, no plausible replacement	3,679,107	100
Never prepaid	12,992,598	100
Total	**102,266,735**	

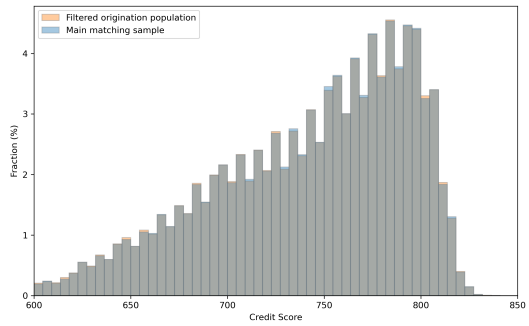
Notes: Fixed-rate, single-family, owner-occupied first-lien GSE mortgages in the fifty states and the District of Columbia, 2000–2025, covering 2,577,037 loans. One-to-one matched loans are retained in full; all other loans are drawn in a uniform one-percent sample. Applying the two design weights reconstructs a population risk set of 3.71 billion loan-months.

Each observation is weighted by the inverse of its sampling rate, so that the weighted sample represents the population of loan-months. The weights correct only for the sampling design;

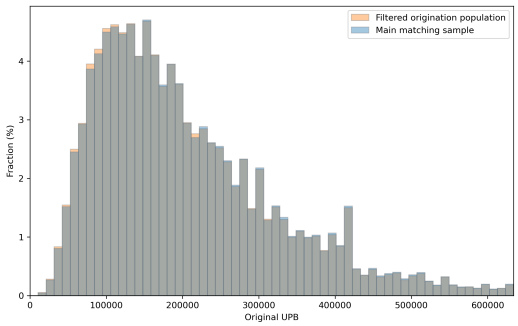
ambiguity in identifying the replacement loan is instead addressed through the definition of the refinancing outcome in Appendix A.3.

Validation of the Analysis Sample

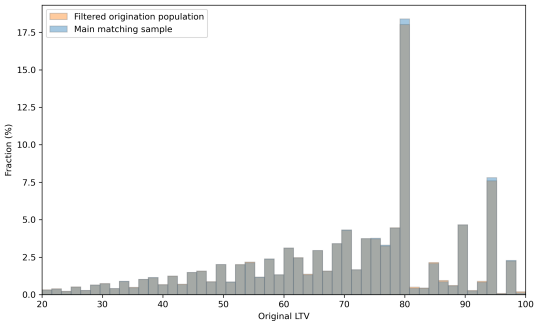
Figure A.1 and Table A.2 show that the weighted sample closely matches the population distributions and moments of credit score, loan balance, leverage, contract characteristics, and geography.



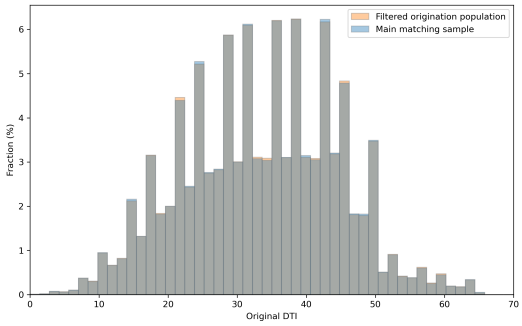
(a) Credit score



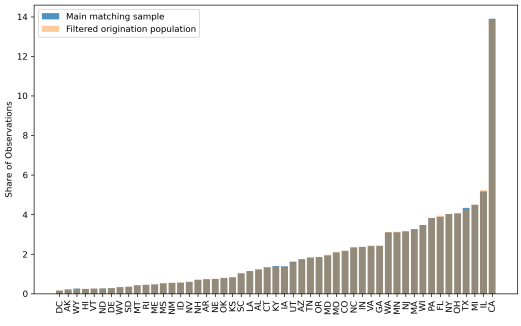
(b) Loan balance



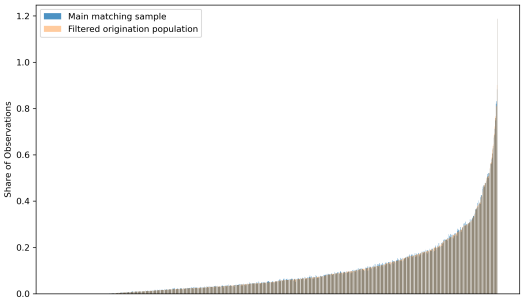
(c) LTV ratio



(d) PTI ratio



(e) Share of observations by state



(f) Share of observations by ZIP code

Figure A.1: GSE Population and Analysis Sample: Distributions of Key Variables

Table A.2: Data Validity Checks on the Matched Panel

	Mean			Median		
Variable	Population	Main sample	Difference	Population	Main sample	Difference
Credit Score	741.7	742.1	0.3	753.0	753.0	0.0
Loan Balance	208,756	210,379	1,624	178,000	179,000	1,000
Loan Term	306.4	306.9	0.5	360.0	360.0	0.0
LTV	70.0	69.9	-0.1	74.0	74.0	0.0
PTI	33.4	33.4	-0.0	34.0	34.0	0.0
Interest Rate	4.95	4.94	-0.00	4.88	4.88	0.00
CLTV	71.3	70.6	-0.7	75.0	75.0	0.0
Number of Borrowers	1.58	1.58	0.00	2.00	2.00	0.00
MI Percent	24.2	24.3	0.0	25.0	25.0	0.0

Notes: The table compares the matched main sample with the full GSE population using the weighted validation output. Main-sample moments apply inverse-probability sampling weights; population moments are unweighted loan-level moments. PTI is the GSE-reported underwriting measure. Population loan count is 67,267,355; main-sample loan count is 2,577,038.

A.2 Identifying Refinancing Events

The GSE performance files record that a loan terminated early, but not why: a zero balance is consistent with a refinance, a home sale, or an outright payoff. I identify refinances by linking each prepaid loan to the new origination that replaced it. This appendix describes that procedure and the classification it produces.

For each prepaid loan I first restrict the set of potential replacement loans that are ever compared to it. A potential replacement must agree exactly on state, three-digit ZIP area, metropolitan area, property type, number of units, and occupancy status, since a mortgage on a different property is not a refinance of that loan. It must also be consistent in timing and in size: its first scheduled payment must fall within three months after the old loan's payoff month, because the new loan pays off the old one at closing and its first payment comes due shortly afterward, and its original balance must lie within an admissible range around the old loan's payoff balance. These restrictions are deliberately permissive. Their purpose is to exclude what cannot be a refinance, not to select what is likely to be one; that is the job of the scoring step.

Each surviving pair then receives a Fellegi–Sunter score. The comparison fields are the timing of the replacement relative to payoff, the new-to-old balance ratio, the change in note rate, the transition in term, differences in credit score and reported PTI, and indicators for agreement in the number of borrowers, origination channel, and GSE. Each field is discretized into agreement bands, and its weight is the log ratio of its frequency among high-confidence links to its frequency among other within-block potential-replacement pairs. The same weights are applied to all pairs nationally.

For each prepaid old loan, the classifier ranks the surviving potential replacements by score, breaking exact score ties by new-loan ID. It compares the top score with the runner-up and counts how many old loans select the same new loan as their top potential replacement. It is not a global one-to-one assignment: a one-to-one status means that the top potential replacement clears the score-gap rule and has no competing top claim, not that the new loan independently selects that old loan as its globally best counterpart.

Table A.3: Refinancing-Match Classes

Class	Description
One-to-one	Top potential replacement clears the score-gap rule and has no competing top claim
One-to-many	Several potential replacements are equally plausible for this loan
Many-to-one	Best potential replacement is claimed by more than one prepaid loan
Many-to-many	Match is ambiguous on both sides
No match	No plausible replacement appears in the GSE data

I call the first four classes link-eligible prepayments. Subject to the coverage of the linked GSE data, a prepayment with no surviving potential replacement is not counted as a match-based refinance. Whether a particular replacement can be selected under the classifier is a much more demanding question than whether one exists, and its answer is not random across borrowers, as Appendix A.3 shows.

A.3 Ambiguous Matches and the Definition of the Refinancing Outcome

The narrowest definition of a refinance is a prepayment classified as a one-to-one match. The difficulty is that one-to-one status is an outcome of the linkage, and it is not achieved at random.

Among link-eligible prepayments, the share resolving to a one-to-one match rises monotonically with income, from 0.79 in the bottom quintile to 0.87 in the top, so a one-to-one matched outcome would drop more genuine refinances for low-income borrowers and steepen the estimated gradient even if propensities were identical. The same share peaks near a zero gap and falls steadily as the gap widens, from roughly 0.86 to 0.69 at gaps above three and a half percentage points, because large gaps arise in refinancing booms, when many similar loans are originated in the same area at once and the best potential replacement is hardest to separate from the runner-up. One-to-one matches are scarcest exactly where refinancing is most intense.

I therefore define the baseline event as a prepayment for which the linkage locates at least one plausible replacement, which does not condition on one-to-one status. In the income-adjusted in-the-money sample, the raw top-to-bottom ratio is 2.14 for this baseline outcome and 3.37 for the one-to-one matched outcome. After applying the full-control specification, the corresponding ratios are 1.74 and 3.18.

Table A.4: Refinancing Outcome Definitions and the Income Gradient

Outcome	Event	Raw Q5/Q1	Adjusted Q5/Q1
One-to-one matched	Prepayment classified one-to-one	3.37	3.18
Link-based *(baseline)*	Any plausible replacement	2.14	1.74
Prepayment	Any prepayment, no linkage used	2.08	–

Notes: Raw ratios compare the design-weighted event rates of the top and bottom income quintiles in the income-adjusted in-the-money sample. Adjusted ratios compare predicted refinancing probabilities from the full-control specification of Equation 2, estimated on the identical risk set with identical controls and design weights. Only the event definition differs across rows.

The ordering is what the first paragraph predicts, and its width measures how much of the gradient under the one-to-one definition the linkage itself produces. Among linkage-based outcome definitions, I report the baseline link-based outcome as a lower bound, with the one-to-one matched outcome as the upper bound: the adjusted gradient is bracketed between roughly 1.7 and 3.2, and every conclusion of the paper holds at the lower bound.

A.4 Payment-to-Income Measurement and the Imputation of Income

Because borrower income is not directly reported in the GSE files, I impute it from the mortgage terms recorded at origination. I first use the original balance, mortgage interest rate, and term to construct the monthly payment, and then divide this payment by the underwriting PTI ratio. This calculation would recover income exactly if the numerator of the ratio were the same payment, but two components create a wedge. A front-end, or housing, ratio also includes property taxes and insurance, while the back-end ratio reported in the GSE files additionally includes non-housing debt service. Income imputed from the mortgage payment alone therefore underestimates true income. This subsection derives a correction for this wedge, estimates it from the SCF by income quintile and year, and applies it to the unrealized-savings calculation.

Let Y be monthly borrower qualifying income, P the mortgage principal-and-interest payment, T property taxes and home insurance, and O non-housing debt service. The GSE ratio is a back-end ratio,

$$\text{PTI}^{\text{BE}} = \frac{P + T + O}{Y}, \tag{32}$$

while the payment recoverable from the loan file is P alone. Imputing income as $\widehat{Y} = P/\text{PTI}^{\text{BE}}$ therefore yields $\widehat{Y} = Y \cdot P/(P + T + O) < Y$, and recovering income requires scaling up by

$$\xi \equiv \frac{P + T + O}{P}, \qquad Y = \widehat{Y} \times \xi. \tag{33}$$

The bias is irrelevant wherever income enters as a ranking, which is the case for the refinancing regressions in Section 3 provided the distortion is close to proportional across the distribution. It matters wherever income enters as a level, as in the unrealized-savings analysis in Section 3.3, where forgone savings are expressed as a share of monthly income.

SCF-based income correction. I estimate ξ from the SCF, using the eight triennial waves from 2001 to 2022. The SCF reports for each household both a housing and a total debt PTI ratio, which is exactly the front-end and back-end pair the correction requires. All statistics are weighted and pool the multiply imputed replicates.

