AEA Papers and Proceedings 2023, 113: 70–75
https://doi.org/10.1257/pandp.20231027

The Trickling Up of Excess Savings[†]

By Adrien Auclert, Matthew Rognlie, and Ludwig Straub[*]

In the wake of the COVID pandemic, households accumulated a very large stock of "excess savings," which they have only recently begun to deplete. Figure 1, panel A shows that the US personal savings rate first rose very rapidly in 2020, more than doubling relative to its long-term average, then started falling below that average in late 2021. Figure 1, panel B shows an estimate of the resulting stock of excess savings by the Federal Reserve Board (Aladangady et al. 2022). This stock has only modestly fallen from its peak. In mid-2022, it still stood at $1.7 trillion, or 6.7 percent of GDP.

Because excess savings and their distribution across the population intuitively matter for aggregate demand, economists have paid a considerable amount of attention to estimating both. In this paper, we provide a tractable heterogeneous agent New Keynesian model that explicitly maps the distribution of excess savings to the path of output, and that explains the process by which their effect dissipates. We use this framework to estimate the likely contribution of excess savings to aggregate spending in the coming years under various assumptions about the marginal propensities to consume (MPCs) of agents holding the savings and scenarios for monetary policy.

Our framework recognizes that one person's spending is another person's income. As we show, taking this fact into account implies that excess savings from debt-financed transfers have much longer-lasting effects than a naive calculation would suggest. In a closed economy, unless the government pays down the debt used to finance the transfers, excess savings do not go away as households spend them down. Instead, the effect of excess savings on aggregate demand slowly dissipates as they "trickle up" the wealth distribution to agents with lower MPCs. Tight monetary policy speeds up this process, but this effect is likely to be quantitatively modest.

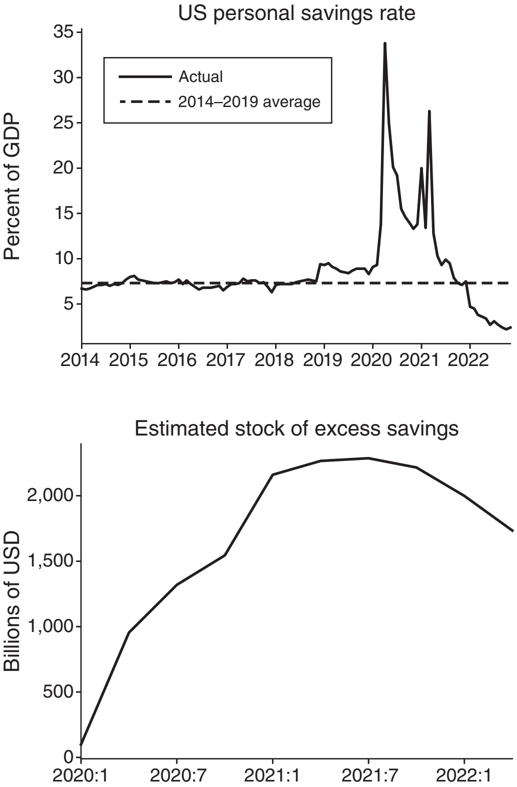

Figure 1. US Personal Savings Rate and Excess Savings

Sources: The personal savings rate is from the Bureau of Economic Analysis (FRED code: PSAVERT). The estimated stock of excess savings is from Aladangady et al. (2022).

I. Model

We consider a continuous time model with N types of households, $i = 1, \ldots N$.

[*] Auclert: Stanford University and NBER (email: aauclert@stanford.edu); Rognlie: Northwestern University and NBER (email: matthew.rognlie@northwestern.edu); Straub: Harvard University and NBER (email: ludwigstraub@fas.harvard.edu). We thank Şebnem Kalemli-Özcan and our discussant Fabrizio Perri for helpful comments.

[†] Go to https://doi.org/10.1257/pandp.20231027 to visit the article page for additional materials and author disclosure statement(s).

