Consumer Payment Choice and the Heterogeneous Impact of India’s Demonetization

Abstract

Consumer payment choice is based on heterogeneous preferences, availability, usage costs, and effective taxes. We examine the consequences of this choice on consumption distribution, aggregate output, welfare and the shadow economy. We employ this framework to analyze India’s unexpected demonetization of 86% of its currency in circulation. We find that this shock to liquidity led to an immediate and temporary fall in aggregate output and welfare. Further, it led to disparate distributional effects as not all consumers could switch to non-cash payments. Using consumption distribution data, we find the lower and middle deciles in rural areas were disproportionately affected.

JEL Codes: D83, E41, E52, E58, O17

*We thank Guillaume Rocheteau, Randall Wright, Pedro Gomis-Porqueraz, Pushkar Maitra, Dilip Mookherjee, Almuth Scholl, Anirudh Tagat, Nikita Naik and discussants Raghbendra Jha and Makoto Watanabe for helpful discussions on earlier versions. We also thank participants at the Bank of Finland; Mid-west Macro Workshop, Wisconsin; 5th HenU/INFER Macro Workshop, China; Deakin University, Monash University, University of Queensland; Summer School in Development Economics, Italy; Australasian Eco-nomic Theory Workshop; Ashoka University, Delhi School of Economics and IIM-Ranchi for useful comments and suggestions. Any errors that remain are ours. Ayushi Bajaj: ayushi.bajaj@monash.edu, Nikhil Damodaran: ndamodaran@jgu.edu.in.
1 Introduction

Consumer payment methods worldwide have undergone fundamental changes over time, with the most recent being a shift away from cash, to other electronic or digital means, such as debit cards and digital wallets. These different instruments involve trade-offs that influence their selection as means of payments. For example, carrying several small bills of cash is cumbersome while larger bills are more susceptible to counterfeiting. Digital payments are usually devoid of such costs but using them requires a bank account in the least, which involves some fees or minimum balance requirements. Besides, infrastructural constraints lead to differential access, especially in developing countries. These factors can potentially create a payments divide based on rates of adoption and access. In addition, anonymous cash payments facilitate tax evasion, and the emergence of a parallel shadow economy. Whereas a shift towards digital payment methods creates difficulties in tax evasion, because such transactions are easier to track.

In this paper, we model these features of payment instruments using a tractable monetary framework based on Lagos and Wright (2005) and Rocheteau and Wright (2005). We also include preference heterogeneity and taxation to characterize equilibrium regimes based on consumer’s choice of means of payments. We analyze the consequences of this choice on outcomes such as consumption distribution, aggregate output, welfare and the size of the shadow economy. If consumers transact in cash, they economize on their money holdings because they face a higher marginal carrying cost. This leads to lower aggregate output and a larger shadow economy. However, the alternative of making digital payments is limited by consumption size, effective taxation and infrastructural constraints. This leads to a payments divide, which affects aggregate welfare and inequality especially following a payments system shock.

We apply this framework to analyze the heterogeneous impact of such a monetary episode in India. On November 8, 2016 the Government of India unexpectedly demonetized the two largest denomination bills comprising of 86% of the existing currency in circulation, effective at midnight. Replacement of the demonetized currency with new notes took time and effort, imposing a significant strain on the payments system for almost three quarters. Demonetization occurred in an otherwise stable macroeconomic environment and led to an immediate fall in aggregate output and welfare, as consumers carrying cash found their bills...
to be no longer accepted for transactions. We also show how a uniform aggregate shock across the economy has differential effects on consumers given their choice of payment methods.

Our framework has three key features that makes it particularly amenable to analyze this monetary episode. First, we model money as a means of payment with explicit micro-foundations i.e. money helps alleviate limited commitment and lack of double coincidence of wants. We employ a tractable environment where money is essential i.e. its presence makes superior allocations possible. Money can be held in two forms. Cash is available in two denominations which involves a trade-off between lower carrying cost and higher probability of counterfeiting. To keep things simple, the cost of counterfeiting is chosen such that agents are indifferent between holding cash in either denomination. The non-cash form of payment is like a digital wallet or a debit card, which requires a fixed cost to setup, independent of transaction size. For instance, consumers need a bank account with fixed operational fees to access these instruments. We assume that this cost falls exclusively on the consumers, and the firm can accept both forms of money costlessly. This implies the existence of a unique equilibrium as there is no strategic complementarity associated with this choice.

The second key feature of our model is heterogeneity on consumer preferences. This enables us to examine their relative cash dependence and analyze the differential impact of a uniform demonetization shock. The disparate effects of the shock is further intensified in the presence of a payments divide as not all consumers could make the switch to non-cash means of payments. Third, we assume that government levies a sales tax, but it cannot be perfectly enforced. This leads to the emergence of a tax evading shadow economy. We model the tax enforcement authority’s efforts to prevent tax evasion and find that the effective tax rate on cash transactions is lower. This is because smaller transactions are often done in cash, so the reward from catching them is lower. These features of the model make it well-suited to analyze this monetary episode, as it explicitly addresses why cash matters and what determines its choice over non-cash payments.

To quantify the impact of demonetization, we calibrate the model parameters to match key features of the Indian economy. We match the relationship between M1/GDP and 91-day Treasury bill rate to obtain our money demand function. We also set the preference heterogeneity parameters for rural and urban consumption deciles from the National Consumer Expenditure Survey (CES). We study the impact of households suddenly finding their high denomination bills to be unacceptable for transactions and find that aggregate output
fell by at least 20%.

We also quantify the impact of the slow and costly remonetization that followed over the next three quarters by adjusting the cost of carrying cash. The effects of slow remonetization are disproportionate and largely focused on the regions and groups that were not able to transition to digital means of payment, which included the rural regions and the lower and middle consumption deciles. Urban households in the upper deciles could switch to non-cash payments more easily because they could bear the associated costs by taking a reduction in their surpluses. We find that in the first quarter, urban output dropped by 3.5% while rural output would have reduced by twice as much. Aggregate welfare which accounts for output, tax revenue and costs of payments fell by 7.7%. The shadow economy would also fall on account of higher digital payments usage on which tax evasion is harder. But, in comparison with the fall in total and especially unorganized sector output which featured a large and persistent negative impact of the slow remonetization, the effect on the shadow economy was of a smaller magnitude and was less persistent.

Related Literature

For the monetary framework this paper builds on Lagos and Wright (2005) and Rocheteau and Wright (2005). The denomination structure is modeled as a simplified version of Lee et al. (2005) and Nosal and Rocheteau (2011). Some other papers using similar monetary frameworks with two means of payments include Li (2011) that finds that checks are used only in big transactions while cash is used in all transactions and monetary policy has differential impacts on the terms of trade in transactions using different means of payment. Kim and Lee (2010) presents a model of debit card where sellers bear a fixed record-keeping cost regardless of transaction size. Lotz and Vasselin (2019) develops a dual payments model with electronic money and cash, and finds that strategic complementarities lead to multiple monetary equilibria as the cost of accepting e-money borne by merchants. Zhu and Hendry (2019) and Williamson (2019) present models with multiple means of payment to analyze the effects of introducing central bank digital currency.

This paper also builds on the theoretical literature on shadow economies, while drawing important insights from its empirical counterpart. A paper working on a similar class of models where a shadow economy arises endogenously is Gomis-Porqueras et al. (2014) which compares cash and trade credit as the two means of payment, with cash being subject to an inflation tax and all credit transactions subject to sales tax. In this paper, since both
means of payment are issued as money, both are subject to inflation. Moreover, we solve
the tax authority’s enforcement problem endogenously on cash and non-cash transactions.
Di Nola et al. (2018) finds that income tax evasion leads to a larger self employment sector
but reduces their productivity by calibrating their heterogeneous agent, incomplete markets
model to US data. Other papers on shadow economy includes Koreshkova (2006) which
again focuses on inflation as tax on underground economy, Camera (2001) takes a search-
theoretic approach and Schneider and Enste (2000) provides a summary. Rogoff (2017) also
offers useful insights and discussion, highlighting cases of tax evasion to make the case for
phasing out large denomination bills.

Lahiri (2020) provides a summary analysis of India’s demonetization and some related lit-
erature. Chodorow-Reich et al. (2020) presents a model of demonetization where agents hold
cash to satisfy a cash-in-advance constraint and for tax evasion. They use cross-sectional
data on distribution of new notes for causal inference and find that districts experiencing
more severe demonetization had relative reductions in economic activity, faster adoption of
alternative payment technologies, and lower bank credit growth. For our quantitative exer-
cise we do not vary the rate of redistribution of new notes, but we assume different capacities
on the part of household to use and switch to alternative payment technologies based on con-
sumption distribution data. Wadhwa (2019) also looks at the effects on consumption through
an empirical exercise using consumer pyramids data. The key finding is that demonetization
led to a higher decline in consumption for richer than poorer households as the rich could
afford to reduce their consumption temporarily due to their high initial consumption levels.
We find that among households who could not switch to non-cash payments, this is true.
But, households at the very top of the distribution could switch, so their consumption did
not fall significantly, and hence the middle deciles experienced the greatest decline in their
consumption.

Crouzet et al. (2019) focuses on technological adoption of electronic payments by retailers
and finds evidence for positive externalities in adoption which limited the costs of the shock.
We do not consider retailer’s adoption decision but in our analysis of demonetization we allow
for increased technological development which could reduce the exogenous costs of using
digital means. Karmakar and Narayanan (2019) uses a panel dataset on Indian households
and finds that the 17 percent of households that did not have bank accounts experienced 2 to
7 percent lower consumption than the control group of households with bank accounts, with
the size of the effect varying by the initial asset levels of the household. Some other papers that analyze this episode include from different angles include Waknis (2017), Agrawal (2018) and Tagat and Trivedi (2020).