I restrict the sample to households that resemble GSE borrowers: owner-occupiers of their primary residence carrying a mortgage on it, with positive income. I exclude jumbo borrowers by dropping households whose primary-residence mortgage balance exceeds that year's national baseline conforming limit. Income quintiles are formed within this mortgagor sample in each wave, matching the relative-rank concept used in the loan panel. This leaves roughly 1,800 to 2,900 households per wave, or 360 to 580 per quintile-wave cell.

Two limitations constrain what can be estimated. The public SCF carries no geographic identifiers, so a region-specific correction would require restricted-access files; and the summary files do not identify single-family structures, so the SCF sample is slightly broader than the loan universe on that dimension.

Table A.5: SCF Front-End and Back-End PTI Correction

Income quintile	Front-end (%)	Back-end (%)	ξ_{SCF}	ξ_{full}
Q1	27.9	37.1	1.15	1.45
Q2	19.2	27.0	1.27	1.60
Q3	15.7	23.0	1.36	1.70
Q4	13.6	19.6	1.35	1.69
Q5	10.0	13.7	1.24	1.56

Notes: Weighted medians pooled across waves. ξ_{SCF} is the ratio of the back-end to the front-end PTI ratio; ξ_{full} rebases it onto principal and interest.

The observable ratio $\xi_{\text{SCF}} = 1+O/R$ has the household's reported payment R in its denominator, and about 62 percent of reported payments include taxes or insurance while the imputation uses principal and interest alone. Two factors close the gap. The housing factor $g \equiv (P+T)/P$ is recovered from the escrow split itself: the ratio of median R/P among escrow-inclusive households to the same median among escrow-exclusive households cancels any common error in the computed P and leaves $g \approx 1.28$, matching the 1.29 obtained by adding reported property taxes directly. The rebasing factor $f \equiv R/P \approx 1.16$ puts non-housing debt service, measured per unit of reported payment, on the same base. Hence

$$\xi_{\text{full}} = g + f\left(\xi_{\text{SCF}} - 1\right), \tag{34}$$

which runs from 1.45 to 1.70 and averages 1.60.

A.5 The ADL Threshold and Identification of In-the-money Loans

As described in the text, an in-the-money loan refers to a situation where the difference between the old and new interest rates after refinancing exceeds a certain threshold. This threshold represents the fixed costs associated with refinancing and option value. Notation in this subsection follows Agarwal et al. (2013) and is local to it.

$$\text{In-the-money} \quad \text{if} \quad \underbrace{(\text{Interest Rate}_i^{\text{old}} - \text{Potential Refi Mortgage Rate}_{i,t}^{\text{new}}) - \bar{g}_{it}^{ADL}}_{\equiv \text{ ADL-adjusted rate gap}} > 0$$

(i and t denote loan and time respectively)

Agarwal et al. (2013) propose the following tractable closed-form solution for the threshold value.

$$\begin{aligned}
\bar{g}_{it}^{ADL} &= \frac{1}{\psi}\left[\phi + W\left(-\exp\left(-\phi\right)\right)\right] \\
\psi &= \frac{\sqrt{2(\rho+\lambda)}}{\sigma} \\
\phi &= 1 + \psi(\rho+\lambda)\frac{\kappa(M)/M}{(1-\tau)}
\end{aligned}$$

Here, $W(\cdot)$ is the Lambert W-function, ρ is the real discount rate, λ is the expected real rate of exogenous mortgage repayment, σ is the standard deviation of the mortgage rate, $\kappa(M)/M$ is the ratio of the tax-adjusted refinancing cost and the remaining mortgage value, and τ is the marginal tax rate. The values of the parameters suggested by Agarwal et al. (2013) are $\sigma = 0.0109$, $\tau = 0.28$, $\rho = 0.05$.

$\kappa(M)$ is the cost entailed in refinancing, which is defined by:

$$\kappa(M) = F + fM\left[1 - \frac{\tau}{\zeta+\rho+\pi}\left\{\frac{1-\exp\left(-(\zeta+\rho+\pi)N\right)}{N}\frac{\rho+\pi}{\zeta+\rho+\pi} + \zeta\right\}\right]$$

F is the fixed cost of refinance, f is the points divided by 100, τ is the marginal tax rate, ζ is the expected arrival rate of a full deductibility event, such as moving and refinancing. Using the parameter suggested by the paper $F = 2000, f = 0.01, \zeta = \mu + 0.1 = 0.2$, the cost can be approximated as follows.

$$\kappa(M) = 2000 + 0.007905M$$

Lastly, λ is defined as the expected rate of decline in the real principal of the mortgage for reasons other than rate-reducing refinancing, and is defined as:

$$\lambda = \mu + \frac{i^{old}}{\exp(i^{old}\Gamma) - 1} + \pi,$$

Γ is the remaining life of the existing mortgage in years, i^{old} is the current (old) mortgage interest rate for existing loan, μ is the hazard of relocation, and π is the inflation rate. The parameter values used are $\mu = 0.1$ and $\pi = 0.03$, following the paper.

B Additional Empirical Results

B.1 Interest-Rate Reductions at Refinancing

Table B.1 reports the interest-rate reductions realized at refinancing by income quintile, estimated from the one-to-one matched old-to-new loan pairs used in Section 3.3. Uncontrolled rate reductions are larger for lower-income borrowers, but the ordering reverses once borrower and loan controls are added: in columns (2)–(5), higher-income borrowers obtain larger reductions.

Table B.1: Interest Rate Reductions by Refinancing

Dependent	Interest Rate Reduction by Refi (pp)				
Equation	(1)	(2)	(3)	(4)	(5)
Income Quintile 2	-0.051[a]	0.021[a]	0.022[a]	0.021[a]	0.021[a]
	(0.007)	(0.003)	(0.002)	(0.003)	(0.003)
Income Quintile 3	-0.087[a]	0.039[a]	0.040[a]	0.050[a]	0.050[a]
	(0.010)	(0.003)	(0.003)	(0.004)	(0.004)
Income Quintile 4	-0.116[a]	0.060[a]	0.064[a]	0.087[a]	0.086[a]
	(0.012)	(0.003)	(0.003)	(0.005)	(0.005)
Income Quintile 5	-0.138[a]	0.106[a]	0.114[a]	0.150[a]	0.148[a]
	(0.015)	(0.004)	(0.005)	(0.007)	(0.007)
Constant	–	–	–	–	–
Control variables					
LTV, credit score, loan age, rate gap, ZIP code		x	x	x	x
Credit tightness			x	x	x
Loan balance				x	x
Estimated current LTV					x
Observations	1,227,447	1,227,421	1,222,936	1,222,936	1,222,883
R-squared	0.172	0.609	0.610	0.611	0.611

	Predicted Interest Rate Reduction by Refi (pp)				
Equation	(1)	(2)	(3)	(4)	(5)
Income Quintile 1	1.0758	0.9343	0.9304	0.9138	0.9144
	(0.0096)	(0.0027)	(0.0028)	(0.0039)	(0.0039)
Income Quintile 2	1.0251	0.9548	0.9520	0.9348	0.9354
	(0.0041)	(0.0013)	(0.0015)	(0.0026)	(0.0025)
Income Quintile 3	0.9890	0.9729	0.9708	0.9639	0.9641
	(0.0022)	(0.0009)	(0.0010)	(0.0013)	(0.0012)
Income Quintile 4	0.9601	0.9938	0.9940	1.0009	1.0007
	(0.0026)	(0.0010)	(0.0011)	(0.0017)	(0.0017)
Income Quintile 5	0.9382	1.0399	1.0449	1.0634	1.0628
	(0.0062)	(0.0020)	(0.0020)	(0.0031)	(0.0030)

Notes: This table presents weighted realized old-to-new interest-rate reductions among one-to-one matched refinances that are in the money under the simple rate-gap definition. The simple rate gap is the existing coupon minus the loan- and borrower-characteristic-adjusted market rate. Columns progressively add borrower and loan controls, ZIP-code fixed effects, credit tightness, loan balance, and estimated current LTV. The predicted reductions are the adjusted predictions used in the unrealized-savings calculation. Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively. Standard errors are in parentheses and clustered by ZIP code and origination vintage year-quarter. Source: Fannie Mae and Freddie Mac loan-level panel, 2000–2025.

B.2 Monthly Refinancing Probability Estimates

Table B.2 reports the monthly LPM coefficient estimates of (2) under the ADL-adjusted in-the-money criterion, the counterpart of Table 2 in the main text. All income-quintile coefficients are positive, monotone, and significant at the 1 percent level. In the full-control specification, the predicted monthly refinancing probability rises from 1.95 percent in the bottom quintile to 3.37 percent in the top, a gap of 1.42 percentage points, so bottom-quintile borrowers refinance approximately 42 percent less often than top-quintile borrowers under this stricter definition.

Table B.2: Monthly Refinancing Probability: ADL Gap

Dependent	Refinancing Indicator × 100			
Equation	(1)	(2)	(3)	(4)
Income Quintile 2	0.721[a]	0.484[a]	0.471[a]	0.496[a]
	(0.083)	(0.048)	(0.047)	(0.047)
Income Quintile 3	1.090[a]	0.717[a]	0.706[a]	0.686[a]
	(0.122)	(0.059)	(0.053)	(0.049)
Income Quintile 4	1.578[a]	1.082[a]	1.054[a]	1.007[a]
	(0.168)	(0.082)	(0.074)	(0.064)
Income Quintile 5	2.204[a]	1.567[a]	1.497[a]	1.424[a]
	(0.188)	(0.104)	(0.095)	(0.071)
Control variables				
LTV, credit, loan age, rate inc., ZIP		x	x	x
Credit tightness, HARP			x	x
Loan balance, Est. LTV				x
Observations	6,249,157	6,249,157	6,247,646	6,247,480

Notes: Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively. Standard errors are in parentheses. This table reports weighted linear-probability-model coefficients on income-quintile indicators relative to the bottom quintile. The sample is restricted to in-the-money observations under the ADL-adjusted rate-gap definition. The dependent variable is monthly refinancing, scaled by 100, so coefficients are percentage-point differences. Column (1) includes origination-vintage year-quarter fixed effects; columns (2)–(4) also include ZIP-code fixed effects and add the controls indicated in the table. Standard errors are clustered by ZIP code and origination vintage year-quarter. Source: Fannie Mae and Freddie Mac loan-level panel, 2000–2025.

Table B.3 and Table B.4 report the predicted monthly refinancing probabilities by income quintile underlying the main-text estimates, for each specification under the simple and ADL-adjusted in-the-money definitions.

Table B.3: Predicted Monthly Refinancing Probability: Simple Rate Gap

Dependent	Refinancing Indicator × 100			
Equation	(1)	(2)	(3)	(4)
Income Quintile 1	1.237[a]	1.318[a]	1.333[a]	1.421[a]
	(0.039)	(0.028)	(0.023)	(0.021)
Income Quintile 2	1.670[a]	1.695[a]	1.706[a]	1.790[a]
	(0.017)	(0.014)	(0.012)	(0.011)
Income Quintile 3	1.984[a]	1.964[a]	1.967[a]	1.961[a]
	(0.012)	(0.009)	(0.009)	(0.009)
Income Quintile 4	2.315[a]	2.260[a]	2.251[a]	2.156[a]
	(0.029)	(0.022)	(0.019)	(0.019)
Income Quintile 5	2.752[a]	2.666[a]	2.633[a]	2.475[a]
	(0.052)	(0.040)	(0.035)	(0.031)
Control variables				
LTV, credit, loan age, rate inc., ZIP		x	x	x
Credit tightness, HARP			x	x
Loan balance, Est. LTV				x
Observations	47,415,612	47,415,612	47,114,678	47,111,510

Notes: Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively. Standard errors are in parentheses. This table reports weighted predicted monthly refinancing probabilities by income quintile from the linear probability models. The sample is restricted to in-the-money observations under the simple rate-gap definition. Entries are percentage points. Column (1) includes origination-vintage year-quarter fixed effects; columns (2)–(4) also include ZIP-code fixed effects and add the controls indicated in the table. Standard errors are clustered by ZIP code and origination vintage year-quarter. Source: Fannie Mae and Freddie Mac loan-level panel, 2000–2025.