Agents with higher i have lower instantaneous MPCs m_i, with agent N having an MPC of 0, $m_1 > m_2 > \cdots > m_N = 0$. Motivated by the empirical evidence on the negative correlation between MPCs and wealth, we think of agents with higher i as being initially richer, with agent N being the richest. While this is a useful interpretation, it is not strictly necessary: what is important is the distribution of m_i across types i.

At $t = 0$, the government distributes a transfer a_{i0} to households, issuing debt $B = \sum_{i=1}^{N} a_{i0}$ to finance the transfer and maintaining a constant debt level thereafter. We first consider an "easy monetary policy" scenario in which the central bank responds by holding constant the real interest rate at its steady-state level of 0, $r = 0$. This implies, in particular, that the additional debt requires no change in taxes.

Each type's behavior is described by a utility function over consumption and assets. Agents understand the central bank's announcements of future real interest rates r_t (here, $r = 0$), but they assume that future aggregate income Y_t remains permanently at its steady-state level.[1] Agent type i earns a fixed proportion $\theta_i \in (0,1)$ of total income Y_t.

We linearize this model around the steady state where each agent type owns a certain stock of assets (with higher-type agents plausibly holding more wealth). This delivers the following equations:

$$(1)\quad c_{it} = m_i a_{it}, \quad \dot{a}_{it} = \theta_i Y_t - c_{it}, \quad Y_t = \sum_{i=1}^{N} c_{it},$$

where Y_t is aggregate demand and income, c_{it} is type i's consumption, a_{it} his asset holdings (all relative to their steady-state level), and $m_i \in [0,\infty)$ his instantaneous MPC out of liquid assets. The θ's, which satisfy $\sum_{i=1}^{N}\theta_i = 1$, are the income shares across the types. The equations in (1) give a tractable version of the intertemporal Keynesian cross (Auclert, Rognlie, and Straub 2018).

An unconventional feature of our model is that it assumes the presence of agents with zero MPC. One can interpret these type N agents as standard permanent-income agents, in the limit where their discount rate goes to zero, but alternative interpretations are possible. First, they could stand in for the rest of the world. Second, they could represent the government receiving a fraction of aggregate income via taxation and using it to pay down the debt. Finally, they could represent zero-MPC financial accounts, such as retained earnings saved by firms or pension funds.

One natural objection to the model in (1) is that it assumes that monetary policy maintains an easy stance of $r = 0$ in the face of high demand. To address this, we extend our model by assuming that monetary policy tightens as it sees higher demand, reacting with a rule $r_t = \phi Y_t$.[2] Since higher demand will naturally be associated with higher inflation, an alternative interpretation of this rule is that monetary policy tightens in reaction to the inflation generated by excess savings. Online Appendix A derives the equations characterizing the model in this case.

Another objection to the model in (1) is that it relies on imperfect foresight by agents. Online Appendix A also derives the equations characterizing the model when agents have rational expectations about interest rates r_t as well as income Y_t.

Partial Equilibrium Analysis.—A naive partial equilibrium approach to calculating the effect of excess savings on spending would be to ignore the endogeneity of output, instead assuming that Y_t remains at its normalized steady-state level of 0 forever. Solving out for (1) in this case, we find that aggregate demand is given by

$$(2)\qquad C_t = \sum_{i=1}^{N-1} m_i e^{-m_i t} a_{i0}.$$

Equation (2) delivers a simple way to map a distribution of MPCs and excess savings by type into an effect on aggregate spending: take type i's initial stock of savings, and apply to it an exponential distribution for spending with

[1] In the words of Farhi and Werning (2019), agents have level k thinking, with $k = 1$. This makes the model particularly tractable. We later consider the case with rational expectations.

[2] To neutralize the income effects of changing interest rates, in this extension we assume that all agents types start with a steady-state level of wealth of 0. We think of this as proxying for the presence of long duration assets, which hedge agents against interest rate risk.