The paper is organized as follows. Section 2 lays down the model environment and Section 3 characterizes the equilibrium with only cash as the means of payments and no taxation. Section 4 includes digital payments and analyzes the steady state equilibrium in the model without a shadow sector, which is finally brought in the full model in Section 5, which is then applied to India’s experience with demonetization and Section 6 does a quantitative analysis. Finally, Section 7 concludes. Proofs to all Lemmas and Propositions are in the Appendix.

2 Model

Time is discrete and continues forever. The economy is populated by households, firms and a government/central bank with a consolidated budget constraint. Households supply labor \( l \) to firms and consume two goods: a general good \( q \) consumed at the end of every period and a special good \( y \) to be consumed earlier with probability \( \alpha \). The general good serves as a numéraire. This alternating consumption sequence is based on the centralized and decentralized market structure in Lagos and Wright (2005) and Rocheteau and Wright (2005).

The lifetime discounted expected utility of a household is:

\[
\mathbb{E} \sum_{t=0}^{\infty} \beta^t [\epsilon u(y_t) + U(q_t) - l_t],
\]

(1)

where \( \beta \equiv (1 + \rho)^{-1} \in (0, 1) \) is the discount factor between periods, \( \epsilon u(y_t) \) is the utility from consumption of the special good at time, \( t \) and \( U(q_t) \) is that from the general good. The dis-utility from labor \( l_t \) is linear. This quasi-linear preference structure follows from Lagos and Wright (2005) which simplifies the analysis since it leads to a degenerate distribution of assets. We also have standard assumptions on both utilities i.e. they are twice continuously differentiable with \( u'(\cdot) > 0, u''(\cdot) < 0 \) and, \( u(0) = 0 \). Similarly, for \( U(\cdot) \).

The multiplicative term in front of \( u(y_t) \) is \( \epsilon \geq 0 \) which represents a preference shock.
parameter for special goods consumption. We allow for these preference shocks to be conditional on household types, such that each household type-\(i\)'s utility from consumption of the special good \(y_i^t\) is given by \(\epsilon^i u(y_i^t)\), where \(i \in \mathbb{I}\) set of integers. We assume that with probability \(\alpha\), \(\epsilon^i > 0\) for household-\(i\) and \(\epsilon^i = 0\) with complementary probability \(1 - \alpha\). And, the proportion of households of type-\(i\)'s is given by \(\pi^i\).

Firms operate in perfectly competitive markets for both goods and labor. They employ a production technology to obtain \(\bar{z}\) units of the general good at the end of every period. Firms can also speed up production of the special good at the beginning of every period, under a linear cost, \(c(y_t) = y_t\). If they choose to speed up the production of special good, then its output in the last stage is \(\bar{z} - y_t\). We assume that \(\bar{z} - y_t \geq 0\) along the equilibrium path. Wages \(w_t\) which along with the remaining revenue (i.e. profits), \(\Delta_t\) are paid out to households at the end of every period.

So far we have not mentioned how transactions take place and the frictions involved, if any, when households get a positive preference shock for early consumption i.e. before wages are paid. Similar to the way meetings take place in the decentralized sub-market in Rocheteau and Wright (2005) (under perfect competition), we assume that buyers are anonymous and cannot commit to repay debt in the early consumption stage. Hence, early consumption cannot be financed with debt, i.e., settlement cannot be delayed. So, there is role for a medium of exchange in this economy.

There is fiat money available as cash in two simple denominations. The low is fully divisible, and the high denomination is available in \(k > 1\) units of the numéraire. There is a central bank which controls the total money supply, and is willing to adjust the supply of each denomination as per demand. Total money supply by the central bank each period is given by \(M_t\) which grows at a constant rate, \(\Pi\) where \(1 + \Pi = M_t / M_{t-1} = \phi_{t-1} / \phi_t\) is the inflation rate and \(\phi\) is the price of money in terms of numéraire. The nominal interest on an illiquid bond \(\iota\) is given by \(\iota = (1 + \Pi)(1 + \rho) - 1\), which represents the opportunity cost of holding money. There is also a carrying cost of \(\gamma\) per bill. Besides, these bills can also be counterfeited at nominal cost \(\delta\) such that, \(1 < \delta < k\) per bill during the late consumption period.\(^1\) We will also introduce a digital or electronic means of payment, like a debit card in Section 4 which will work exactly like the cash fiat money, except that they do not incur

\(^1\)The counterfeiting structure has been greatly simplified as it is not key to the analysis as will be seen. See for example Li et al. (2012) or Quercioli and Smith (2015) for a more rigorous analysis on counterfeiting.
a carrying cost and cannot be counterfeited. But, there is a fixed usage fee for holding the
digital means, \( \kappa \) which falls on households. So, agents decide how much money balances
to hold each period, \( m_t \) and then decide whether to carry it as cash, \( m^c_t \) (and in which
denominations) or as digital, \( m^d_t \).

Finally, we will also allow for the emergence of a shadow economy in Section 5 by intro-
ducing sales tax \( \tau \) levied by the government on firms. We will assume that the government
cannot perfectly enforce tax collection, and has to exert effort to increase enforcement. Tax
enforcement agent exerts effort \( \omega \) into increasing enforcement/tax compliance.\(^2\) This affects
the probability of paying taxes by firms, \( P(\omega) \), where \( P'(\omega) > 0, P''(\omega) < 0, P(0) = 0, \)
\( P(\infty) = 1 \). The per-period consolidated government and central bank budget constraint is
given by:

\[
0 = T_t + P(\omega)\tau y_t + \phi_t M_{t+1} - \phi_t M_t, \text{ or } 0 = T_t + P(\omega)\tau y_t + \Pi \phi_t M_t,
\]

where \( T_t \) is lumpsum tax which adjusts every period to maintain equality.

3 Cash equilibrium

In this section we will characterize equilibrium with cash as the only means of payment; we
introduce the digital medium in the next section.

Households

We begin with households. Let \( W_t(\cdot) \) denote the value function of households at the
beginning of late consumption sub-period (in which general goods is consumed). In each such
sub-period, the state variable is the current money balances \( m_t \), and the choice variables are
general goods consumption \( q_t \), labor supply \( l_t \) and money balances to be carried to the next
early consumption sub-period \( m_{t+1} \). There is discounting between these two sub-periods so
the continuation value of early consumption, \( V_{t+1} \) which depends on the money balances
brought forward \( m_{t+1} \) is pre-multiplied by \( \beta \). This gives us the following problem:

\[
W_t(m_t) = \max_{q_t, l_t, m_{t+1}} \{ U(q_t) - l_t + \beta V_{t+1}(m_{t+1}) \},
\]

\(^2\) The authority has to exert effort to access the transaction during the early consumption sub-period as
agents and transactions are anonymous then. Moreover, taxes are collected in the special goods \( y \).
\[ \begin{align*}
q_t &= l_t w_t + \Delta_t + T_t + \phi_t m_t - \phi_t m_{t+1}, \\
\text{s.t.} \\
\end{align*} \]

where \( w_t \) is wages per unit of labor supplied, \( \Delta_t \) is firm profits transferred to households, \( T_t \) is lumpsum taxes and \( \phi_t \) is the value of money in terms of the general good. Substitute for \( l_t \) and normalize \( w_t = 1 \) or adjust the disutility from labor to get the following problem for money holdings. This simplification makes the problem tractable, which follows from the quasi-linear structure of preferences:

\[ W_t(m_t) = \max_{m_{t+1}} \{-\phi_t m_{t+1} + \beta V_{t+1}(m_{t+1})\}. \tag{2} \]

This portfolio decision problem that determines the total money balances held at the end of every period is common to both cases with or without digital means. But, to solve this we need to know \( V_t(m_t) \) which depends on the choice of means of payment. If the household can only carry cash (assume for now \( \epsilon = 1 \), when we introduce heterogeneous types we will add \( \epsilon^i \) for \( i \in \mathbb{I} \)), \( V_t(m_t) \) is equal to:

\[ \begin{align*}
= &-\phi_t \gamma m_t + \phi_t \gamma (k-1) \left[ \frac{m_t}{k} \right] + \alpha \max_{y_t} \{ \epsilon u(y_t) + W_t \left( \phi_t m_t - k \eta \phi_t \left[ \frac{m_t}{k} \right] - p_t y_t \right) \} + (1-\alpha) W_t \left( \phi_t m_t - k \eta \phi_t \left[ \frac{m_t}{k} \right] \right) + (1-\alpha) W_t \left( \phi_t m_t - k \eta \phi_t \left[ \frac{m_t}{k} \right] \right), \\
\text{s.t.} \\
p_t y_t &\leq \phi_t m_t - k \eta \phi_t \left[ \frac{m_t}{k} \right].
\end{align*} \tag{3} \]

This value includes the carrying cost \( \gamma \) of each bill held and the net utility from consumption of the special good if the household receives an early consumption shock with probability \( \alpha \). The carrying cost depends on the denominations held. Let \( [x] \) denote the integer part of \( x \) and recall that \( k > 1 \) is the unit of high denomination bills. In nominal terms, the cost of carrying the portfolio \( m_t \) given that the maximum number of high-denomination bills are held (equal to the integer part of \( m_t/k \)) and the remaining in low-denomination bills simplifies to:

\[ -\gamma \left[ m_t/k \right] - \gamma k \left( m_t/k - \left[ m_t/k \right] \right) = -\gamma m_t + \gamma (k-1) \left[ m_t/k \right]. \]

Along with the carrying cost of cash, the value function also includes the net utility from consumption of the special good if the household receives an early consumption shock with probability \( \alpha \). In this case, she consumes \( y_t \) and if she does not receive this shock (with probability \( 1 - \alpha \)), she carries forward the value of genuine bills in her portfolio to
the late consumption period. Finally, note that households consumption $y_t$ is constrained by the value of genuine bills in her portfolio as a high-denomination bill is a counterfeit with probability $\eta$ which is detected in the next stage. Thus, she gets $1 - \eta$ times the value held in high denomination bills, i.e. consumption of special goods is constrained by:

$$k(1 - \eta)\phi_t[m_t/k] + k\phi_t (m_t/k - [m_t/k]) = \phi_t m_t - k\eta\phi_t[m_t/k].$$