Table B.4: Predicted Monthly Refinancing Probability: ADL Gap

Dependent	Refinancing Indicator × 100			
Equation	(1)	(2)	(3)	(4)
Income Quintile 1	1.576[a]	1.908[a]	1.929[a]	1.949[a]
	(0.105)	(0.051)	(0.045)	(0.039)
Income Quintile 2	2.297[a]	2.392[a]	2.400[a]	2.445[a]
	(0.040)	(0.030)	(0.029)	(0.027)
Income Quintile 3	2.666[a]	2.624[a]	2.635[a]	2.634[a]
	(0.026)	(0.020)	(0.020)	(0.021)
Income Quintile 4	3.153[a]	2.989[a]	2.984[a]	2.956[a]
	(0.068)	(0.039)	(0.037)	(0.035)
Income Quintile 5	3.780[a]	3.474[a]	3.426[a]	3.373[a]
	(0.091)	(0.062)	(0.059)	(0.046)
Control variables				
LTV, credit, loan age, rate inc., ZIP		x	x	x
Credit tightness, HARP			x	x
Loan balance, Est. LTV				x
Observations	6,249,157	6,249,157	6,247,646	6,247,480

Notes: Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively. Standard errors are in parentheses. This table reports weighted predicted monthly refinancing probabilities by income quintile from the linear probability models. The sample is restricted to in-the-money observations under the ADL-adjusted rate-gap definition. Entries are percentage points. Column (1) includes origination-vintage year-quarter fixed effects; columns (2)–(4) also include ZIP-code fixed effects and add the controls indicated in the table. Standard errors are clustered by ZIP code and origination vintage year-quarter. Source: Fannie Mae and Freddie Mac loan-level panel, 2000–2025.

B.3 Reporting-Time Fixed Effects

Table B.5 re-estimates the full-control specification with reporting-time fixed effects at the month, quarter, or year level in place of origination-vintage fixed effects, as discussed in Section 3.1.

Table B.5: Baseline LPM with Reporting-Time Fixed Effects

Dependent	Refinancing Indicator × 100					
In-the-money definition	Simple Rate Gap			ADL Gap		
Time FE	Month	Quarter	Year	Month	Quarter	Year
Income Quintile 2	0.34[a]	0.34[a]	0.35[a]	0.46[a]	0.46[a]	0.46[a]
	(0.02)	(0.02)	(0.04)	(0.04)	(0.02)	(0.06)
Income Quintile 3	0.49[a]	0.49[a]	0.50[a]	0.66[a]	0.66[a]	0.65[a]
	(0.03)	(0.03)	(0.06)	(0.05)	(0.04)	(0.07)
Income Quintile 4	0.64[a]	0.64[a]	0.65[a]	0.92[a]	0.92[a]	0.91[a]
	(0.03)	(0.05)	(0.08)	(0.06)	(0.07)	(0.12)
Income Quintile 5	0.91[a]	0.91[a]	0.92[a]	1.31[a]	1.31[a]	1.29[a]
	(0.05)	(0.08)	(0.11)	(0.08)	(0.08)	(0.15)
Control variables	x	x	x	x	x	x
ZIP-code FE	x	x	x	x	x	x
Reporting-time FE	x	x	x	x	x	x
Observations	47,111,510	47,111,510	47,111,510	6,247,480	6,247,480	6,247,480

Notes: Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively. Standard errors are in parentheses. This table re-estimates the weighted full-control monthly refinancing linear probability model after replacing origination-vintage fixed effects with reporting-time fixed effects. Standard errors are clustered by ZIP code and the corresponding reporting-time fixed effect. Source: Fannie Mae and Freddie Mac loan-level panel, 2000–2025.

B.4 Twelve-Month Refinancing Probability Estimates

The monthly hazards in Section 3.1 measure the flow probability of refinancing conditional on being in the money. Small differences in monthly hazards compound over time, and a policy-relevant summary is how many borrowers act on a refinancing opportunity within a fixed horizon. For each loan, let T_i^{ITM} denote the first month in which loan i becomes in the money, and define

$$Refi_Indicator_12m_i = 100 \times \mathbf{1}\left\{\sum_{\tau=0}^{11} Refi_Indicator_{i,T_i^{ITM}+\tau} > 0\right\}.$$

I regress this indicator on income-quintile dummies with the same borrower and loan controls, ZIP-code fixed effects, and origination vintage year-quarter fixed effects as the baseline monthly specification, so that each loan contributes a single observation indexed by its first in-the-money entry. I report results under both the simple and ADL-adjusted rate-gap definitions.

Table B.6: Twelve-Month Cumulative Refinancing Probability by Income Quintile

Dependent	Refinancing within 12 Months × 100	
In-the-money definition	Simple Rate Gap	ADL Gap
Income Quintile 1	14.64[a]	28.23[a]
	(0.41)	(0.51)
Income Quintile 2	15.78[a]	31.23[a]
	(0.24)	(0.32)
Income Quintile 3	17.30[a]	32.51[a]
	(0.13)	(0.30)
Income Quintile 4	19.27[a]	34.33[a]
	(0.25)	(0.30)
Income Quintile 5	22.01[a]	37.39[a]
	(0.45)	(0.56)
Q5–Q1 gap	7.37	9.16
Q5/Q1 ratio	1.50	1.32
Observations	1,977,302	411,238

Notes: Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively. Standard errors are in parentheses. This table reports weighted predicted probabilities of refinancing within twelve months after a loan first becomes in the money. Each loan contributes one observation at its first in-the-money month. Entries are percentage points. The full-control specification includes borrower and loan controls, ZIP-code fixed effects, and origination-vintage year-quarter fixed effects. Standard errors are clustered by ZIP code and origination vintage year-quarter. Source: Fannie Mae and Freddie Mac loan-level panel, 2000–2025.

Table B.6 reports the main estimates. Under the ADL-gap definition, the predicted twelve-month refinancing probability rises monotonically from 28.2 percent in the bottom income quintile to 37.4 percent in the top quintile, a top-to-bottom ratio of approximately 1.32. The coefficient on each higher income quintile is positive, monotone, and statistically significant at the 1 percent level. Results under the simple rate-gap definition show a qualitatively similar top-to-bottom ratio. The predicted twelve-month probability rises from 14.6 percent in the bottom quintile to 22.0 percent in the top quintile, a ratio of 1.50. In this specification, all higher income quintiles are statistically different from the bottom quintile. Full regression estimates are reported in Appendix Table B.7.

These estimates have an economically intuitive interpretation. Within one full year after a refinancing opportunity first arises, lower-income borrowers are substantially less likely to have acted on it. Under the ADL-gap definition, roughly three quarters of bottom-quintile borrowers remain in their original mortgage twelve months after first becoming in the money, compared

with about two thirds of top-quintile borrowers. This pattern indicates that the income gradient in monthly hazards documented above does not reflect small per-month differences alone. Those differences compound into a large cross-income inequality in the cumulative take-up of refinancing opportunities.

Table B.7: Twelve-Month Refinancing Probability: Regression Estimates

Dependent	Refinancing within 12 Months × 100	
In-the-money definition	Simple Rate Gap	ADL Gap
Income Quintile 2	1.14[a]	3.00[a]
	(0.29)	(0.57)
Income Quintile 3	2.66[a]	4.28[a]
	(0.44)	(0.60)
Income Quintile 4	4.64[a]	6.10[a]
	(0.63)	(0.69)
Income Quintile 5	7.37[a]	9.16[a]
	(0.83)	(0.98)
Observations	1,977,302	411,238

Notes: Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively. Standard errors are in parentheses. This table reports weighted coefficients on income-quintile indicators relative to the bottom quintile. The dependent variable equals 100 if a loan has a refinancing event within twelve months after first becoming in the money, and zero otherwise. The full-control specification includes borrower and loan controls, ZIP-code fixed effects, and origination-vintage year-quarter fixed effects. Standard errors are clustered by ZIP code and origination vintage year-quarter. Source: Fannie Mae and Freddie Mac loan-level panel, 2000–2025.

C Robustness Checks

C.1 Robustness to Income-Rank Misclassification

Income is not observed directly in the GSE data but inferred from the reported back-end PTI ratio. Because the back-end ratio includes non-mortgage debt payments, I use the SCF to construct an income correction factor ξ and apply its income-quintile average to the GSE measure. Appendix A.4 describes this correction in detail. Applying the correction at the quintile level leaves open the possibility that household-level variation in ξ changes some borrowers' income ranks. I examine this possibility using four exercises.

The SCF estimates show substantial within-quintile variation in ξ, with coefficients of variation of roughly 0.4–0.5. I use this variation directly in the first exercise and then examine the sensitivity of the estimates to alternative changes in income ranks. Each exercise re-estimates the full-control specification of the baseline linear probability model in Table B.3. To limit computation, the robustness specifications use the same fixed 30% loan subsample; the baseline row reports the full-sample estimate. I report results for both the simple rate-gap and ADL-adjusted in-the-money samples.

(1) Misclassification Monte Carlo. For each loan, I draw a household-specific correction factor ξ_i from the SCF distribution for its income quintile and rescale backed-out income as $\tilde{y}_i = y_i \times \xi_i$. I then assign income quintiles using the original within-origination-year cutoffs and re-estimate the model. I repeat the procedure 25 times and report the mean and the 5th–95th percentile range of the top-to-bottom gap. About half of the loans change quintile in each replication. Because the measured dispersion in ξ_i also contains survey sampling variation and small-denominator noise, these simulations allow for substantial reassignment across income groups.

(2) Boundary trim. I drop loans whose within-year income percentile lies within ±5% of a quintile cutoff and re-estimate the model using the original income assignment for the remaining loans. These are the observations whose quintile assignments are most sensitive to small changes in

measured income. If assignments near the cutoffs drive the estimated gradient, removing these observations should reduce the gap.

(3) Boundary swap. I move the top and bottom 10% of each income quintile into the neighboring quintile and re-estimate the model. This exercise reverses the ordering of observations near each cutoff by moving higher-income observations downward and lower-income observations upward.

(4) Placebo. I randomly assign income quintiles while preserving their marginal shares and re-estimate the model. Random assignment removes the information in income rank while leaving the number of observations in each quintile unchanged.

Table C.1 reports the results. In the misclassification Monte Carlo, the top-to-bottom gap falls from 1.05 to 0.77 percentage points under the simple rate gap and from 1.42 to 0.98 percentage points under the ADL-adjusted gap, despite about half of the loans changing quintile. The 5th–95th percentile ranges are narrow. Dropping observations near the quintile cutoffs increases the gaps to 1.12 and 1.59 percentage points, respectively. The 10% boundary swap produces gaps close to the baseline estimates, while random quintile assignment reduces them to approximately zero. The income gradient is therefore present under each reassignment that retains information about income rank. In these exercises, errors in income rank reduce the estimated gradient rather than account for it.