TABLE 1—DURATION OF OUTPUT AND EXCESS SAVINGS BY TYPE UNDER ALTERNATIVE SCENARIOS (IN QUARTERS)

Scenario	Duration of output and excess savings		
	Output Y	Middle-class a_1	Rich a_2
Partial equilibrium	3	2	4
Benchmark	20	19	22
Lower MPCs ($mpc_1 = 0.3, mpc_2 = 0.1$)	38	34	43
More excess savings to rich ($a_{10} = a_{20} = 0.45B$)	21	20	22
More earnings to rich ($\theta_1 = 0.3, \theta_2 = 0.55$)	23	19	26
Rational expectations	8	6	10
Tight monetary policy ($\phi = 1.5$)	8	7	11

Notes: The time unit is a quarter. Given that $r = 0$, the duration of a variable X_t is defined as $\int tX_t dt / \int X_t dt$. Our benchmark calibration has $mpc_1 = 0.4$, $mpc_2 = 0.2$, with $m_i = -\log(1 - mpc_i)$; income shares $\theta_1 = 0.47, \theta_2 = 0.38$, $\theta_3 = 0.15$; and initial assets $a_{10} = 0.6 \cdot B, a_{20} = 0.3 \cdot B$, with $B = 6.7\%$ of GDP. For the monetary response scenario, we assume that agents have an elasticity of intertemporal substitution of 1/2.

mean $1/m_i$.[3] A simple back-of-the-envelope calculation using this equation suggests that for the United States, the remaining excess savings might only affect aggregate demand for a few quarters (see Table 1).

This approach, however, fails to recognize that one agent's spending is another agent's income. Ignoring this fact has important consequences: if agents simply spent down their excess savings without raising anyone else's income, then no one would be purchasing the assets they sold in the process. But this is inconsistent with the government keeping its debt constant. As we show next, recognizing this fact implies a much greater persistence of excess savings and output than equation (2) suggests.

II. The Trickling-Up Effect

We now explicitly solve the dynamical system in (1). We begin with a simple observation about the steady state of this system.

PROPOSITION 1 (Long-Run Trickling Up): *In the long run, type N owns all the debt:* $\lim_{t\to\infty} a_{Nt} = B$.

This result follows immediately from the fact that type N has $m_N = 0$, so that its asset dynamics are given by $\dot{a}_{Nt} = \theta_N \left(\sum_{i=1}^{N-1} m_i a_{it}\right)$. Hence, as long as other agents have excess savings, they spend them down, increasing the income and therefore the savings of the richest type. Since the government keeps its debt position constant ($\sum_{i=1}^{N} a_{it} = B$ at all times), in the long run all types have zero assets except for type N, which owns all of B. At this point, excess savings have "trickled up" to agents with the highest i. Given our interpretation of type N as being initially the richest agent, we see that any initial transfer, no matter how targeted it is to the poor, eventually ends up raising wealth inequality.

PROPOSITION 2 (Trickling-Up Dynamics): *Assume that $m_i a_{i0}/\theta_i$ decreases in i. Then the distribution of assets across types i at any later date t′ first-order stochastically dominates the distribution at any earlier date $t < t'$: $\sum_{i=1}^{n} a_{it'} < \sum_{i=1}^{n} a_{it}$ for all $n < N$.*

This result, proved in online Appendix B, shows the exact sense in which excess savings trickle up: no matter where we look in the distribution of excess savings, as time passes, the wealth held by all lower types is falling, and the wealth held by all higher types is rising. The only necessary condition is that excess savings initially cause a larger percentage increase in spending among poorer agents, which is easily satisfied since they have higher MPCs.