We can simplify the above value of the first stage consumption $V_t(\cdot)$ by using the linearity of $W_t(\cdot)$ i.e.

$$V_t(m_t) = -\phi_t\gamma m_t + \phi_t\gamma(k - 1)\left[\frac{m_t}{k}\right] + \alpha \max_{y_t}[\epsilon u(y_t) - pt y_t] + \phi_t m_t - k\eta\phi_t \left[\frac{m_t}{k}\right] + W(0), \quad (4)$$

s.t.

$$pt y_t \leq \phi_t m_t - k\eta\phi_t \left[\frac{m_t}{k}\right].$$

Now to get the households choice of money holdings, take the early consumption value function one period forward and plug in to the $W(\cdot)$ value function in (2). Ignore constants to get the following portfolio choice maximization problem for a buyer each period:

$$\max_{m_t} \left\{ -\phi_{t-1}m_t + \beta \left\{ -\phi_t\gamma m_t + \phi_t\gamma(k - 1)\left[\frac{m_t}{k}\right] + \alpha \max_{y_t \leq \mathbb{C}_t}[\epsilon u(y_t) - pt y_t] + \phi_t m_t - k\eta\phi_t \left[\frac{m_t}{k}\right] \right\} \right\}, \quad (5)$$

where $\mathbb{C}_t \equiv (\phi_t/p_t)m_t - k\eta(\phi_t/p_t) [m_t/k]$ is the constraint on early consumption. Each buyer takes the prices $p_t$, $\phi_t$ and the degree of counterfeiting $\eta$ as given and maximizes the above. Before solving this problem we obtain $\eta$ and $p_t$, then we will derive $m_t$ given $\phi_t$.

Counterfeiting, $\eta$

In discussing the above portfolio maximization problem, we assumed that to carry $m_t$ money balances the maximum number of high-denomination bills are held (equal to the integer part of $m_t/k$) and the remaining in low-denomination bills. The choice of this division and subsequently the decision of total money balances to be held, $m_t$ depends on $\eta$, an endogenous object. This proportion of genuine bills or the degree of counterfeiting in the economy, depends on the decision of potential counterfeiters.

To keep things simple, recall that we assumed the nominal cost of counterfeiting, $\delta$ per bill is low enough to counterfeit any number of the high-denomination bills but high
enough to not counterfeit any low denomination ones. In real terms and given that the low-denomination bill is fully divisible, we get, $\phi_t < \delta \phi_t < k \phi_t$, i.e. the counterfeiter would like to counterfeit as many high-bills as she likes in any period. But, there will be an upper limit on counterfeiting because if $\eta$ is too high, households will not demand any high-bills. In fact, $\eta$ will be such that households are indifferent between holding their portfolio $m_t$ fully in low denomination bills or in a mixed form (with $[m_t/k]$ bills held in the high-denomination and $m_t - k[m_t/k]$ in low where $m_t$ is the nominal value of bills held).

We can thus replace the portfolio choice problem in (5) with high and low denominations with the following problem with only low denominations:

$$\max_{m_t} \{-\phi_{t-1} m_t + \beta \{-\gamma \phi_t m_t + \alpha \max_y [\epsilon u(y_t) - p_t y_t] + \phi_t m_t\}\}. \quad (6)$$

Thus, the portfolio choice problem does not directly depend on $\eta$ and more importantly our problem is now differentiable. We first obtain the optimal early consumption demand by households $y_t$ given $\phi_t$ and $p_t$ by solving $\max_y [\epsilon u(y_t) - p_t y_t]$ subject to $p_t y_t \leq \phi_t m_t$. We get that,

$$y_t = \min\left\{\frac{\phi_t m_t}{p_t}, u^{-1}(p_t)\right\}. \quad (7)$$

**Firms**

We now solve for $p_t$ from the problem of each perfectly competitive firm which is relatively straightforward. Firm’s expected revenue in terms of the numéraire is,

$$z_t = \bar{z} + \max_{y_t} [-c(y_t^*) + p_t y_t^*]. \quad (8)$$

Under the assumption of linear cost i.e. $c(y) = y$, we get $p_t = c'(y_t^*) = 1$. Thus, $z_t = \bar{z}$. Market clearing in the early-consumption stage given equal number of firms and households is $y_t^* = \alpha y_t$.

**Money demand**

We now go back to the household’s problem to solve its portfolio choice problem given by (6) to derive money demand. We get the following first order condition (using $p_t = 1$ and $y_t$ from (7)):

$$- \phi_{t-1} + \beta \phi_t \{-\gamma + \alpha [\epsilon u'(\phi_t m_t) - 1] + 1\} = 0, \quad (9)$$
Note that we used $y_t = \phi_t m_t$ from (7) (this is the standard result that households do not bring more real balances than they need in trade as money is costly to hold). The first-order condition simplifies to:

$$\alpha [\epsilon u'(\phi_t m_t) - 1] - \gamma = \frac{\phi_t - 1}{\phi_t} (1 + \rho) - 1.$$  \hspace{1cm} (10)

Taking $\phi_t$ as given, households choice of $m_t$ i.e. money demand satisfies the above.

**Preference heterogeneity**

So far we have assumed that there is one type of household, but it easy to extend the model to incorporate different types of households with varying preferences. We let $\epsilon$ vary across the set of households given by $\mathbb{I}$ such that each type-$i$’s utility is pre-multiplied by $\epsilon^i$. We also assume that the proportion of type-$i$’s is given by $\pi^i$. The first-order condition that determines $m^i_t$ for type-$i$ is given by:

$$\alpha [\epsilon^i u'(\phi_t m^i_t) - 1] - \gamma = \frac{\phi_t - 1}{\phi_t} (1 + \rho) - 1.$$  \hspace{1cm} (11)

It is straightforward to see that the following result on the relation of money balances held by different types of agents holds.

**Lemma 1.** Given $\phi_t > 0$, if $\epsilon^i > \epsilon^j$ for all $\{i, j\} \in \mathbb{I}$ then $m^i_t > m^j_t$.

**Market clearing**

The market clearing condition for money gives us the price of money, $\phi_t$. In the last stage, real money demand and supply are equal,

$$\sum_{i \in \mathbb{I}} \pi^i y^i_t = \phi_t \sum_{i \in \mathbb{I}} \pi^i m^i_t = \phi_t M^s_t,$$  \hspace{1cm} (12)

since firms do not carry money. Due to counterfeiting, the supply of money, $M^s_t$ will reflect the counterfeits on the high-denomination bills. We get,

$$M^s_t = M_t + \frac{\eta}{(1 - \eta) k} \left[ \frac{M_t}{k} \right],$$  \hspace{1cm} (13)

where $M_t$ is the supply of money by the central bank and the term on the right is supply of counterfeit bills. Since $[M^s_t/k]$ number of bills are held in the high-denomination which
are the only ones counterfeited, \( \eta k[M^* / k] \) is the supply of high-denomination bills that are counterfeits. Since the nominal value of high-denomination bills in supply, \( M^*_h = M_h / (1 - \eta) \) (i.e. the genuine money supply by the central bank of high denomination bills \( M_h \), augmented by counterfeits), we get \( \eta k[M^* / k] = \eta / (1 - \eta) k[M_t / k] \) as the supply of counterfeits.

For the qualitative analysis that follows, we will ignore counterfeiting i.e. assume that \( \delta \) is very high such that \( \eta = 0 \), \( M^*_s = M_t \) and the cash decision reduces to only a single denomination. This is equivalent to the one with two denominations and counterfeiting under our earlier assumptions. This parameter will again feature in the quantitative and applications section, when we calibrate the parameter \( \gamma \), by using data on counterfeiting.

Finally, we can define a monetary cash equilibrium as a sequence \( \{ \phi_t \}_{t=0}^{\infty} \) which solves the first order differential equation given by (11), where \( \phi_t \) is bounded.

### 4 Cash and digital payments in equilibrium

The portfolio decision problem that determines the total money balances held at the end of every period, \( m_t \) is exactly of the same form as before as given by (2). The value of the early stage consumption, \( V_t \) however has to be modified to reflect the choice between cash and digital means of payments, given that the household brings the portfolio \( m_t \). Recall that we also introduced preference heterogeneity, where the utility of type-\( i \in I \) household is pre-multiplied by \( \epsilon^i \). So, each household type-\( i \) chooses her money balance, \( m^i_t \) as cash, \( m^ic_t \) or digital, \( m^id_t \) depending on,

\[
V^i_t = \max \{ V^{ic}_t, V^{id}_t \},
\]

(14)

where \( V^{ic}_t \) is the value of the early stage consumption if household-\( i \) brings in cash and \( V^{id}_t \) is the same value if the digital means of payment is carried. The former, \( V^{ic}_t \) is exactly the same as before as given by (3). Thus, if \( V^i_t = V^{ic}_t \), then the same analysis also carries, and the first-order condition determining \( m^i_t \) will be given by (11) for given prices, \( \phi_t \). If instead, \( V^i_t = V^{id}_t \), then \( V^{id}_t \) is given by (15) below,

\[
V^{id}_t(m^i_t) = -\kappa + \alpha \max_{y^i_t} \{ \epsilon^i u(y^i_t) + W^i_t(\phi_t m^i_t - p_t y^i_t) \} + (1 - \alpha) W_t(\phi_t m_t),
\]

(15)
s.t.  \[ p_t y_t^i \leq \phi_t m_t^i. \]

The value of the early stage consumption if household-\( i \) brings in digital means includes the fixed usage cost \( \kappa \) and the net utility from consumption as before. Next, using the same intermediate steps as before used to obtain (5) or (6), the household-\( i \)’s portfolio choice problem becomes:

\[
\max_{m_t^i} \{-\phi_{t-1}m_t^i + \beta\{-\kappa + \alpha[\epsilon^i u(y_t^i) - y_t^i] + \phi_t m_t^i\}\},
\]

(16)

where \( y_t^i = \min\{\phi_t m_t^i, u^{-1}(1/\epsilon^i)\} \) as derived from (7) using \( p_t = 1 \) from the firm’s problem which remains exactly as before.