Table C.1: Robustness of the Income Refinancing Gradient to Income-Quintile Misclassification

	Simple rate gap		ADL-adjusted gap	
	Q5−Q1 (pp)	Q5/Q1	Q5−Q1 (pp)	Q5/Q1
Baseline (full sample)	1.05	1.74	1.42	1.73
(1) Misclassification Monte Carlo	0.77	1.50	0.98	1.45
[5–95% band]	[0.75, 0.80]		[0.85, 1.09]	
(2) Boundary trim (±5%)	1.12	1.81	1.59	1.84
(3) Adversarial boundary swap (10%)	1.06	1.75	1.50	1.76
(4) Placebo (random quintiles)	≈ 0.00	≈ 1.00	≈ 0.00	≈ 1.00

Notes: Each entry is the top-minus-bottom (Q5−Q1) difference, in percentage points, and the ratio (Q5/Q1) of the predicted monthly refinancing probability by income quintile from the full-control linear probability model. The baseline row is the full-sample estimate; every other row re-estimates the model on a fixed 30% loan subsample. In the misclassification Monte Carlo, each loan draws a household-specific back-to-front income-correction factor from the SCF within-quintile distribution, the backed-out income is rescaled and loans are re-ranked into quintiles, and the model is re-estimated (25 replications; approximately 46–49% of loans change quintile per replication); the reported value is the mean and the bracketed range is the 5th–95th percentile across replications. The boundary trim drops loans within ±5% (income percentile) of each quintile cutoff; the adversarial swap moves the top and bottom 10% of each quintile into the neighboring quintile; the placebo assigns income quintiles at random while preserving the marginal quintile shares.

C.2 Credit Score and PTI Bins

Table C.2: Refinancing Probability by Income Quintile within PTI Bins

PTI bin	< 36	$[36, 43)$	$[43, 45)$	$[45, 50)$	$[50, 60)$	≥ 60
Income Q2	0.389^a	0.584^a	0.521^a	0.638^a	0.724^a	0.848^a
	(0.023)	(0.033)	(0.054)	(0.049)	(0.069)	(0.117)
Income Q3	0.656^a	0.974^a	0.897^a	1.026^a	1.120^a	1.182^a
	(0.029)	(0.050)	(0.086)	(0.096)	(0.127)	(0.215)
Income Q4	0.898^a	1.349^a	1.247^a	1.420^a	1.387^a	1.596^a
	(0.042)	(0.094)	(0.120)	(0.132)	(0.160)	(0.316)
Income Q5	1.244^a	1.759^a	1.379^a	1.599^a	1.347^a	1.118
	(0.060)	(0.114)	(0.148)	(0.158)	(0.205)	(0.682)

Notes: The sample is partitioned by PTI bins corresponding to common underwriting thresholds: PTI < 36, $[36, 43)$, $[43, 45)$, $[45, 50)$, $[50, 60)$, and ≥ 60. Within each PTI bin, I estimate the weighted full-control monthly refinancing linear probability model in the simple rate-gap in-the-money sample. The specification includes borrower and loan controls, ZIP-code fixed effects, and origination-vintage year-quarter fixed effects. Reported coefficients are percentage-point differences relative to the bottom income quintile. Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively; standard errors are clustered by ZIP code and origination vintage year-quarter.

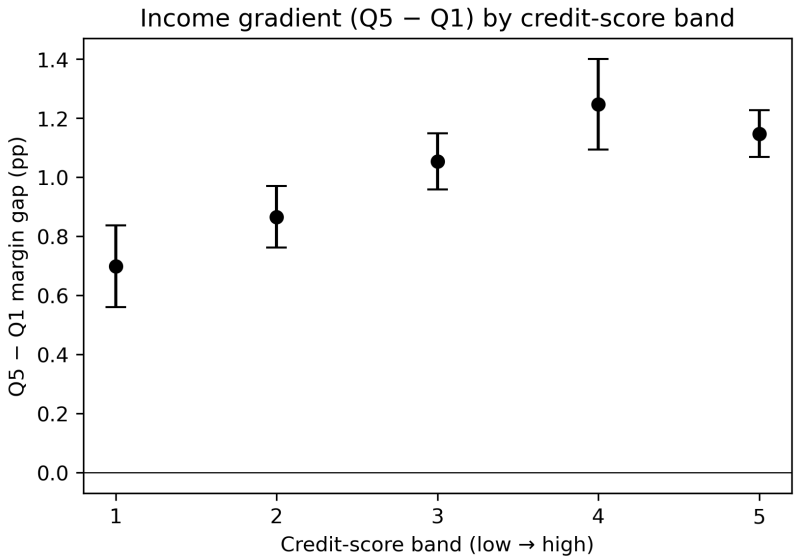


Figure C.1: Income Gradient in Refinancing by Credit-Score Band

Notes: The sample is divided into five bands of the origination credit score. Within each band, the figure reports the estimated Q5–Q1 difference in monthly refinancing probability from the full-control linear-probability specification using the simple rate-gap in-the-money sample. The specification includes the baseline borrower and loan controls, ZIP-code fixed effects, and origination-vintage year-quarter fixed effects, except that credit score itself is omitted because it is constant within a band. Dots are point estimates and bars are 95% confidence intervals based on standard errors clustered by ZIP code and origination vintage year-quarter.

C.3 Origination Lender-Year Fixed Effects

This section uses the Fannie Mae– and Freddie Mac–HMDA linked panel to examine whether the refinancing gradient reflects sorting into different origination lenders or different lender-year vintages. The HMDA match identifies each GSE loan's origination lender, which allows me to compare borrowers within the same origination lender-year cell while also controlling for the origination quarter and borrower location.

Table C.3: HMDA–GSE Origination-Lender Match Coverage

Statistic	Count
Panel coverage	2000–2025
Monthly GSE loan observations	102,266,735
Monthly observations with origination lender ID	9,020,395
Unique GSE loans	2,577,037
Matched Freddie Mac loans	160,506
Matched Fannie Mae loans	83,917
Unique GSE loans with origination lender ID	244,423
Unique origination lenders	11,058
Rate-gap in-the-money observations with origination lender ID	4,716,116
ADL-gap in-the-money observations with origination lender ID	850,827

Notes: The table summarizes the national 2000–2025 GSE loan-month panel and its Fannie Mae– and Freddie Mac–HMDA origination-lender links. Monthly observations are loan-month observations. The in-the-money rows are the samples used in the simple rate-gap and ADL-gap origination-lender-year fixed-effect regressions. Sources: Fannie Mae and Freddie Mac loan-level panel, 2000–2025, and HMDA.

The estimating equation is

$$Refi_{it} = \sum_{q=2}^{5} \beta_q IncomeQuintile_{iq} + X_{it}'\Gamma + \alpha_{z(i)} + \lambda_{v(i)} + \mu_{\ell(i)\times y(i)} + \epsilon_{it}, \tag{35}$$

where $Refi_{it}$ is the monthly refinancing indicator multiplied by 100, $IncomeQuintile_{iq}$ denotes income-quintile indicators relative to the lowest quintile, and X_{it} contains the baseline loan and borrower controls. The fixed effects include ZIP-code fixed effects $\alpha_{z(i)}$, origination year-quarter fixed effects $\lambda_{v(i)}$, and origination lender-year fixed effects $\mu_{\ell(i)\times y(i)}$, defined by HMDA origination lender ID interacted with origination year. Standard errors are clustered by ZIP code and origination year-quarter, as in the HMDA regression output.

Table C.4 shows that the income gradient remains after adding origination lender-year fixed effects. In the full-control specification, borrowers in the highest income quintile are 1.04 percentage points more likely to refinance than borrowers in the lowest quintile under the simple rate-gap definition, essentially unchanged from the 1.05 percentage point benchmark estimate. The coefficients for income quintiles 2 through 5 are positive and statistically significant.

Because the comparison is within origination lender-year cells, the near-unchanged estimate substantially weakens the explanation that the benchmark gradient is generated by sorting into

original lenders or lender-vintage-specific conditions. The data do not observe lender outreach, screening, or approval at refinancing, so this test does not rule out lender-side channels operating after origination.

Table C.4: Monthly Refinancing Probability with Origination Lender-Year Fixed Effects (Simple Rate Gap)

Dependent variable	Refinancing indicator × 100
Specification	Full controls
Income Quintile 2	0.482[a]
	(0.055)
Income Quintile 3	0.630[a]
	(0.055)
Income Quintile 4	0.742[a]
	(0.071)
Income Quintile 5	1.043[a]
	(0.084)
Control variables and fixed effects	
ZIP-code FE	x
Origination year-quarter FE	x
Origination lender-year FE	x
LTV, credit score, loan age, rate incentive	x
Credit tightness, loan balance, estimated LTV	x
Observations	4,700,385

Notes: Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively. Standard errors are in parentheses and clustered by ZIP code and origination year-quarter. The weighted dependent variable is monthly refinancing, scaled by 100; coefficients are percentage-point differences relative to income quintile 1. The sample is restricted to in-the-money observations with a positive income-rate incentive and a matched HMDA origination lender. Origination lender-year fixed effects interact the HMDA-linked origination lender with origination year. The table reports the full-control estimate. Sources: Fannie Mae and Freddie Mac loan-level panel, 2000–2025, and HMDA.

C.4 Income Mobility

The benchmark assigns income rank at origination, whereas the refinancing decision occurs later. Table C.5 therefore uses the PSID to measure two-year changes in income rank among owner-occupier mortgagors. The 21,217 household-pairs span the 2005–2021 biennial waves, and the two-year interval brackets the 21-month median and 32-month mean loan age at refinancing in the matched GSE loans. Income quintiles are defined within the PSID mortgagor distribution in each wave, so the table measures the stability of relative rank. Transitions between Q1 and Q5 are uncommon: 0.9 percent of households initially in Q1 move to Q5, while 1.4 percent initially in Q5 move to Q1. More broadly, 80.6 percent of households initially in the bottom two quintiles remain in Q1 or Q2 after two years, and 82.6 percent of those initially in the top two remain in Q4 or Q5. Income rank can therefore change over the horizon relevant for refinancing, but most households remain within the same broad part of the income distribution. Moves into adjacent quintiles can still attenuate the Q5–Q1 gradient, which motivates the rank-perturbation exercises reported above.

Table C.5: Changes in Income Quintile: Two-Year Transition Matrix (PSID)

	Income quintile after two years				
Income quintile at t	Q1	Q2	Q3	Q4	Q5
Q1 (lowest)	71.3	20.0	5.8	1.9	0.9
Q2	19.4	50.5	22.2	6.3	1.6
Q3	5.5	19.8	47.5	23.1	4.1
Q4	2.3	5.7	19.3	53.4	19.3
Q5 (highest)	1.4	2.1	4.1	15.7	76.7

Notes: Two-year income-quintile transition matrix for owner-occupier mortgagor households in the PSID (biennial waves 2005–2021; 21,217 household-pairs). The sample keeps PSID reference persons who own their home (A19=1) and have a mortgage (A23=1), the closest available match to the single-family, owner-occupied, mortgaged GSE panel. Income quintiles are constructed within the PSID mortgagor cross-section each wave and each household is followed from wave t to wave $t+1$ (two years). Two years is chosen to mirror the average time-to-refinancing of the one-to-one matched GSE loans (mean loan age at refinancing ≈32 months; median ≈21 months). Cells report the row-normalized probability (%) of moving from the origin quintile (row) to the destination quintile (column). Rows sum to 100.

C.5 Freshness of Origination Controls

One concern is that several controls can change after origination, whereas the benchmark uses the values recorded when the loan was originated. Table C.6 therefore repeats the full-control regression after first restricting the sample to in-the-money loan-months under the simple rate-gap definition and then imposing maximum loan ages of 12, 24, 36, and 60 months. For younger loans, origination income, credit, and leverage measures have had less time to drift away from their values at the refinancing decision, so this exercise asks whether the income gradient remains when that concern is less acute.

The coefficients are not directly comparable across age caps. Because the sample is restricted to loans that are in the money, a short cap can overrepresent cohorts that entered that region soon after origination, including cohorts with relatively high original coupons. A longer cap also adds loans that have remained in the in-the-money region without refinancing, potentially selecting borrowers with low refinancing attention or unobserved constraints.

With these composition changes in mind, Table C.6 shows a positive Q5–Q1 coefficient significant at the 1 percent level in every column. The coefficient is 0.92 percentage points under the 12-month cap and remains between 1.2 and 1.3 percentage points for the 24-, 36-, and 60-month caps. The smaller estimate under the tightest cap may reflect its cohort composition and the shorter window over which refinancing differences can appear. The persistence of the gradient provides supporting evidence that changes in the origination controls do not mechanically generate the benchmark result.