PROPOSITION 3 (Slow Dissipation): *In the long run, $Y_t \sim e^{-\lambda t}$: aggregate demand and excess savings dissipate at rate λ, where $\lambda < m_{N-1}$. Hence, excess savings have a strictly longer-lasting effect on demand than the naive partial equilibrium calculation in* (2) *would suggest.*

[3] This functional form characterizes the intertemporal MPCs of agents with assets in the utility; once multiple types of such agents are mixed together, the model's aggregate dynamics are similar to those of alternative heterogeneous agent models. See Auclert, Rognlie, and Straub (2018).

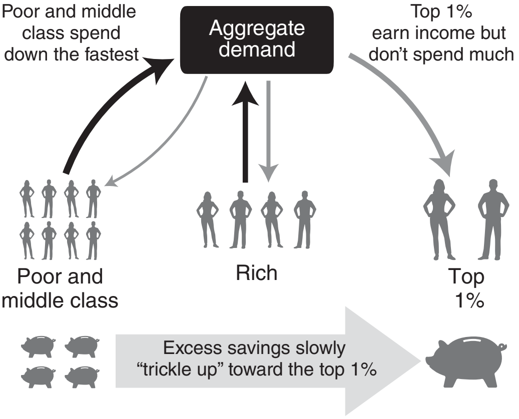
Poor and middle class spend down the fastest
Aggregate demand
Top 1% earn income but don't spend much
Poor and middle class
Rich
Top 1%
Excess savings slowly "trickle up" toward the top 1%

Figure 2. The Trickling-Up Effect

In the partial equilibrium calculation from equation (2), spending eventually becomes dominated by type $N-1$ agents, decaying at rate m_{N-1}. Proposition 3 shows that general equilibrium spending dissipates strictly more slowly than this. Intuitively, this is because the spending from any type sustains income from any other type as the wealth of all agents goes to zero.

Figure 2 illustrates the adjustment process characterized by Propositions 1–3. Dark arrows flow from agent types to aggregate demand Y_t via their spending (m_i), with lower types spending down their assets faster. Gray arrows flow from aggregate demand to the income of these agents, and they are more equally distributed across the population (θ_i), with type N agents receiving a significant share. Running this system forward, we see that excess savings slowly trickle up the wealth distribution, until type N agents own all of the assets.

Three-Type Example.—We now specialize the model to a case with $N = 3$ types. This case is simple to analyze graphically, and provides additional analytical insights into the trickling up of excess savings. We think of type 1 as representing the poor and the middle class, type 2 as representing the rich, and type 3 as representing the superrich. Manipulating the equations in (1), we see that the dynamics of excess savings for the first two types satisfy

$$\begin{pmatrix}\dot{a}_1\\ \dot{a}_2\end{pmatrix} = \begin{pmatrix}-m_1(1-\theta_1) & \theta_1 m_2\\ \theta_2 m_1 & -m_2(1-\theta_2)\end{pmatrix}\begin{pmatrix}a_1\\ a_2\end{pmatrix}.$$

Once we have solved for (a_{1t}, a_{2t}), it is easy to back out $a_{3t} = B - (a_{1t} + a_{2t})$.