Each buyer type-\( i \) takes the price \( \phi_t \) as given and maximizes the above. We get the following first-order condition, assuming an interior solution,

\[
\alpha[\epsilon^i u'(\phi_t m_t^i)] - 1 = \frac{\phi_t^{-1}}{\phi_t} (1 + \rho) - 1.
\]

(17)

It is straightforward to see that Lemma 1 will hold when money is held in digital means as well.

The key difference between (11) and (17) is the carrying cost of cash, \( \gamma \) in the former. This implies that a household holding cash carries less money balances and consumes less goods in the early consumption period, but also does not have to pay the fixed cost \( \kappa \) (which does not show up in the first order conditions). This result is in line with the general observation that individuals economize on cash holdings more so than on digital means, which may lead to a lower consumption when using cash if the money constraint binds. It is also consistent with the finding in Runnemark et al. (2015) where they find that consumers pay more using debit cards than cash. This result is formalized in the Lemma below.

**Lemma 2** (Cash versus Digital). Given \( \phi_t > 0 \) and \( \gamma > 0 \), we get \( m_t^{id} > m_t^{ic} \) as given in (17) and (11) respectively for all \( i \in I \).

Finally, we can derive \( \phi_t \) from the market clearing condition as in section 3 as given by (12).
4.1 Steady state equilibrium

We now focus on stationary monetary equilibria where $1 + \Pi = M_t/M_{t-1} = \phi_{t-1}/\phi_t$ is the inflation rate and $\iota = (1 + \Pi)(1 + \rho) - 1$ is the nominal interest on an illiquid bond which represents the opportunity cost of holding money.

The first-order condition that determines the choice of real balances held by households carrying cash in the steady state with preference denoted by $ic$ becomes:

$$\alpha[\epsilon^{ic}u'(y^{ic}) - 1] = \iota + \gamma,$$

and, for households carrying digital means of payment in the steady state with preference denoted by $id$, it becomes:

$$\alpha[\epsilon^{id}u'(y^{id}) - 1] = \iota.$$  \hspace{1cm} (19)

The above two first-order conditions equate the households marginal utility of consumption to its cost. Note that there is a monetary wedge equal to $(\iota + \gamma)/\alpha$ in (18) and $\iota/\alpha$ in (19) between the households marginal utility of consumption $[\epsilon u(y)]$ and price of the good (= 1).

The market clearing is,

$$\sum_{i \in I} (\pi^{ic}y^{ic} + \pi^{id}y^{id}) = \phi M.$$ \hspace{1cm} (20)

Whether household-$i$ carries cash ($-ic$) or digital ($-id$) i.e. the choice of means of payment depends on relative costs. So, we can now obtain thresholds on cost of holding digital means, $\bar{\kappa}^i$ above which household-$i$ carries cash, as given by (21) derived in the following lemma. Note that we have defined, $s(x^i) \equiv \epsilon^i u(x^i) - x^i$ for brevity. If $\kappa = \bar{\kappa}^i$, then household-$i$ is indifferent between holding cash and digital, and if $\kappa > \bar{\kappa}^i$, then prefers cash. From Lemma 2, since $m^{id} > m^{ic}$, this threshold on the cost of digital means, not only depends on the carrying cost of cash but also the net benefit from carrying extra money balances when using digital means. So we get that $\bar{\kappa}^i > \gamma \phi m^{ic}$ because there has to be adjustment for the net benefit.

**Lemma 3** (Indifference thresholds). (i) Household-$i$ holds cash if and only if $\kappa > \bar{\kappa}^i$, where $\bar{\kappa}^i$ is given by:

$$\bar{\kappa}^i = \frac{\gamma y^{ic}}{\alpha[s(y^{id}) - s(y^{ic})]} - \frac{\iota(y^{id} - y^{ic})}{\alpha}. $$ \hspace{1cm} (21)
(ii) $\bar{\kappa}_i$ is increasing $\epsilon_i$.

Furthermore, we get that $\bar{\kappa}_i$ is higher for types with a higher $\epsilon_i$ as shown in Lemma 3 (ii). This implies that if type $i < k$, \{i, k\} $\in \mathbb{I}$ prefers to hold digital means of payment, then $k$ also prefers the same. This follows from the fixed cost structure of digital means of payment (also, note that while we assume fixed types of households, we can think of them as a household desiring different consumption quantities such that ones with a desire to consume lower quantities, prefers cash while if the shock is for a larger consumption bundle then digital).

4.2 Equilibrium payments regimes

For the analysis that follows, we will assume that there are three types of agents, $l, m, h \in \mathbb{I}$ with preferences given by $\epsilon_h > \epsilon_m > \epsilon_l$ and proportions $\pi^l + \pi^m + \pi^h = 1$. This is without loss of generality, and assumed primarily for a clear exposition of results. Define $\pi_d \in [0, 1]$ as the proportion of households using digital means.

**Definition 1.** Define a steady state monetary equilibrium with,

(i) only cash payments, $\pi_d = 0$ as a tuple $(\phi, y^{lc}, y^{mc}, y^{hc}) \in \mathbb{R}_+^4$ where $y^{lc}, y^{mc}, y^{hc}$ solve (18) and $\phi$ is derived from (20), if $\bar{\kappa}_h < \kappa$, where $\bar{\kappa}_h$ solves (21).

(ii) partial digital payments, $\pi_d = \pi^h$ as a tuple $(\phi, y^{lc}, y^{mc}, y^{hd}) \in \mathbb{R}_+^4$ where $y^{lc}, y^{mc}$ solve (18), $y^{hd}$ solves (19) and $\phi$ is derived from (20), if $\bar{\kappa}_m < \kappa < \bar{\kappa}_h$, where $\bar{\kappa}_m, \bar{\kappa}_h$ solve (21).

(iii) partial digital payments, $\pi_d = \pi^h + \pi^m$ as a tuple $(\phi, y^{lc}, y^{md}, y^{hd}) \in \mathbb{R}_+^4$ where $y^{lc}$ solves (18), $y^{md}, y^{hd}$ solve (19) and $\phi$ is derived from (20), if $\bar{\kappa}_l < \kappa < \bar{\kappa}_m$, where $\bar{\kappa}_l, \bar{\kappa}_m$ solve (21).

(iv) only digital payments, $\pi_d = 1$ as a tuple $(\phi, y^{ld}, y^{md}, y^{hd}) \in \mathbb{R}_+^4$ where $y^{lc}, y^{md}, y^{hd}$ solve (19) and $\phi$ is derived from (20), if $\kappa < \bar{\kappa}_l$, where $\bar{\kappa}_l$ solves (21).

**Proposition 1** (Existence and uniqueness of equilibrium in each payments regime). There exists a unique steady state monetary equilibrium for each payments regime as defined in Definition 1.

Each regime gives the fraction of population using digital means of payments $\pi_d$ as shown in Figure 1. These regimes are unique depending on the cost of digital payments indifference.
thresholds $\bar{\kappa}^i$. Thus, there is no multiplicity or strategic complementarity unlike some other papers with multiple means of payments such as Lotz and Vasselin (2019). We will discuss this, and other aggregate outcomes for each payments regime after we include a shadow sector in the next section.

5 Equilibrium with shadow economy

We now include a more active role for government/taxation in the baseline model with the two payment instruments. Since we are interested in exchange/trade between agents, we focus exclusively on sales tax $\tau$ levied by government on firms in the early consumption period on goods sold by them. Recall that we assumed the government cannot perfectly enforce tax collection, and has to exert effort to increase enforcement. There is a tax enforcement agent who exerts effort, $\omega$, into increasing enforcement/tax compliance. This affects the probability of firms paying taxes $P(\omega)$ where $P'(\omega) > 0$, $P''(\omega) < 0$, $P(0) = 0$, $P(\infty) = 1$. If $P(\omega) < 1$, then firms do not pay taxes in full which will then make them part of the shadow economy.

Now we modify the firms’ and households’ problem to include taxes, taking the probability of paying taxes, $P(\omega)$ as given. Firm’s now have to pay taxes which they pass on to consumers by increasing their prices to $p^{\tau} = 1/(1 - P(\omega)^{\tau})$. The household-$i$’s portfolio choice problem if she carries digital means is still given by (6) but now the output is modified to $y_{i,t}^{\tau} = \phi_t m_{i,t}^{\tau}/p^{\tau} = \min \{\phi_t m_{i,t}^{\tau}[1 - P(\omega_t^{\tau})], u^{-1}(1/[1 - P(\omega_t^{\tau})])\}$. This gives us the choice of $m_{i,t}^{\tau}$ by the following first order condition,

$$\alpha [\epsilon^{i} u'(\phi_t m_{i,t}^{\tau}(1 - P(\omega_t^{\tau})) - 1)] [1 - P(\omega_t^{\tau})] = \frac{\phi_t - 1}{\phi_t} (1 + \rho) - 1.$$  \hfill (22)

Similarly, the first order condition if the household carries cash is modified from (11) to:

$$\alpha [\epsilon^{i} u'(\phi_t m_{i,t}(1 - P(\omega_t^{i})) - 1)] [1 - P(\omega_t^{i})] - \gamma = \frac{\phi_t - 1}{\phi_t} (1 + \rho) - 1,$$  \hfill (23)

(22) and (23) above give the choice of money balances $m_{i,t}^{\tau}$ when the household decides to carry digital or cash means respectively for a given $\phi_t$ and $\omega_t^{\tau}$. It is straightforward to see that the results from both Lemmas 1 and 2 will continue to hold given $\phi_t$ and $\omega_t^{i}$. Price $\phi_t$ is derived from the market clearing condition as before and is given by (12). We will
now derive the tax enforcement agent’s effort $\omega^i_t$ which will give us $P(\omega^i_t)$, the probability of compliance.