Table C.6: Regression by Loan Age Restrictions

Dependent	Refinancing Indicator × 100			
Loan Age ≤ (months)	12	24	36	60
Income Quintile 2	0.157[a]	0.344[a]	0.426[a]	0.422[a]
	(0.034)	(0.036)	(0.039)	(0.030)
Income Quintile 3	0.278[a]	0.559[a]	0.645[a]	0.627[a]
	(0.055)	(0.054)	(0.054)	(0.040)
Income Quintile 4	0.540[a]	0.825[a]	0.912[a]	0.855[a]
	(0.082)	(0.079)	(0.075)	(0.059)
Income Quintile 5	0.920[a]	1.244[a]	1.316[a]	1.233[a]
	(0.112)	(0.100)	(0.096)	(0.076)
Constant	–	–	–	–
Observations	12,430,584	21,261,591	26,775,282	33,786,801

Notes: Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, respectively. Standard errors are in parentheses. Each column reports the weighted full-control refinancing linear probability model for the indicated maximum loan age in the simple rate-gap in-the-money sample. The specification includes borrower and loan controls, ZIP-code fixed effects, and origination-vintage year-quarter fixed effects. The sample is first restricted to in-the-money loan-months. Changing the age cap therefore changes both the cohorts that enter the sample and the survivor composition. Standard errors are clustered by ZIP code and origination vintage year-quarter. Source: Fannie Mae and Freddie Mac loan-level panel, 2000–2025.

C.6 Stability Across Calendar Windows

Table C.7 estimates the full-control model in overlapping five-year calendar windows. This exercise examines whether the income gradient appears across different refinancing cycles. It differs from both the income-mobility and origination-freshness exercises. The top-to-bottom gradient is positive and statistically significant in every displayed window, although its magnitude varies with the rate environment and the cohorts that are in the money in each period.

Table C.7: Five-Year Rolling Window Estimates

Rolling window end year	2005	2007	2009	2011	2013	2015	2017	2019
Income Quintile 2	0.643[a]	0.554[a]	0.243[a]	0.296[a]	0.374[a]	0.409[a]	0.276[a]	0.182[a]
	(0.066)	(0.057)	(0.025)	(0.021)	(0.020)	(0.031)	(0.019)	(0.016)
Income Quintile 3	0.958[a]	0.820[a]	0.343[a]	0.444[a]	0.540[a]	0.513[a]	0.371[a]	0.235[a]
	(0.071)	(0.066)	(0.032)	(0.026)	(0.024)	(0.042)	(0.029)	(0.021)
Income Quintile 4	1.201[a]	1.062[a]	0.425[a]	0.585[a]	0.707[a]	0.755[a]	0.481[a]	0.277[a]
	(0.090)	(0.084)	(0.040)	(0.030)	(0.028)	(0.053)	(0.031)	(0.021)
Income Quintile 5	1.490[a]	1.370[a]	0.508[a]	0.942[a]	1.056[a]	1.149[a]	0.738[a]	0.461[a]
	(0.091)	(0.075)	(0.069)	(0.043)	(0.050)	(0.076)	(0.046)	(0.032)

Notes: Superscripts a, b, c denote significance at 1%, 5%, and 10% levels, and standard errors, clustered by ZIP code and origination vintage year-quarter, are in parentheses. Each column reports the weighted full-control linear probability model of Equation (2) estimated on a five-year window ending in the indicated year, on in-the-money observations under the simple rate-gap definition. The dependent variable is monthly refinancing, scaled by 100.

C.7 Estimating Unrealized Savings by Constructing Counterfactual Rate Gaps

As an alternative approach to quantifying unrealized savings, I estimate income-group-specific rate gaps directly from the population of in-the-money loan-months, without relying on the matched-loan sample. This method constructs a counterfactual refinancing interest rate for each borrower and translates it into an implied monthly payment reduction.

Step 1: Estimating the predicted rate gap by income group. I regress the rate gap $RateGap_{it} \equiv m^*_{it} - m_{it}$ on income quintile indicators and the full set of controls other than the rate gap itself, restricting the sample to in-the-money observations ($gap > 0$):

$$RateGap_{it} = \alpha + \sum_{q=2}^{5} \hat{\beta}_q \cdot \mathbf{1}\{IncomeQ_i = q\} + \Gamma' \cdot X_{it} + \nu_{z(i)} + \lambda_{v(i)} + \epsilon_{it} \tag{36}$$

where m^*_{it} is borrower i's contracted coupon rate and m_{it} is the prevailing obtainable rate. The coefficients $\hat{\beta}_q$ capture the income-quintile gradient in the rate gap after controlling for borrower and loan characteristics, ZIP fixed effects, and vintage fixed effects.

Step 2: Constructing the counterfactual refinancing interest rate. Using the estimated income-group predicted rate gap $\widehat{RateGap}_{q(i)}$, I construct a counterfactual new coupon rate for each

in-the-money loan-month:

$$\hat{m}_{it}^{new} = m_{it}^{*} - \widehat{RateGap}_{q(i)} \tag{37}$$

This counterfactual rate represents what a borrower in income quintile q would pay if they refinanced and captured the income-group-specific rate benefit. I then translate this into an implied monthly payment reduction using loan-specific balance F_{it} and current coupon m_{it}^{*}. The existing payment uses remaining term N_{it}, while the counterfactual refinance resets maturity to 360 months:

$$\widehat{\Delta Pay}_{it} = PMT(F_{it},\, m_{it}^{*},\, N_{it}) - PMT\left(F_{it},\, \hat{m}_{it}^{new},\, 360\right) \tag{38}$$

Step 3: Aggregating unrealized savings by income group. For each in-the-money loan-month, I divide $\widehat{\Delta Pay}_{it}$ by the borrower's SCF-corrected monthly income. As in the main analysis, the correction uses income-quintile-by-origination-year estimates of the relationship between front-end and back-end PTI in the SCF. I then average these loan-month-level ratios within each income quintile.

The results are reported in Table C.8. Consistent with the matched-loan estimates, the implied monthly payment reduction is largest relative to income for the lowest-income borrowers, amounting to approximately 6.9 percent of corrected monthly income for the bottom quintile compared with 3.8 percent for the top quintile. The similarity of these estimates to those from the matched-loan approach corroborates the robustness of the main finding. Forgone refinancing savings are disproportionately large relative to income for lower-income households, precisely those with the highest MPC.

Table C.8: Unrealized Savings under the Counterfactual Rate-Gap Approach

Income Quintile	Avg. Income ($)	Avg. Balance ($)	Est. Rate Gap (pp)	Mthly. Pymt. Reduction ($)	Reduction / Income (%)
1	24,052	76,692	0.97	133.11	6.85
2	45,175	119,868	0.93	199.64	5.36
3	66,583	156,384	0.91	255.87	4.66
4	91,971	196,869	0.88	323.84	4.26
5	153,473	240,069	0.85	448.77	3.76

Notes: Rate gaps are weighted adjusted predictions from the full-control counterfactual rate-gap regression estimated among in-the-money loan-months. The regression includes the same controls and fixed effects as column (4) of Tables 2 and B.2, except that the rate-gap control is omitted because rate gap is the dependent variable. Predictions are applied to in-the-money at-risk loan-months defined using the income-adjusted obtainable-rate gap. Payment savings compare the existing payment at remaining maturity with a counterfactual refinance that resets maturity to 30 years. Average income and the reduction relative to income use SCF-corrected borrower income. Sources: Fannie Mae and Freddie Mac loan-level panel, 2000–2025, and the Survey of Consumer Finances.

D Additional Results on Differential Refinancing Response to Monetary Policy Shocks

This appendix reports additional results for the monetary-policy refinancing exercise in Section 3.4. The purpose is twofold. First, I verify that the high-frequency monetary-policy shocks from Swanson (2021) generate a strong first-stage movement in mortgage rates and a sizable average refinancing response. Second, I show that the short-run income gradient documented in the main text is robust to alternative refinancing-response measures and to a high-versus-low income grouping.

First stage and average refinancing response. I first report first-stage diagnostics for the endogenous mortgage-rate interactions used in the income-gradient local-projection instrumental-variable (LP-IV) specifications. Table D.1 shows that the high-frequency monetary-policy instruments are jointly strong across the relevant quarterly lags. The Wald statistics are large and the associated p-values are effectively zero, indicating that the monetary surprises have substantial explanatory power for the residualized endogenous interactions used in the second stage.

Table D.1: First-stage Diagnostics for $\Delta m_{t-k}\times$ Income Score

	Quintile rank				Grouped high-low			
$\Delta m_{t-k}\times$ income score	N	R^2	Wald χ^2	p-value	N	R^2	Wald χ^2	p-value
t	205,939	0.483	132.43	< 0.001	123,546	0.484	133.44	< 0.001
t-1	205,939	0.449	109.33	< 0.001	123,546	0.451	110.55	< 0.001
t-2	205,939	0.504	232.50	< 0.001	123,546	0.505	236.25	< 0.001
t-3	205,939	0.476	271.37	< 0.001	123,546	0.475	276.72	< 0.001
t-4	205,939	0.473	195.43	< 0.001	123,546	0.473	198.05	< 0.001
t-5	205,939	0.495	162.42	< 0.001	123,546	0.495	163.28	< 0.001
t-6	205,939	0.434	82.75	< 0.001	123,546	0.436	83.57	< 0.001

Notes: This table reports first-stage diagnostics for the endogenous regressor $\Delta m_{t-k} \times s_i$, where Δm_{t-k} is the lagged change in the 30-year market mortgage rate and s_i is the income score. Quintile rank uses $s_i = (q_i - 1)/4$, so Q1 equals 0 and Q5 equals 1. Grouped high-low uses $s_i \in \{0, 0.5, 1\}$. Denote the Swanson high-frequency monetary-policy factors (federal funds rate, forward guidance, and large-scale asset purchases) by $\mathcal{H}_t = (FFR_t, FG_t, LSAP_t)'$. The excluded instruments are $\mathcal{H}_{t-\ell} \times s_i$ for $\ell = 0, \ldots, 6$. The Wald statistic is the joint test that these excluded instruments have zero coefficients in the first stage. Specifications residualize the endogenous regressor, instruments, and controls X with respect to income-by-ZIP fixed effects and date fixed effects; they use outstanding-loan weights and two-way clustered standard errors by ZIP3 and date.

I next estimate the average refinancing response, pooling across income groups. This exercise is not used to identify the income gradient, but it provides a useful benchmark for the magnitude of the differential response in the main text. The specification is the same LP-IV design used in Section 3.4, except that the dependent variable is the aggregate regional refinancing response. The preferred dependent variable is the cumulative sum of quarterly refinancing rates,

$$\mathcal{R}_{zt}^{h} = \sum_{k=0}^{h} 100 \times \frac{x_{z,t+k}}{y_{z,t+k}},$$

where x_{zt} is the number of refinances in region z and quarter t, and y_{zt} is the corresponding outstanding mortgage stock.