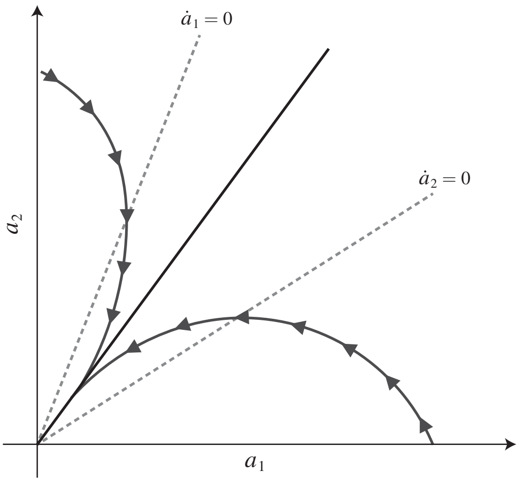

Figure 3. Phase Diagram for $N = 3$ Case

Figure 3 visualizes this dynamical system using a phase diagram for (a_1, a_2). The locus for $\dot{a}_1 = 0$ is given by $a_2 = \frac{\theta_2+\theta_3}{1-(\theta_2+\theta_3)}\frac{m_1}{m_2}a_1$; to the right of this locus, the assets of type 1 agents decline. The locus for $\dot{a}_2 = 0$ is flatter, at $a_2 = \frac{\theta_2}{1-\theta_2}\frac{m_1}{m_2}a_1$; to the right of this locus, type 2 assets increase. The dynamics of the wealth distribution are then given by the arrows on the graph, splitting the positive quadrant into three regions: two regions close to the axes in which agents' assets move in opposite directions, and a middle cone in which both agents' assets decline together. In the scenario where initial a_2 is low relative to a_1, type 2 agents initially increase their assets, as the spending by type 1 agents initially boosts their incomes and savings before reaching a second phase in which both types' assets decline as the superrich accumulate. We formalize this situation in the following proposition:

PROPOSITION 4: *Assume that type* 1 *agents initially own a sufficiently large share of assets,* $\theta_2 m_1 a_{10} > (1-\theta_2) m_2 a_{20}$. *Then, type* 2 *agents first accumulate assets before spending them down.*

The hump-shaped response of savings of type 2 agents is a simple manifestation of the trickling-up effect from Proposition 2.

III. Application to the United States

We use our model to quantify the likely impact of the stock of excess savings estimated by Aladangady et al. (2022) on aggregate demand and its likely duration. We follow the three-type classification outlined in Section II. We set the time units so that $t = 1$ corresponds to a quarter. The parameters of the model are θ_i, m_i, and a_{i0} for each i.

We interpret types as follows: type 1 is the bottom 80 percent of the US wealth distribution, type 2 is the next 19 percent, and type 3 is the top 1 percent. In the 2019 Survey of Consumer Finances, the bottom 80 percent of the US wealth distribution earns 47 percent of income, the next 19 percent earns 38 percent, and the top 1 percent earns 15 percent. We assume that marginal income is distributed like average income; this implies our θ_i's. Next, we assume realistically high quarterly MPCs for the middle class and the rich: $mpc_1 = 0.4$ and $mpc_2 = 0.2$, respectively. We then convert these numbers to instantaneous MPCs using the formula $1 - e^{-m_i} = mpc_i$. Finally, we assume that the excess savings have only started to trickle up the wealth distribution, with the middle class owning 60 percent and the rich owning 30 percent of the stock of excess savings. Finally, we take the total stock to be $B = 6.7\%$ of GDP, as estimated by Aladangady et al. (2022). While the exact numbers entering our calculations are highly uncertain, Table 1 shows that our results are robust to reasonable alternative calibrations.