The tax enforcement agent takes the tax rate $\tau$ and early consumption output of household-$i$, $y^i_t = \phi_t m^i_t$ as given and maximizes tax revenue net of (linear) cost of effort $\omega^i$, to solve the following problem (suppress the time subscript),

$$\max_{\omega^i} \{ P(\omega^i) \tau y^i - \omega^i \}. \quad (24)$$

The first-order condition gives,

$$P'(\omega^i) \tau y^i = 1. \quad (25)$$

Assume that the probability of compliance/collection as a function of enforcement agent’s effort, $\omega \geq 0$ is $P(\omega) = (1 - \exp^{-\lambda \omega})$ where $\lambda > 0$ is the rate parameter on the compliance probability. This functional form satisfies the assumptions on $P(\cdot)$ as laid out previously. Thus, we get

$$\omega^i = \frac{\ln(y^i \tau \lambda)}{\lambda}. \quad (26)$$

Note that we need $y^i \tau \lambda \geq 1$ for $\omega^i \geq 0$, else we set $P(\omega^i) = 0$. We assume for the qualitative analysis that $\lambda$ is sufficiently large so we get $P(\omega^i) = 1 - 1/y^i \tau \lambda$. It is straightforward to see that enforcement effort is increasing in output $y^i$ and tax rate $\tau$.

### 5.1 Steady state shadow equilibrium

We now focus on stationary monetary equilibria as before where $1 + \Pi = M_t/M_{t-1} = \phi_{t-1}/\phi_t$. The first-order condition that determines the choice of real balances held by households carrying cash in the steady state with preference denoted by $ic$ becomes:

$$\alpha \left[ \epsilon^{ic} u' \left( y^{ic} \left[ 1 - P(\omega^{ic}) \tau \right] \right) - 1 \right] [1 - P(\omega^{ic}) \tau] = \iota + \gamma, \quad (27)$$

where $y^{ij} = \phi m^{ij}$ for $j \in \{c, d\}$ and, for households carrying digital means of payment in the steady state with preference denoted by $id$, it becomes:

$$\alpha \left[ \epsilon^{id} u' \left( y^{id} \left[ 1 - P(\omega^{id}) \tau \right] \right) - 1 \right] [1 - P(\omega^{id}) \tau] = \iota. \quad (28)$$

17
Note that if $\tau = 0$, the above two first order conditions reduce to (18) and (19) respectively. Recall that $\omega^i$ is given by (26), and the probability of paying taxes $P(\omega^{ij}), j \in \{c, d\}$ was derived as,

$$P(\omega^{ij}) = 1 - \frac{1}{y^{ij} \tau \lambda}, j \in \{c, d\}.$$  (29)

The market clearing is,

$$\sum_{i \in I} \left( \pi^{ic} y^{ic} + \pi^{id} y^{id} \right) = \phi M.$$  (30)

This gives us the full system of equations that determines the steady state shadow equilibrium with cash or digital means. Before we describe the different equilibrium regimes, consider the difference in tax rates across households.

Due to the difference in tax enforcement agent’s efforts $\omega^{ij}$, the effective tax rate will vary across households by their preferences and also by their choice of means of payments. We can now define the effective tax rate as $\tilde{\tau}^{ij} = P(\omega^{ij}) \tau$ where $i \in I$ and $j \in \{c, d\}$, using (29) we get,

$$\tilde{\tau}^{ij} = \tau - \frac{1}{y^{ij} \lambda}, j \in \{c, d\}.$$  (31)

Thus, clearly the effective tax rate is increasing in $y^{ij}$. In this respect, this tax has a progressive nature. And, since $y^{ij}$ is higher for higher types and on using digital means then the effective tax rate is higher on digital transactions.$^3$ These results are formalized in the following lemma.

**Lemma 4** (Effective tax rate, $\tilde{\tau}$). (i) If $\epsilon^i > \epsilon^k$ for all $\{i, k\} \in I$ then $\tilde{\tau}^{ij} > \tilde{\tau}^{kj}, j \in \{c, d\}$, (ii) $\tilde{\tau}^{id} > \tilde{\tau}^{ic}$ for all $i \in I$.

We now obtain thresholds on cost of holding digital means, $\tilde{\kappa}^i$ above which household-$i$ carries cash in the presence of taxation, which will help determine the existence of different shadow equilibrium regimes. The following lemma gives this threshold, and shows that the previous result from Lemma 3 (ii) also holds here.

$^3$Note that we could also assume that $\lambda_d \geq \lambda_c \geq 0$ i.e. the digital transactions are easier to enforce. If we assume that $\lambda_d$ tends to infinity (implying $\tilde{\tau}^{id} = \tau$) and $\lambda_c$ tends to zero (such that $P(\omega^i) = 0$), then cash transactions do not pay taxes at all, and digital transactions always pay. But, we do not make this assumption, instead we let the $\lambda$s be equal and the difference in enforcement be fully endogenous.
Lemma 5 (Thresholds). (i) Household-i holds cash if and only if $\kappa > \tilde{\kappa}_i$, where $\tilde{\kappa}_i$ is given by:

$$
\tilde{\kappa}_i = \underbrace{\gamma y_{ic}}_{\text{carrying cost}} + \alpha \left[ s^i \left( y^{id} [1 - P(\omega^{id}) \tau] \right) - s^i \left( y^{ic} [1 - P(\omega^{id}) \tau] \right) \right] - \underbrace{\iota (y^{id} - y^{ic})}_{\text{cost of extra money}}
$$

(ii) $\tilde{\kappa}_i$ is increasing in $\epsilon_i$.

5.2 Equilibrium regimes with shadow

Now, as before we can define a steady state monetary equilibrium as in Definition 1 with indifference thresholds on digital payments, $\tilde{\kappa}^{ij}$ instead of $\bar{\kappa}^{ij}$ and modified equilibrium conditions. Replace (18), (19), (20) and (21) with (27), (28), (30) and (32) respectively and, (29) gives $P(\omega_{ij})$. Proof of existence and uniqueness also follows the same steps as in the previous section by using Lemmas 4 and 5.

Figure 1 shows the different equilibrium regimes given by the fraction of households using digital means of payments $\pi_d$ as a function of the cost of using digital payments, $\kappa$. This fraction depends on where $\kappa$ lies with respect to its thresholds and it is decreasing in digital payments cost, both with and without taxation.

The indifference thresholds on the cost of digital payments under taxation $\tilde{\kappa}$ is lower than in a model without tax $\bar{\kappa}$ for all the numerical examples we worked with, including our baseline calibration. This is fairly intuitive as well. Consider for example the case when $\tilde{\kappa}^m < \kappa < \bar{\kappa}^m$. In this case, if there were no taxes, then types-$m$ and $h$ would use digital payments, i.e. the case in Definition 1 (iii) with $\kappa < \bar{\kappa}^m$. But, if there is taxation (and given that $\kappa > \tilde{\kappa}^m$), then types-$m$ prefer not to hold digital payments as it is now relatively cheaper for them to use cash. Thus, taxation has reduced usage of digital payments as the effective tax rate on digital payments is higher due to higher enforcement.

Now we can compare the outcomes under different payments regimes including prices, output and the size of the shadow economy. Define aggregate early consumption output as $Y = \alpha \sum_{i \in I} \pi_i y_i$. From Lemma 2, $Y$ is maximized in the only digital payments equilibrium as given by Definition 1(iv). The size of the shadow i.e. the output generated on which tax is evaded as a fraction of output on which tax is not evaded (i.e. measured output) is defined
Figure 1: Indifference thresholds on cost of digital payments
Note: Red dashed line is for the model with taxation and green full line is for without taxation

as,

$$S \equiv \frac{\alpha \sum_{i \in I} \pi^i [1 - P(\omega^i)]}{} \frac{y^i}{\sum_{i \in I} \pi^i q^i + \alpha \sum_{i \in I} \pi^i P(\omega^i) y^i}.$$ (33)

The size of the shadow economy is minimized in the only digital payments equilibrium as given by Definition 1(iv). To see this, plug in for $P(\omega^i)$ from (29) in (33) to reduce the expression for the shadow size to,

$$S = \frac{\alpha (\tau \lambda)^{-1}}{} \frac{\pi^i q^i + \alpha \sum_{i \in I} \pi^i y^i}{\sum_{i \in I} \pi^i q^i + \alpha \sum_{i \in I} \pi^i y^i - \alpha (\tau \lambda)^{-1}}.$$ (34)

Thus, in an all digital payments equilibrium, which occurs if $\kappa$ is low enough, the early consumption output is maximized as households carry higher money balances as given by Lemma 2. And, since the effective tax rate on digital transactions is higher, the size of the shadow also decreases in an all digital payments equilibrium. However, the equilibrium is determined by payment costs and ultimately what happens to surplus i.e. utility net of costs and welfare depends on these costs. And, it matters whether $\gamma$ increases or $\kappa$ decreases for the economy to move to an equilibrium with higher digital payments as we will see.
5.3 Application to India’s demonetization

We will now apply the model to India’s demonetization and study its impact on macroeconomic aggregates. In an overnight surprise move, the two highest denomination bills in India were demonetized, that is 86% of currency in circulation ceased to be legal tender.