Figure D.1 shows the average refinancing response to a 1 percentage point decline in the 30-year mortgage rate. The cumulative response peaks at 16.6 percentage points after eight quarters. After the peak, the cumulative response declines as refinance originations fall and new mortgage originations expand the outstanding-loan stock, mechanically lowering the cumulative refinancing rate. The magnitude is broadly comparable to Eichenbaum et al. (2022), though the comparison is not one-for-one. They report cumulative refinancing responses to a 50 basis point mortgage-rate

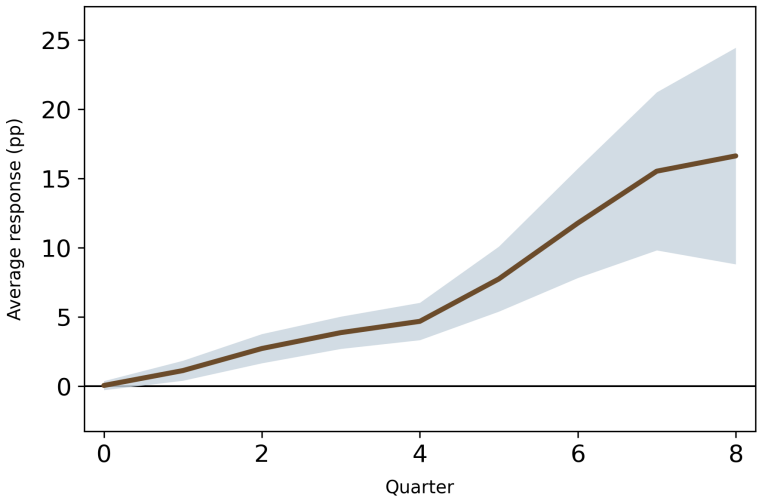


Figure D.1: Average Refinancing Response to a Monetary Policy Easing

Notes: The figure reports local-projection instrumental-variable estimates of the average cumulative quarterly refinancing response in the regional panel. The specification includes region and quarter fixed effects, six lags of refinancing and local macroeconomic controls, and is weighted by the outstanding mortgage stock. High-frequency monetary-policy surprises instrument mortgage-rate changes. The shock is normalized to a 1 percentage point decline in the 30-year mortgage rate. Shaded regions are 95% confidence intervals based on two-way clustering by region and quarter.

cut that peak at roughly 2–10 percent after about five quarters, depending on borrowers' initial rate gaps. Scaled linearly to the same 50 basis point cut, my estimate implies a peak of about 8 percentage points, which falls inside their range. Differences in sample coverage, mortgage types, time period, and response construction further limit a direct comparison.

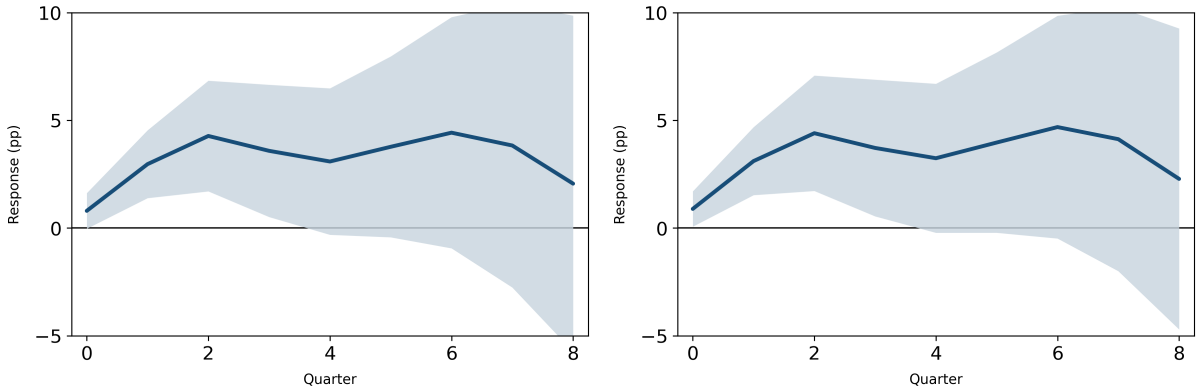
Alternative refinancing-response measures. The main text uses the cumulative sum of quarterly refinancing rates because it closely matches the refinancing-rate construction in Eichenbaum et al. (2022). I now verify that the short-run income gradient is not driven by this normalization. I consider two alternative measures. The first normalizes cumulative refinances by the average outstanding stock over the response window:

$$\mathcal{R}^{h,1}_{zqt} = 100 \times \frac{\sum_{k=0}^{h} x_{zq,t+k}}{\left(\sum_{k=0}^{h} y_{zq,t+k}\right)/(h+1)}. \tag{39}$$

The second normalizes cumulative refinances by the pre-shock outstanding stock:

$$\mathcal{R}^{h,2}_{zqt} = 100 \times \frac{\sum_{k=0}^{h} x_{zq,t+k}}{y_{zq,t-1}}. \tag{40}$$

The first measure allows the outstanding stock to evolve over the response horizon, while the second fixes the denominator at the pre-shock stock. Both measures are estimated using the same date fixed effects, ZIP-by-income fixed effects, controls, instruments, weights, and two-way clustered standard errors as in the main specification.



(a) Average-stock denominator (b) Pre-shock denominator

Figure D.2: Robustness: Q5–Q1 Differential Response under Alternative Refinancing Measures

Notes: The figure reports local-projection instrumental-variable estimates of the Q5–Q1 differential cumulative refinancing response in the quarterly regional panel. Panel (a) divides cumulative refinances by the average outstanding mortgage stock over the response window. Panel (b) divides them by the pre-shock stock. Both specifications include ZIP-by-income-quintile and quarter fixed effects, six lags of refinancing and local macroeconomic controls, and are weighted by the outstanding mortgage stock. High-frequency monetary-policy surprises instrument mortgage-rate changes interacted with income rank. The shock is normalized to a 1 percentage point decline in the 30-year mortgage rate. Shaded regions are 95% confidence intervals based on two-way clustering by region and quarter.

Figure D.2 shows that the short-run income gradient is essentially unchanged under both alternative response measures. Under the average-stock denominator, the Q5–Q1 differential response reaches 4.28 percentage points after two quarters. Under the pre-shock denominator, it reaches 4.41 percentage points. In both cases, the response is positive and statistically significant over the first several quarters, closely matching the preferred specification in the main text.

Alternative income grouping. As a final robustness check, I replace the linear income-rank specification with a coarser grouping of income cells. I define a low-income group as Q1–Q2, a

middle-income group as Q3, and a high-income group as Q4–Q5. I then estimate the same LP-IV specification after replacing $\tilde{q}$ with a group-rank variable equal to zero for the low-income group, one-half for the middle group, and one for the high-income group. The coefficient is therefore interpretable as the high-minus-low differential response to the monetary-policy shock.

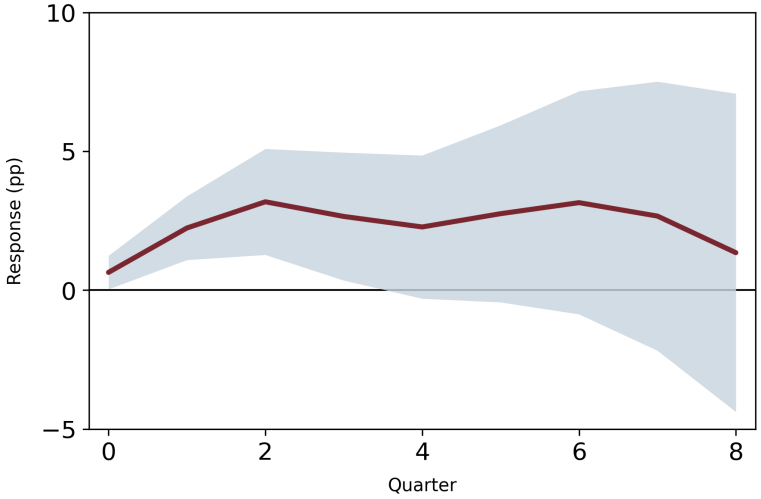


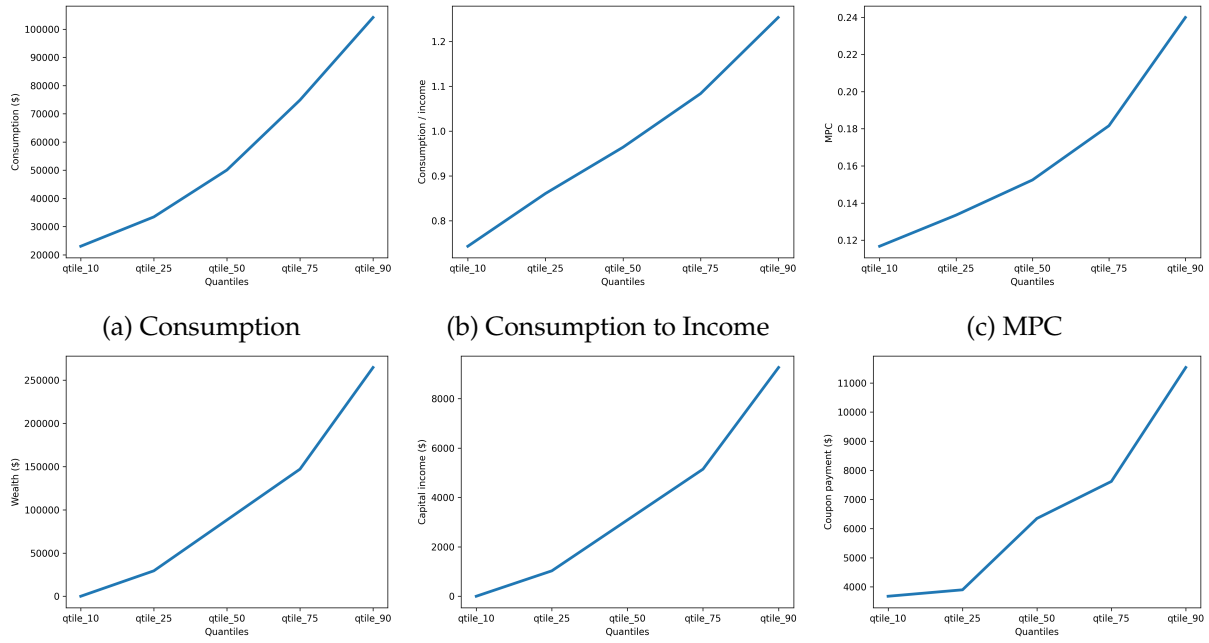
Figure D.3: Robustness: High–Low Differential Refinancing Response

Notes: The figure reports local-projection instrumental-variable estimates of the high-minus-low differential cumulative refinancing response in the quarterly regional panel. The low-income group is Q1–Q2, the middle group is Q3, and the high-income group is Q4–Q5. The specification includes ZIP-by-income-group and quarter fixed effects, six lags of refinancing and local macroeconomic controls, and is weighted by the outstanding mortgage stock. High-frequency monetary-policy surprises instrument mortgage-rate changes interacted with income-group rank. The shock is normalized to a 1 percentage point decline in the 30-year mortgage rate. Shaded regions are 95% confidence intervals based on two-way clustering by region and quarter.

Figure D.3 confirms the same pattern. The high-minus-low differential response is positive on impact, rises to about 3.2 percentage points after two quarters, and remains statistically significant through horizon three. The magnitude is slightly smaller than the Q5–Q1 specification because the grouped comparison averages over the bottom two and top two quintiles, but the timing and sign are nearly identical. Overall, the appendix results show that the main monetary-policy finding is robust. Refinancing responses to easing shocks are concentrated among higher-income borrowers in the first year after the shock.

E Cross-Sectional Moments of the Baseline Model

This appendix reports the cross-sectional distributions of the baseline model referenced in Section 5.1.



(a) Consumption (b) Consumption to Income (c) MPC

(d) Wealth (e) Capital Income (f) Mortgage Coupon Payment

Figure E.1: Cross-Sectional Moments of the Baseline Model

Notes: Each panel plots the named variable across percentiles of its distribution (the 10th, 25th, 50th, 75th, and 90th) in the baseline model's simulated cross-section: consumption, the consumption-to-income ratio, the MPC, liquid wealth, capital income, and the mortgage coupon payment.

F Numerical Solution

This appendix describes the numerical solution of the household problem and the evolution of the cross-sectional distribution.

State space. The individual state is liquid wealth W, mortgage balance F, existing coupon m^*, and income state s; the aggregate state is the short rate r. Liquid wealth is discretized on 18 points and the short rate on 8. Income follows the five-state discretization of Section 5.1. Because collateral is fixed at origination, the household problem is solved separately for each of the three collateral types of Section 4. The balance grid is finer for larger collateral, with 5, 7, and 13 points in the three blocks, since a larger loan can take a wider range of balances between origination and full amortization. Aggregates combine the blocks using the population weights in Table 4.