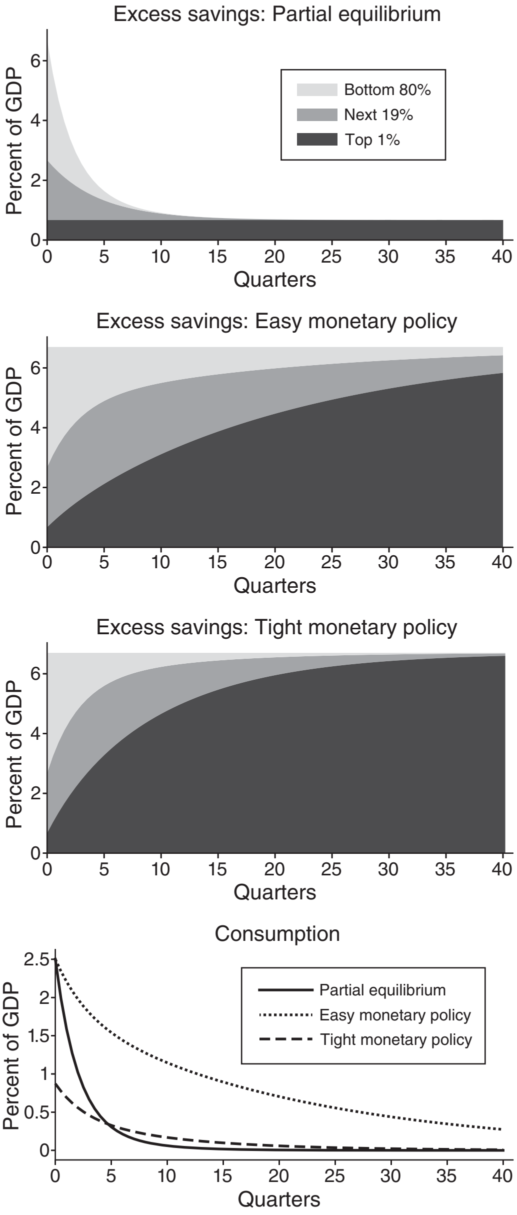

Figure 4. Dynamic Evolution of the Distribution of Excess Savings and Consumption

Figure 4 reports the evolution of the distribution of savings across types in three alternative scenarios. The top panel shows the outcome of a partial equilibrium analysis: all types except the first quickly run down their excess savings, and after a few years, only 10 percent of the US debt is held by superrich US residents. The second panel from the top shows our general equilibrium benchmark instead, in which the debt is continuously held domestically.[4] This visualizes the trickling-up phenomenon: the share of wealth held by the rich initially rises (the parametric restriction for a hump shape is satisfied), and the superrich keep accumulating assets until they hold all of the excess savings. The third panel from the top shows what happens under a tight monetary policy scenario, with $\phi = 1.5$. The qualitative trickling-up patterns are unchanged, but the monetary response does speed up the adjustment process. The bottom

[4] Aggarwal et al. (2022) consider an intermediate case where the United States is a partially open economy. With home bias in spending, the outcome is similar to our closed-economy simulations, except that we can interpret the top 1 percent as the foreigners.

panel summarizes the effect of excess savings on aggregate consumption. These effects are long-lasting and significant. In addition to speeding up the adjustment, the monetary response brings down the level of demand.

Table 1 summarizes our results by displaying the duration of output and excess savings for the middle class and the rich under each of our scenarios. The partial equilibrium scenario summarizes the conventional wisdom, according to which the effect of excess savings will dissipate in a few quarters. By contrast, our benchmark scenario suggests that these effects will stick around for roughly five years. These numbers are larger if MPCs are lower, and they are robust to plausible alternative calibrations. Rational expectations about the future boom make the response much larger on impact due to current spending out of anticipated income, which turns out to speed up the trickling-up process. Tight monetary policy, on the other hand, also speeds up trickling up, but it does so by mitigating the effects of excess savings on demand. In either case, however, the duration of excess savings and output remains more than twice as long as the conventional wisdom suggests.

REFERENCES

Aggarwal, Rishabh, Adrien Auclert, Matthew Rognlie, and Ludwig Straub. 2022. “Excess Savings and Twin Deficits: The Transmission of Fiscal Stimulus in Open Economies.” In *NBER Macroeconomics Annual 2022*, Vol. 37, edited by Martin Eichenbaum, Erik Hurst, and Valerie A. Ramey.

Aladangady, Aditya, David Cho, Laura Feiveson, and Eugenio Pinto. 2022. “Excess Savings during the COVID-19 Pandemic.” *FEDS Notes*, Board of Governors of the Federal Reserve System, October 21. https://doi.org/10.17016/2380-7172.3223.

Auclert, Adrien, Matthew Rognlie, and Ludwig Straub. 2018. “The Intertemporal Keynesian Cross.” NBER Working Paper 25020.

Farhi, Emmanuel, and Iván Werning. 2019. “Monetary Policy, Bounded Rationality, and Incomplete Markets.” *American Economic Review* 109 (11): 3887–928.