The immediate impact of this shock in the model at the beginning of the early consumption period, implies that firms will no longer accept cash as media of exchange. This implies an immediate fall in output for households who carried cash, $y^{ic}$ as well as their welfare, $W^{ic}$, and tax revenues paid by them equal to $\bar{\tau}^{ic} y^{ic}$. The output, welfare and taxes paid by households who carried digital means remain unaffected. Per-period welfare for household-$ij$, for $j \in \{c,d\}$ is defined as,

$$W^{ic} \equiv U(q^{ic}) - l^{ic} - \gamma y^{ic} + \alpha [u(y^{ic}_r) - y^{ic}_r],$$

$$W^{id} \equiv U(q^{id}) - l^{id} - \kappa + \alpha [u(y^{id}_r) - y^{id}_r],$$

where $y^{ij}_r = y^{ij}[1 - P(\omega^{ij})\tau]$. Recall that Lemma 5 determines if $j = c$ or $d$ i.e. if household-$i$ holds cash or digital means. The per-period aggregate welfare $\mathbb{W}$ includes tax revenues and is defined below,

$$\mathbb{W} \equiv \sum_{i \in I} \pi^{i} W^{i} + \pi^{i} P(\omega^{i})\tau y^{i}.$$  

(36)

On account of the cash holding households, aggregate welfare $\mathbb{W}$ falls after this policy shock.

After the sudden demonetization move, the demonetized bills had to be replaced with new bills. But, this remonetization process was slow and it took almost three quarters to get the money supply back to its trend growth, as shown in Figure 5 below. The institutional realities of the time meant that cash, as a means payments became more expensive as people had to line up at banks to get their bills exchanged and they could do so only in limited amounts at a time.\(^5\) In the model, this would imply an increase in the carrying cost of cash $\gamma$. This would in turn lead to a rise in the threshold of using digital payments for all types as given in Lemma 5.

\(^4\)We will analyze the effects from denomination change in the quantitative section.

\(^5\)While people who used digital means of payments before also found it costlier to use them given that accessing ATMs and using debit cards became cumbersome and costly. We assume in this section that this cost remained unchanged, in the calibration section we will adjust this cost as well and focus on the relative cost between cash and digital.
If it is now relatively cheaper for some types to use digital means of payment, there is an increase aggregate consumption for those types since $y_{ic} < y_{id}$ from Lemma 2. But, their surplus (utility net of costs) or household welfare, $W^i$ is still lower than before as they have to pay the fixed cost for using digital means, so this may not mean higher surplus or household welfare, even though it means higher consumption for some. Besides, the types for whom it is still costly to use digital means, do not switch and end up consuming less. Thus, aggregate private welfare $\sum_{i \in I} \pi^i W^i$ falls.

Due to the switch to digital means, there will be also be an increase in tax compliance, as $\tilde{\tau}_d > \tilde{\tau}_c$ from Lemma 4, but the effect on the fraction of the shadow economy $S$ is ambiguous. Using (34) note that since the types who use cash consume less, output $\sum_{i \in I} \pi^i y^i$ might fall and hence the size of the shadow can increase or decrease. Similarly, tax revenue may still fall as the types who use cash, consume less leading to a lower base and lower revenue, so the overall effect on tax revenue remains ambiguous.

Finally, as cash became available, the carrying cost of cash reduces back to its original level, and the thresholds above which cash is held for each household also returns back to their pre-shock levels. If there is no change in the cost of using digital, all variables return back to their pre-shock levels. But, if this cost $\kappa$ falls permanently\(^6\) such that a few types can switch to digital then output, inverse of prices, household welfare and aggregate welfare would all (weakly) rise and shadow $S$ reduce.

As discussed, the specific results depend on the relative change in costs and cost relative to thresholds for each type. In the next section we will calibrate the model to the Indian economy using consumption distribution data. This will help us quantify the aggregate and disaggregate effects of these shocks. For this section, we present a qualitative summary of the above results in the following proposition, by assuming that there are three types of agents, with preferences as $\epsilon_h > \epsilon_m > \epsilon_l$ and proportions $\pi^t + \pi^m + \pi^h$.

**Proposition 2 (Effects of Policy).** If $\gamma = \gamma_0$, $\kappa = \kappa_0$ where $\tilde{\kappa}^m(\gamma_0) < \kappa_0 < \tilde{\kappa}^h(\gamma_0)$, we get that,

---

\(^6\)Since there was also a large influx in new digital payment methods post demonetization, it may imply a fall in $\kappa$ in the model (or at least that $\kappa$ falls at a faster rate than $\gamma$). While we do not model such strategic complementarities, but if many households switch or if many firms accept digital means as possibly happened due to the increase in the relative cost of cash, there might have been incentives for floating more digital means in the market making them more accessible and hence $\kappa$ could fall. For example Crouzet et al. (2019) focuses on technological adoption of electronic payments by retailers and finds evidence for positive externalities in adoption.
(i) Sudden demonetization of currency leads to a fall in $y_l, y^m$ with $y_h$ unchanged, and a fall in aggregate welfare, $\mathbb{W}$ in (36).

(ii) If the increase in the carrying cost of cash $\gamma_1 > \gamma_0$ is such that $\bar{\kappa}_l(\gamma_1) < \kappa_0 < \bar{\kappa}_m(\gamma_1)$, it implies that $y^m$ increases, $y_l$ falls, $y_h$ is unchanged, $W^l$ and $W^m$ fall, $W^h$ is unchanged.

(iii) If the carrying cost of cash decreases to the original $\gamma_0$, and

(a) cost of digital payments $\kappa$ does not change, then all variables return to their pre policy shock levels,

(b) if the cost of digital payments decreases i.e. $\kappa_1 < \kappa_0$ such that $\bar{\kappa}_l(\gamma_0) < \kappa_1 < \bar{\kappa}_m(\gamma_0)$, then compared to their pre policy shock levels $y^m, W^m, W^h$ and $\mathbb{W}$ rise and $S$ is smaller. Other variables are unchanged.

Thus, the overall positive impact of the policy shock fell on digital users (new and old) only, if at all. This when compared to the clear negative consequences of this policy, seems an inefficient way to increase digital payments use. Policies to reduce the cost of digital payments usage directly would have these positive effects without any of the negative consequences of forcing such a change.

6 Quantitative Results

6.1 Parameter Calibration

The model is calibrated to the Indian economy from Q2:1996 to Q3:2016. We calibrate the baseline parameters in the model that are common across household types to match aggregate data observations. We then use the monthly per capita consumption expenditure (MPCE) data from the Consumer Expenditure Survey (CES) to pin down household-$i$’s preference parameter $\epsilon_i$ and its proportion $\pi_i$.

The assumed utility functions are standard in the literature. For the early consumption goods market [or decentralized market in Lagos and Wright (2005)], we assume a generalized version of the standard constant relative risk aversion preferences, as $u(y) = [(y + b)^{(1-\sigma)} - b^{(1-\sigma)}] / (1 - \sigma)$, where $\sigma > 0$ and $b \approx 0$. The late consumption goods market utility (or centralized market) is assumed to be $U(x) = A \log(x)$, which implies $x^* = A$.

The parameters of the utility functions $(A, \sigma)$ are typically calibrated to match the relationship between money demand as a fraction of nominal GDP, $M1/PY$ and nominal
interest rate, $\iota$ in data. We use data on M1 and the return on 91-day Treasury Bill from the Database for Indian Economy (DBIE), RBI. We choose the weightage of late consumption in utility i.e. the level parameter $A$ to match average money demand and the degree of risk aversion in utility, $\sigma$ to match the money demand elasticity, see Figure 2. The money demand function for the average household as a fraction of measured GDP in the model is given by:

$$\frac{M/p}{Y^m} = \frac{y}{\alpha P(\tau, y, \lambda)y + A},$$

which is a function of nominal interest rate $\iota$ through $y$ as given in (27), ignoring the $i$-superscript when considering the average household. For the functional forms assumed and for the average household using cash i.e set $\epsilon^i = 1$ and (27) becomes,

$$y^{-\sigma} [1 - \tau P(\tau, y, \lambda)]^{1-\sigma} = \frac{\iota + \gamma}{\alpha} + 1,$$

(37)

where $P(\tau, y, \lambda) = \min\{0, 1 - 1/y\tau\lambda\}$. We set the average sales tax rate, $\tau$ equals to 10.3% which is the ratio of sales tax revenue to GDP for India over fiscal year 2014-15, obtained from an OECD survey in 2017. The informal output, $y$ will also depend on parameters $\lambda$, $\alpha$ and $\gamma$. We now discuss the calibration strategy of each of these. We choose the rate parameter on tax enforcement/compliance probability $\lambda$ such that the measure of shadow economy’s output as a fraction of total turnover in the model given by $[1 - P(\tau, y, \lambda)]\alpha y/(A + \alpha y)$ matches the share of turnover in non tax net equal to 20.6% as reported in the Economic Survey 2017-18.

Now consider the calibration strategy of $\alpha$, the probability of receiving a positive preference shock for early consumption. A plausible method to calibrate this parameter is to match it to an observed share of output in the early consumption sub-period as a fraction of total output in an economy i.e. $\alpha y/(\alpha y + A)$. This early consumption sub-period is characterized by decentralized/unorganized markets. Since we calibrate the model to India, a natural target for this is the share of unorganized sector output in total output.\(^7\) We rely on

\(^7\)The unorganized sector in the Indian context is defined by the National Commission for Enterprises in the Unorganized Sector as: “consisting of all unincorporated private enterprises owned by individuals or households engaged in the sale or production of goods and services operated on a proprietary or partnership basis and with less than ten total workers.” This sector is also referred as the informal sector in the Indian
data from the Economic Survey 2017-18, which estimates that 48% of total turnover as not being in the social security net, which we will use as the share of unorganized sector output.