Household problem and attention fixed point. I solve (16) by implicit finite differences in pseudo-time. Each iteration forms $B = \text{diag}(\chi^* + \delta + 1/\Delta) - \mathcal{A}$, where $\mathcal{A}$ is the generator under the current policies and Δ is the pseudo-time step, and solves $BV' = \chi^* \widetilde{V}^{refi} + u(C) - q(\chi^*) + V/\Delta$. The wealth drift uses upwind differences, with the direction chosen by comparing the Hamiltonians implied by the forward and backward derivatives. The balance drift needs no upwinding, because $b_F = -\alpha F$ is signed, so amortization enters as a fixed downward advection operator. Income and short-rate transitions enter through their transition-rate matrices, and exogenous coupon resets and refinancing enter as jump blocks.

Endogenous attention adds a fixed point, because χ^* depends on the refinancing surplus while the surplus depends on V and V_W. The value function and the attention policy are updated within the same iteration rather than in nested loops. Given the current value function I compute the optimal cash-out choice, form $G^{\$}(r, x)$ from (12), and set $\chi^* = \min\{(G^{\$}/\theta_s)^{\varepsilon}, \bar{\chi}\}$, so the cap enters the update itself rather than being imposed afterwards. Both the execution indicator and the attention intensity are damped with weight one half before entering the next linear solve, which prevents the execution region from oscillating between iterations. Iteration stops when the scaled value-function change falls below 10^{-7}, with a ceiling of 60 iterations.

Distribution dynamics. Given the converged policies, the same generator defines the forward equation (20), and the distribution is advanced by its adjoint. The drift and transition blocks are those used in the household problem. The jump blocks move mass at intensity ν_m following exogenous coupon resets and at $\lambda^{refi}(r,x) = \chi^*(r,x)D(r,x)$ following refinancing. The post-refinancing state generally does not lie on the grid, because cash-out is a continuous choice, so each refinancing transition distributes mass across neighboring nodes with interpolation weights that sum to one. Typically four nodes receive mass. I verify that the rows of the generator sum to zero, so the operator conserves mass.

Stationary objects solve the forward equation with one equation replaced by the normalization that the distribution integrates to one. For transition experiments I evolve the distribution forward from a common initial distribution at weekly steps over a fifty-year horizon, under the policies implied by each environment, and compute aggregates by integrating the household policy functions over the evolving distribution.

G The Analytical Benchmark

This appendix proves Proposition 2 and shows how the S-shape of the hazard in the rate gap arises. The benchmark simplifies the model of Section 4 to obtain closed forms, and the full model is solved numerically.

G.1 Environment and the Surplus Equation

Assumptions. Following Berger et al. (2025), the benchmark makes four simplifying assumptions. The household values its mortgage liability separately from its wealth, $V(x) \approx V_0(W,s) - u'(\bar{C}_s)L(g,F)$, and the rate gap g is fixed between refinancing events. There is no cash-out, and refinancing extinguishes the liability, $L(0,F) = 0$. Under these assumptions $V_W \approx u'(\bar{C}_s)$ and $\widetilde{V}^{refi} - V = u'(\bar{C}_s)[L(g,F) - \kappa]$, so the marginal utility cancels in (12) and $G^{\$} = L(g,F) - \kappa_{eff}$, which does not depend on W. The liquidity channel in Section 4.7 relaxes the first and third.

Discounting and breakeven. Refinancing saves gF_t per unit of time. The saving is discounted at δ, ends when the household moves at rate ν_m, and shrinks as the balance amortizes at rate α, so its present value is

$$\int_0^\infty e^{-(\delta+\nu_m+\alpha)t}\, gF\, dt = \frac{gF}{\tilde{\rho}}, \qquad \tilde{\rho} \equiv \delta + \nu_m + \alpha. \tag{41}$$

Setting (41) equal to κ_{eff} gives the breakeven $\underline{g}(F) = \tilde{\rho}\kappa_{eff}/F$ in (21), a zero-net-present-value threshold. The optimal threshold of Agarwal et al. (2013) lies above it because a household facing a stochastic gap gains from waiting for a larger one. The two coincide when the gap is fixed, as in the benchmark.

Surplus. Under optimal monitoring, the liability L is the present value of the extra interest paid plus monitoring costs, less the expected gain from refinancing later. It solves

$$\tilde{\rho}\, L = gF + \min_{\chi \geq 0}\Big\{\theta_s \frac{\chi^{1+1/\varepsilon}}{1+1/\varepsilon} - \chi\, G^{\$}\Big\}. \tag{42}$$

The inner minimization is the problem of Proposition 1, with value $-(G^{\$})^{1+\varepsilon}/[(1+\varepsilon)\theta_s^{\varepsilon}]$. Substituting this value, $L = G^{\$} + \kappa_{eff}$, and $\tilde{\rho}\kappa_{eff} = F\underline{g}$ gives

$$\frac{(G^{\$})^{1+\varepsilon}}{(1+\varepsilon)\,\theta_s^{\varepsilon}} + \tilde{\rho}\, G^{\$} = F\big(g - \underline{g}(F)\big). \tag{43}$$

The right side is the flow value of the raw gain. The first term on the left is the flow value of the option to refinance at a later opportunity, which refinancing now gives up, and it grows with $G^{\$}$ because a larger surplus induces more monitoring. The left side rises from zero in $G^{\$}$, so the root is unique for $g > \underline{g}(F)$.

Proof of Proposition 2. At $\varepsilon = 1$, $G^{\$} = \theta_s\chi^*$ and (43) becomes $\chi^{*2} + 2\tilde{\rho}\chi^* = 2F(g - \underline{g})/\theta_s$, whose positive root is (22). For general ε, near the breakeven $G^{\$} \approx F(g - \underline{g})/\tilde{\rho}$, and far from it $G^{\$} \approx [(1+\varepsilon)\theta_s^{\varepsilon}F(g - \underline{g})]^{1/(1+\varepsilon)}$, so $\chi^* = (G^{\$}/\theta_s)^{\varepsilon}$ gives (23). □

G.2 The Shape of the Hazard

Hazard of a single loan. With $R \equiv \sqrt{\tilde{\rho}^2 + 2F(g - \underline{g})/\theta_s}$ and $\varepsilon = 1$, differentiating (22) above the breakeven gives

$$\frac{\partial \chi^*}{\partial g} = \frac{F}{\theta_s R} > 0, \qquad \frac{\partial^2 \chi^*}{\partial g^2} = -\frac{F^2}{\theta_s^2 R^3} < 0, \qquad \frac{\partial \chi^*}{\partial F}\Big|_g = \frac{g}{\theta_s R} > 0, \qquad \frac{\partial \chi^*}{\partial \theta_s} = -\frac{F(g - \underline{g})}{\theta_s^2 R} < 0,$$

and implicit differentiation of (43) gives the same signs for general ε. Attention is zero below $\underline{g}(F)$, turns on there with slope $F/(\theta_s \tilde{\rho})$, and rises at a decreasing rate above it. The monthly hazard (15) inherits each property, so for one loan the curve is an inaction region followed by a concave rise, with no inflection point.

Aggregation within an income group. Balances, and with them breakevens, differ within an income group. Let $\mathbb{E}_s$ denote the average over loans in income state s. The group hazard is $\bar{h}_s(g) = \mathbb{E}_s[h(g; F)]$, and a loan is active when $F \geq \tilde{\rho}\kappa_{eff}/g$, so larger loans turn on first. Because h is continuous at the breakeven, the slope of the group hazard is the average slope over active loans,

$$\bar{h}_s'(g) = \mathbb{E}_s\left[h_g(g; F)\,\mathbf{1}\{F \geq \tilde{\rho}\kappa_{eff}/g\}\right]. \tag{44}$$

Just above the smallest breakeven in the group, few loans are active. As g rises, more loans cross their breakevens and add their liftoff slopes, so the group hazard is convex. Farther in the money, most loans are already active, and the concavity of each loan dominates, so the group hazard is concave. The group hazard therefore has an inflection point and an S-shape, even though each loan's hazard is concave above its breakeven. The breakeven $\tilde{\rho}\kappa_{eff}/F$ sets the location of the rise, the dispersion of balances and the liftoff slope $F/(\theta_s \tilde{\rho})$ set its steepness, and θ_s and ε set its height far in the money through (23).

G.3 Identification

The expressions above organize the estimation in Section 5. The location of the rise identifies κ_{eff}, so I fix κ from closing-cost data and read $\Phi = \kappa_{eff} - \kappa$. The curvature above the breakeven

identifies ε, and given ε the height far in the money identifies θ. The level of the hazard across income groups at a common balance composition identifies ω, given the liquidity channel that the model generates from its own wealth distribution.

H The Berger et al. (2021) Benchmarks as Boundary Cases

This appendix supports the claim in Section 4.4 that the model generalizes the specification it builds on. Both refinancing technologies in Berger et al. (2021), the Calvo benchmark of their main text and the menu-cost specification they report alongside it, arise from the model of Section 4 at boundary values of the attention parameters.

Berger et al.'s refinancing technologies. Berger et al. (2021) write their model with two attention channels: a costless Calvo arrival at rate χ_c, on which the household refinances whenever it is in the money, and a costly arrival at rate χ_f, on which it pays the closing cost and refinances only if that beats waiting, which generates an endogenous gap threshold. Their main-text benchmark sets $\chi_f = 0$, leaving a Calvo specification in which the closing cost is never paid and no gain threshold applies.

Constant-arrival limits. Two restrictions on the attention block eliminate the gain sensitivity of endogenous monitoring and produce a constant-arrival technology. The first drives the elasticity to zero. As $\varepsilon \to 0$ the exponent $1 + 1/\varepsilon$ in the cost (9) diverges while its coefficient $\varepsilon/(1+\varepsilon)$ vanishes, so monitoring becomes free below unit intensity and unboundedly costly above it. The optimal intensity $\chi^* = (G^\$/\theta_s)^\varepsilon$ of Proposition 1 converges to one wherever the surplus is positive and to zero elsewhere, and consideration arrives at a single rate that responds neither to the dollar gain nor to income. Both θ and ω drop out with it. The main text rejects this restriction because the elasticity is the parameter that generates the income gradient.

The second restriction lets the price of attention grow without bound. As $\theta \to \infty$ the same intensity converges to zero in every state, for every value of ε and ω, and consideration arrives only through the free rate $\lambda_{free} = \bar{\lambda}$, constant across households and independent of the gain. The two restrictions imply different levels. The second leaves $\bar{\lambda}$ free, so the benchmark can be matched to an average refinancing rate, while the first fixes the rate at one.

What remains under either restriction is the execution technology, and its two cost parameters select the two benchmarks. Setting $\Phi = \kappa = 0$ and shutting cash-out makes execution costless, so

an opportunity is taken exactly when the coupon exceeds the market rate, and the monthly hazard in (15) becomes

$$h_{\text{Calvo}}(x) = 1 - \exp\Big[-\tfrac{1}{12}\big(\nu_m + \bar{\lambda}\,\mathbf{1}\{m^* > m(r)\}\big)\Big]. \tag{45}$$

Retaining $\kappa > 0$ with $\Phi = 0$ instead reproduces the costly channel. Opportunities arrive at the constant rate $\bar{\lambda}$, the closing cost is paid at execution, and the comparison of $\widetilde{V}^{refi}$ with the value of waiting generates the endogenous threshold.

Proposition 3 (Nesting)**.** *Shut down cash-out and suppose $\bar{\chi} \geq 1$. There are two routes to the constant-arrival refinancing technologies of Berger et al. (2021). First, as $\varepsilon \to 0$, endogenous monitoring converges to unit intensity on the execution region and zero elsewhere. Second, fixing $\lambda_{free} = \bar{\lambda} > 0$ and letting $\theta \to \infty$ shuts down private monitoring and leaves an exogenous consideration rate $\bar{\lambda}$.*

Under either route, the resulting refinancing rule is equivalent to a constant-arrival model with rate $\bar{\lambda}$ (with $\bar{\lambda} = 1$ under the first route). If $\Phi = \kappa = 0$, it reduces to the Calvo benchmark (45). If $\Phi = 0$ and $\kappa > 0$, it reduces to the costly menu-cost specification with closing cost κ.

Proof. By Proposition 1,

$$\chi^*(x) = \min\left\{\left(\frac{G^{\$}(x)}{\theta_s}\right)^{\varepsilon}, \bar{\chi}\right\}.$$

Hence, for $\bar{\chi} \geq 1$,

$$\lim_{\varepsilon \to 0} \chi^*(x) = \mathbf{1}\{G^{\$}(x) > 0\},$$

while, for any $\varepsilon > 0$ and bounded $G^{\$}$,

$$\lim_{\theta \to \infty} \chi^*(x) = 0.$$

The first limit therefore gives unit refinancing intensity on the execution region. The second, with $\lambda_{free} = \bar{\lambda}$, gives refinancing intensity

$$\lambda_R(x) = \bar{\lambda} D(x).$$

Thus both limits are behaviorally equivalent to a constant-arrival technology, with $\bar{\lambda} = 1$ in the first case.

With no cash-out and $\Phi = \kappa = 0$,

$$D(x) = \mathbf{1}\{m^* > m(r)\},$$

which yields (45). With $\Phi = 0$ and $\kappa > 0$, $D(x)$ instead selects the states in which the refinancing value exceeds the value of waiting after paying κ, yielding the costly menu-cost specification. □

Corollary 1. *For every $\bar{\lambda}$, the Calvo hazard (45) is flat in both income and the size of the rate gap among in-the-money borrowers. Conditional on the rate gap, every household has the same refinancing hazard; conditional on being in the money, the hazard depends only on the sign of the gap and not on its magnitude or on mortgage balance.*

Both implications conflict with the empirical patterns in Section 3: refinancing hazards rise with income within rate-gap bins and with the rate gap within income quintiles. No choice of $\bar{\lambda}$ can generate either gradient in the Calvo benchmark.

State-dependent limit. At the opposite extreme, as $\varepsilon \to \infty$, the smooth attention choice converges to a state-dependent threshold rule. Monitoring vanishes below the attention threshold and jumps to full consideration above it. As $\theta \to 0$, this attention threshold disappears, leaving only the execution threshold. This limiting case corresponds to the hassle-cost-only specification considered in the main text.

I IRF Construction

This appendix describes how I construct the impulse responses reported in the quantitative section. The calculation compares two transition paths within a fixed model environment, an unshocked path and a path with a 150 basis point decline in the policy rate at time zero, one step of the model's rate grid. Let j index the model environment, where $j = B$ denotes the baseline calibration and $j = CF$ denotes a counterfactual environment such as the notification economy. Let $z \in \{0, 1\}$ index the interest-rate path, with $z = 0$ denoting the unshocked path and $z = 1$ denoting the rate-cut path.

The state vector is

$$S \equiv (W, F, m^*, s),$$

where W is liquid wealth, F is the outstanding mortgage balance, m^* is the coupon on the current mortgage, and s is the income state. The initial cross-sectional distribution is denoted by $\mu_0(S)$. In all policy experiments, the shocked and unshocked economies start from the same initial distribution μ_0. The short rate in the unshocked path is r_t^0, while the shocked path begins from $r_0^1 = r_0^0 - 0.015$ and then follows the same law of motion for interest rates. The market mortgage rate is the model-implied function $m(r_t^z)$.

For each model environment j and interest-rate path z, the cross-sectional distribution evolves according to the Kolmogorov forward equation

$$\frac{\partial \mu_t^{j,z}(S)}{\partial t} = \mathcal{L}_{j,z,t}^* \mu_t^{j,z}(S), \qquad \mu_0^{j,z}(S) = \mu_0(S),$$

where $\mathcal{L}_{j,z,t}^*$ is the adjoint of the household-state generator. This generator incorporates saving behavior, income transitions, mortgage amortization, exogenous moving, and endogenous refinancing. Endogenous refinancing occurs with intensity

$$\lambda_{j,z,t}^{refi}(S) = \chi_j^*\left(m^* - m(r_t^z), s\right) D_j\left(r_t^z, S\right),$$

where $\chi_j^*(\cdot)$ is the monitoring intensity of (14) and $D_j(r_t^z, S)$ is the execution indicator of (12). Thus, the interest-rate shock affects the transition path by changing rate gaps and the household policy functions evaluated along the rate path. When refinancing occurs, the coupon is reset to the prevailing market mortgage rate $m(r_t^z)$, and households may also adjust mortgage balances through the cash-out choice.

Aggregate consumption in environment j under path z is

$$\bar{C}^{j,z}(t) = \int C_j(r_t^z, S)\, \mu_t^{j,z}(S)\, dS,$$

where $C_j(r_t^z, S)$ is the household consumption policy. The consumption impulse response for environment j is the percentage deviation of aggregate consumption in the shocked path from the unshocked path:

$$IRF_C^j(t) = \frac{\bar{C}^{j,1}(t)}{\bar{C}^{j,0}(t)} - 1.$$

The figures report $100 \times IRF_C^j(t)$.

I summarize the dynamic response using cumulative impulse responses. For horizon T, the undiscounted cumulative response is

$$CIRF_C^j(T) = \int_0^T IRF_C^j(t)\, dt,$$

and the discounted cumulative response is

$$CIRF_{C,\delta}^j(T) = \int_0^T e^{-\delta t} IRF_C^j(t)\, dt,$$

where I set $\delta = 0.14$ in the baseline reporting. Because $IRF_C^j(t)$ is measured as a proportional deviation, $100 \times CIRF_C^j(T)$ is reported in percentage-point-years. The average response over the horizon is $CIRF_C^j(T)/T$.

For counterfactual comparisons, I compare the cumulative response in the counterfactual economy to the cumulative response in the baseline economy:

$$\Delta CIRF^{CF}_{C,\delta}(T) = CIRF^{CF}_{C,\delta}(T) - CIRF^{B}_{C,\delta}(T),$$

and report the amplification relative to the baseline response as

$$Amplification(T) = \frac{CIRF^{CF}_{C,\delta}(T) - CIRF^{B}_{C,\delta}(T)}{CIRF^{B}_{C,\delta}(T)}.$$

In the exercises, this statistic is evaluated at $T = 5$ years.

J External Evidence on Attention across Income

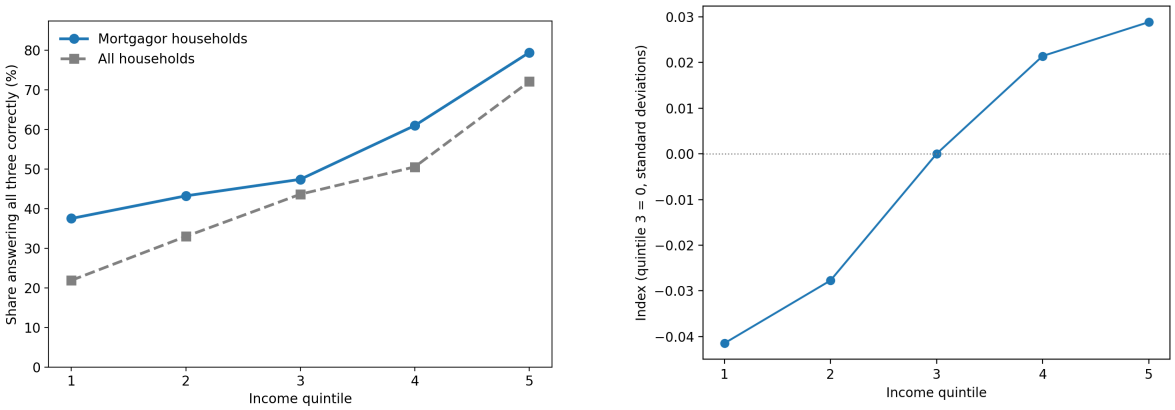
This appendix describes the two external series referenced in Section 5.1. Neither is a calibration target.

Panel (a) of Figure J.1 uses the 2022 SCF. It reports the weighted share of households answering all three standard financial-literacy questions, on compound interest, inflation, and diversification, correctly. "Don't know" responses are coded as incorrect. The series is shown separately for mortgagor households and for all households.

Panel (b) uses the 2013–2023 NSMO public use file. The index averages two components of lender search at origination. The first is the number of lenders the borrower seriously considered before applying (X11). The second is the number applied to (X12), which enters only for borrowers who report that they applied to more than one lender to search for better loan terms (X13A). Each component is rescaled from its one-to-five response range to the unit interval and converted to a weighted z-score before averaging, so the index is measured in standard deviations. The sample is restricted to respondents who report that obtaining a low interest rate was important (X37B), and an observation is retained when at least 60 percent of its components are non-missing.

NSMO reports income in six bins. I map them to the model's five income states by fitting a weighted cubic B-spline to binned mean log income and evaluating it at the model's income states.

The lowest state pools the two lowest NSMO bins, which enter with weights 0.33 and 0.67. The index is reported relative to the third state, and its spread across income states is 0.07 standard deviations.



(a) Financial Literacy by Income (SCF 2022) (b) Mortgage-Search Index by Income Quintile

Figure J.1: External Evidence on Attention across the Income Distribution

Notes: Panel (a) reports the weighted share of households answering all three standard financial-literacy questions correctly, by income quintile. Panel (b) reports the mortgage-search index in standard deviations relative to the third income state. Sources: 2022 SCF and NSMO.

K Robustness for Model Counterfactuals

Table K.1 examines whether the effect of the notification policy depends on the model's MPC gradient. Rather than imposing MPCs externally, I vary the income profile of the discount rate, $\delta_s = \delta\,(z_s/\bar{z})^{\psi}$, and re-solve the model at the baseline friction estimates and the converged arrival rate $\lambda_{free} = 0.263$. This preserves the endogenous relationship between liquidity and MPC within each income group. Setting $\psi = -0.25$ makes lower-income households more impatient and steepens the MPC profile, while $\psi = +0.25$ flattens it. Across these specifications, the bottom-quintile MPC ranges from 0.208 to 0.251 and the Q1–Q5 gap from 0.086 to 0.133, compared with 0.108 in the baseline.

The effect of the notification policy changes little across these economies. The five-year consumption gain ranges from 12.48 to 12.49 percent of the corresponding baseline response, compared with 12.35 percent in the baseline calibration. The dollar gain per household is similarly stable,

ranging from $364 to $367, and the bottom quintile's own response rises by 21.5–21.6 percent, compared with 20.4 percent in the baseline. The additional consumption response to the notification policy is therefore not sensitive to substantial variation in the model-generated MPC profile.

Table K.1: Notification Counterfactual under Alternative MPC Profiles

Scenario	MPC Q1	MPC Q5	MPC gap (Q1−Q5)	Notification gain (%)	Gain ($/hh)	Baseline resp. ($/hh)	Q1 gain (%)
Flatter gradient ($\psi = +0.25$)	0.208	0.122	0.086	12.5	364	2,916	21.6
Baseline ($\psi = 0$)	0.228	0.120	0.108	12.3	361	2,927	20.4
Steeper gradient ($\psi = -0.25$)	0.251	0.117	0.133	12.5	367	2,941	21.5

Notes: Each row is a fully re-solved economy in which the discount rate is tilted by income, $\delta_s = \delta\,(z_s/\bar{z})^{\psi}$, so the income gradient in MPCs moves endogenously while liquidity, balance, and attention heterogeneity remain model objects. Friction parameters are held at their baseline estimates and the notification arrival rate at $\lambda_{free} = 0.263$ per year. MPCs are exposure-weighted quintile means from each economy's simulated cross-section. The notification gain is the five-year discounted cumulative consumption response of the notification counterfactual relative to that economy's own baseline response to the 150 basis point rate-cut experiment; the last column is the bottom quintile's percentage gain. Source: model simulations.