We calibrate the cost of carrying cash $\gamma$ using data on the percentage of counterfeiting $\eta$ from the Reserve Bank of India (RBI)'s data on Fake Indian Currency Note for 2014-15. Their estimates are most likely a lower bound on counterfeit currency in circulation as the data would include only the detected counterfeits. The proportion of counterfeits of the high denomination bills is $\eta = 0.002\%$ if we include Rs.500 and Rs.1000 as high denomination. For the model we collapse the denominations into two: high being the Rs. 500 as $x_2$ and low as the Rs. 100 bill as $x_1$. So, $k = x_2/x_1 = 5$.

Recall that the nominal value of money held in the large denomination bills is $m_2 = k[m/k]$ and the remaining value in low denominations is $m_1 = m - k[m/k]$. Here $m$ is the nominal money supply held and $[\cdot]$ is the integer component. In equilibrium, money demand equals supply i.e. $M/p = \phi m = y$, where $\phi$ is the price of money in terms of the numéraire. Further, the denomination structure also implies that for both the denominations, we have $\phi m_2 = M_2/P(1-\eta)$ and $\phi m_1 = M_1/P$, where $M_1, M_2$ stand for supplies of the two denominations by the RBI. Note the we augmented the supply of $M_2$ given the proportion of counterfeits $\eta$. Define surplus from early consumption as $s(y) \equiv \epsilon u(y) - py$. Equate (5)
Table 1: Key calibration targets and parameter values

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Target</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM utility, $\sigma$</td>
<td>MD/GDP and i elasticity = -0.19</td>
<td>0.24</td>
</tr>
<tr>
<td>CM utility, $A$</td>
<td>MD/GDP average = 0.74 and i = 0.07</td>
<td>0.51</td>
</tr>
<tr>
<td>DM share, $\alpha$</td>
<td>Informal share = 0.48</td>
<td>0.81</td>
</tr>
<tr>
<td>Cost of cash, $\gamma$</td>
<td>counterfeiting data, $\eta = 0.002%$</td>
<td>3e-5</td>
</tr>
<tr>
<td>Tax rate, $\tau$</td>
<td>sales tax/GDP = 0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>Enforcement pr, $P$</td>
<td>Share of turnover in non tax net = 0.2</td>
<td>0.58</td>
</tr>
</tbody>
</table>

and (6) and re-write as,

$$\frac{\gamma M_2 k - 1}{k} \frac{1}{1 - \eta} + \alpha s \left( \frac{M}{P} - \frac{\eta}{1 - \eta} \frac{M_2}{P} \right) - \eta \frac{M_2}{1 - \eta} \frac{1}{P} = \alpha s \left( \frac{M}{P} \right),$$

The fraction of high denomination bills $M_2/M$ as reported by RBI equals 0.85 for 2014-15. So, we can substitute $M/P = y$ and $M_2/P = 0.85y$ in the above to obtain a closed form expression of $\gamma$ as a function of $y$, $\alpha$ and $\sigma$ after substituting the data on $\eta$ and $k$. Thus, using this expression for $\gamma$, the parameters $A, \sigma, \alpha, \lambda$ will be jointly determined as described previously.

We use the monthly per capita consumption expenditures (MPCE) for rural and urban regions in India from Consumer Expenditure Survey (CES) data for 2011-2012 to calibrate $\epsilon^i$ and $\pi^i$, the consumer types by preferences and their population proportions.\(^8\) We use data on fraction of rural population shares from World Bank staff estimates based on the United Nations Population Division’s World Urbanization Prospects: 2018 Revision equal to 0.67. Using this we compute the population weighted average $MPCE_{avg}$ for India. We obtain $\epsilon^i$ by matching the ratio of decile $i$’s consumption to the average i.e. $MPCE^i/MPCE_{avg}$ with its model counterpart, $[\alpha y^i(\epsilon^i) + \epsilon^i A]/(\alpha y + A)$.\(^9\)

In the left panel of Figure 3 below, the red asterisk indicates the population weighted average monthly per capita consumption $MPCE_{avg}$ for the economy as a whole. The blue solid line denotes the observed MPCEs for the urban population divided into deciles. The

\(^8\)This is the latest consumption data survey available as the 2017-18 survey was withdrawn.

\(^9\)Note that here we also pre-multiply the late consumption goods utility by $\epsilon^i$ i.e. $U_i(x^i) = \epsilon^i A \log(x_i)$, thus the utility obtained from late consumption for a household type $\epsilon^i$ is given by $x^i = \epsilon^i A$. And, since we considering fractions, the time period or prices will not matter.
dotted green line denotes the same for the rural population. The calibrated values of the parameter $\epsilon_i$, for urban and rural population deciles is given in the right panel of the figure.

For the baseline calibration, we assumed that the average household transacts only using cash. Thus, the cost of using digital payments, $\kappa$ was irrelevant to the average household except that we implicitly assumed that this cost is higher than $\tilde{\kappa}_{\text{avg}}$ such that the average household chooses to use cash instead of digital payments. For the general model, we can pin down this cost by using the thresholds on making digital payments for household type-$i$ $\tilde{\kappa}_i$ as defined in (32). Figure 4 gives this cost as a percentage of average early consumption $y$ for each population decile, separately for urban and rural. For example, the top urban consumption decile uses digital payments if this cost is less than 0.025% of average early consumption. As expected, this threshold is higher for the higher consumption deciles in both regions.

To find where $\kappa$ lies we first define digital payments very broadly as the demand deposits component of M1.\footnote{Households use these demand deposits directly for payments by writing checks, bank transfers, debit cards and increasingly through mobile payments. Since they need a bank account to initiate such digital payments, we consider demand deposits as part of digital financial architecture.} For our calibration period, demand deposits as a fraction of $M_1$ fluctuated between 0.3 and 0.4. And, the fraction of money holdings of the top two urban deciles in our model is 0.32 of total money holdings. Since their $\tilde{\kappa}$ is higher than the top rural decile’s, we conclude that $\kappa$ is somewhere between $\tilde{\kappa}^{U9}$ and $\tilde{\kappa}^{R10}$, and we set $\kappa = \tilde{\kappa}^{U9}$. The
implied cost is equal to 0.01% of average early consumption. This completes the calibration of the model and enables us to quantify the impact of India’s demonetization next.

6.2 Quantifying the impact of India’s demonetization

Now we quantify the impact of a sudden demonetization of high denomination bills that happened on November 8, 2016 in India and its subsequent, gradual re-monetization over the next four quarters. Using the calibrated model, we will examine the effects of this policy shock on the unorganized sector i.e. early consumption sector – output given by $Y = \alpha \sum_{i \in I} \pi^i y^i$ and surplus (utility net of cost) by $\sum_{i \in I} \pi^i \{\alpha[u(y^i) - y^i] - \gamma y^i\}$ if using cash or $\sum_{i \in I} \pi^i \{\alpha[u(y^i) - y^i] - \kappa\}$ if digital. We also quantify the impact on total output $\alpha Y + A$, total surplus $\sum_{i \in I} \pi^i W^i$, aggregate welfare $\mathbb{W}$ as defined in (36) and the size of the shadow economy $S$ as defined in (33). Using the monthly per capita consumption expenditure data as described in the previous section, we also consider the impact on rural and urban regions separately and the differential impacts on the respective consumption deciles. We apply the consumption distribution data to also analyze the effects on consumption inequality with a Lorenz curve.

Demonetization: We model the demonetization of high denomination bills in the early consumption period such that, households find their high denomination bills to be no longer acceptable for transactions. We use the consumption distribution data to derive the model
implied denomination structure. First, we find the nominal money balances held, by defining the average price level for the model as the ratio of nominal consumption per decile in the data to real consumption in the model. This gives us decile-wise money demand. Second, we divide the nominal money balances into the two denominations $x_1 = Rs.100$ and $x_2 = Rs.500$ by setting $m_2 = k[m_1/k]$ and allocating the remainder to $m_1$, where $i$ is the household type based on rural and urban consumption deciles. Using this process, we find that the average share of high denomination bills held by consumers in the model is 65% as compared to 85% in the data for 2014-15. So, the effect of demonetization in the model will only provide a lower bound for the results.

We find that the adverse effect of demonetization on rural consumption, output and surplus is greater than its urban counterpart, since consumers in the rural regions mostly use cash as their primary payments system. Unorganized sector output falls by 52% in rural regions and 40% in urban whereas surplus in this sector falls by 33% in rural compared to 20% in urban. The greater adverse effects on the rural economy are observable even at the aggregate level. Aggregate total output falls by 23% in rural and 18% in urban. The effect of a demonetization shock on the aggregate economy as measured by a fall in aggregate welfare by 16% is also perceptible. This is again primarily led by the effect of demonetization on unorganized rural consumption. Finally, the impact of the shock will be felt differently across the different deciles, depending on the proportion of cash used for transactions as well as the composition of higher denominations in their cash portfolio.

**Slow Remonetization:** The overnight demonetization shock was not immediately reversed as it took almost three quarters for money supply to go back to its pre-shock levels. As shown in Figure 5, $M_1$ at the end of November 2016 was 24% below its level in October 2016 and it took almost three quarters for it to normalize back to pre-demonetization shock levels. Furthermore, the composition of $M_1$ also changed. In October 2016 currency/$M_1$ stood at 0.6, while in November 2016 this ratio fell to 0.4 and gradually increased to 0.5, eventually going back to 0.6 after August 2017.

The process of this slow remonetization slowed the recovery process by significantly increasing the cost of payments. Cash became costlier, as not all old bills were immediately replaced with new ones, and the process to do so also imposed significant hardship on consumers. The increased strain on accessing payment methods also permeated to digital means
– defined broadly as any form of demand deposit – as the time spent accessing bank accounts, ATMs, setting up and processing mobile payments increased in the interim. However, the latter increase is still smaller than the much higher cost of cash.

We model this slow remonetization by adjusting the cost of using cash, $\gamma$ to match the remonetization rate from Figure 5 for the four quarters denoted by $Q_i$, $i = 1, 2, 3, 4$, starting in November 2016. We also adjust the cost of using digital payments system, $\kappa$ by incorporating the temporary substitution of currency with demand deposits over this period as described above. We determine the groups that switched to non-cash payments in the model, which is in turn used to update $\kappa$ relative to the new $\gamma$ every period. According to the model, households in a particular decile either use digital or cash payments but not both. So, given that in November 2016 the ratio of demand deposits to $M_1$ increased to 0.6 we can conclude that in the first quarter, $Q_1$ the top four urban deciles and the top rural decile switched to digital payments as they together held 58% of the total money holdings. We therefore, set $\kappa = \tilde{\kappa}^{U7}$. Using this method for the remaining periods, we find that $\kappa = \tilde{\kappa}^{U8}$ in $Q_2$ and $\kappa = \tilde{\kappa}^{U9}$ in $Q_3$ and $Q_4$.

The impact of the slow remonetization on aggregate and average sectoral outcomes based on the transition in payments methods as described previously is summarized in Table 2. The key aspect to note is that the effects of slow remonetization are also disproportionate.
and largely focused on the regions and groups that were not able to transition to digital means of payment.

Table 2: Percentage changes compared to pre-shock levels

<table>
<thead>
<tr>
<th></th>
<th>$Q_1$</th>
<th>$Q_2$</th>
<th>$Q_3$</th>
<th>$Q_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Aggregate output (% change)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>-5.4</td>
<td>-4.5</td>
<td>-1.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>Rural</td>
<td>-7.1</td>
<td>-5.3</td>
<td>-1.5</td>
<td>-0.3</td>
</tr>
<tr>
<td>Urban</td>
<td>-3.5</td>
<td>-3.7</td>
<td>-1.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>2. Unorganized output (% change)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>-10.2</td>
<td>-8.6</td>
<td>-2.7</td>
<td>-0.6</td>
</tr>
<tr>
<td>Rural</td>
<td>-16.2</td>
<td>-12.1</td>
<td>-3.3</td>
<td>-0.7</td>
</tr>
<tr>
<td>Urban</td>
<td>-5.7</td>
<td>-5.9</td>
<td>-2.3</td>
<td>-0.5</td>
</tr>
<tr>
<td>3. Surplus (% change)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>-6.4</td>
<td>-5.4</td>
<td>-1.8</td>
<td>-0.4</td>
</tr>
<tr>
<td>Unorganized</td>
<td>-14.2</td>
<td>-11.9</td>
<td>-3.9</td>
<td>-0.8</td>
</tr>
<tr>
<td>Rural Unorganized</td>
<td>-18.3</td>
<td>-14.8</td>
<td>-4.7</td>
<td>-0.9</td>
</tr>
<tr>
<td>Urban Unorganized</td>
<td>-11.2</td>
<td>-9.9</td>
<td>-3.3</td>
<td>-0.7</td>
</tr>
<tr>
<td>4. Aggregate welfare (% change)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-7.7</td>
<td>-6.5</td>
<td>-2.1</td>
<td>-0.4</td>
</tr>
<tr>
<td>5. Shadow economy (% change)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-5.2</td>
<td>-2.7</td>
<td>-0.4</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

$^1$ $Q_1$ refers to Nov’16, $Q_2$ to Feb’17 and so on.

Table 2 shows that the recovery in total output after the demonetization shock as compared to its pre-demonetization level was slow as the economy was remonetized only gradually with significant cost to consumers. But, the recovery in rural regions is slower than its urban counterpart owing to the slow remonetization as the rural regions primarily remained cash dependent. For instance, in $Q_1$ urban output drops by 3.5% while rural output drops by more than twice as much i.e. by 7.1%. Urban recovery in output is relatively faster as urban consumers were better equipped to switch to digital payments as their transactions were large enough to justify the fixed costs of making this switch. Thus, the primary source of differences in recovery for the two regions lies in the different capacities of the two groups to
switch to digital payments.

Second, transactions in the unorganized sector were slower to recover acting as a drag on the rest of the economy. The output deficit in $Q_2$ for the overall economy was 4.5% while that of the unorganized sector was 8.6% in comparison to their own pre-demonetization levels. Further, even in the unorganized sector, the rural unorganized sector was much slower to recover than its urban counterpart. While the urban unorganized sector accounts for 62% of total urban output, the rural unorganized sector only accounts for 44% of total rural output. Despite the predominance of urban unorganized sector in its output, the rural total output fall is greater due to magnitude of the fall in the rural unorganized output as it fell 2.5 times more than its urban counterpart.

Third, both aggregate and unorganized surpluses declined sharply due to the slow remonetization and took a long time to approach its pre-demonetization levels. Note that surplus is defined as utility net of costs, where the latter includes the increased cost of payments. While total surplus declined by 6.4% in $Q_1$ which is 18% more than its output decline for that quarter, the unorganized surplus reduced by 14.2% which is 38% more than its output decline for that quarter. Similarly, while total output was 0.3% below its pre-demonetization levels in $Q_4$, total surplus was 0.4% below its corresponding level. These comparisons are starker for unorganized surplus. Additionally, regional composition of surplus depicts similar contrast between a sharper fall in rural surplus than its urban counterpart.

Fourth, we observe that aggregate welfare – total output net of costs adjusted for tax revenue by the government – tracks unorganized surplus closely and is slow to recover as the economy remonetized slowly. Finally, the shadow economy recovers as the economy is remonetized. There is a fall in the size of the shadow economy owing to fewer cash payments. Since digital payments are easier to track it is harder to evade taxes on such transactions. But, in comparison with total and especially unorganized output which features a large and persistent negative impact of the slow remonetization, the effect on the shadow economy is of a smaller magnitude and is even less persistent. By $Q_2$, unorganized output is still 8.6% less than its pre-denomnetized levels. However, after falling to 5.2% in $Q_1$, the shadow economy adjusts quickly to minimize the shortfall by $Q_2$ to 2.7%.

Thus, the benefits of this policy shock owing to a temporary slowdown in the shadow economy and potentially higher tax collection is outweighed by the direct costs on payments and indirect effects on consumption and production of such a large policy shock as the falling
aggregate welfare measure shows.

**Disaggregate Effects:** While the effects across sectors and regions are contingent upon the prevalence of cash as a payments method, these effects are not uniform across consumption deciles as well due to their differential reliance on cash payments. We analyze the disaggregate effect of the above shock on different groups of people. Figures 6 and 7 show the effect of the same shock in $Q_1$ on the different consumption deciles ranked by their monthly per capita consumption expenditure where 1.0 denotes the highest consumption decile. We demonstrate the effects on unorganized and total output, and unorganized surplus (utility net of costs).

It can be seen from Figure 6 that for the upper deciles, rural output/consumption (both unorganized and total) falls by more than that for urban. This is because the upper urban deciles switch to digital payments as mentioned previously. But, even if consumption of the upper deciles remained unaffected, their unorganized surplus (utility net of costs) fell as they had to pay a higher cost to access payments be it cash or digital (see Figure 7).

Unorganized sector output/consumption falls between 20-30% for the bottom and middle deciles, and slightly more for the lower than middle deciles as seen from the left panel of Figure 6. This is because even though the level of $\gamma$ is the same for all deciles who continue to use cash, the consumption of the lower deciles (i.e. for households that consume less or ones with low $\epsilon^i$) is more responsive to changes in $\gamma$. The fall in unorganized surplus is similar (closer to 20%) for these deciles as seen in Figure 7. It accounts for the higher costs of payments (i.e. higher $\gamma$ for those who did not switch or higher $\kappa$ relative to the pre-shock level of $\gamma$ for those who did) and so is less heterogeneous as it takes into account these costs, which increased across the board, though the middle and lower were more severely impacted than the upper deciles.

The scale on both panels of Figure 6 is the same, so it is easy to see that unorganized output fell by much more than total output for all deciles. And, unlike unorganized output, the fall in total output is most severe for the middle deciles and in fact for these mid and low consumption deciles, the fall is higher for the urban population than rural. The percent of unorganized output as a fraction of total output is higher for the upper urban deciles, and the fall in unorganized output for the middle is only slightly different than the lower deciles. So the former effect dominates and total output for the middle deciles is hit the most. Thus, the higher average fall in total output for the rural region as a whole compared to urban, as
Figure 6: Fall in output across consumption deciles relative to pre-demonetized levels.
Note: 1.0 denotes the highest consumption decile.

As noted previously, the fall in output is driven by the top deciles given their higher shares in total consumption and output. The top deciles, primarily the urban groups, were able to maintain the same level of consumption as before because they could switch to using digital means though at a higher cost.

Finally, we use the monthly per capita consumption expenditure data by population deciles for rural and urban sectors to plot a Lorenz curve in Figure 8 (the left two panels are for rural and right two for urban). The green dotted line is the 45 degree line or the line of perfect equality. The further away the Lorenz curve is from this line, the higher is the inequality. The blue solid line labeled ‘pre’ represents the Lorenz curve before the shock and the orange dashed line is the ‘post’ line for $Q_2$. We find that there is a slight increase in both rural and urban unorganized consumption inequalities owing to the large and differential impact of the shock on the different groups as described above in particular to the middle and lower deciles. Total consumption inequality remains roughly the same due to the lower percent declines.
Figure 7: Percentage Fall in Unorganized Surplus across Regions

Figure 8: Lorenz Curve

Note: The blue solid line ‘pre’ is before the shock and the orange dashed line ‘post’ is $Q_2$.

7 Conclusion

This paper examines the relationship between the choice of payment methods and aggregate economic outcomes. These payment methods involve trade-offs that dictate their selection as means of payments based on preferences and access. We model the choice of payment methods using a tractable monetary framework based on Lagos and Wright (2005) and
Rocheteau and Wright (2005) with preference heterogeneity. We show that the use of cash has two distinct features. First, since there is a carrying cost for using cash, consumers economize on their money holdings when employing cash for payments. Second, since cash led transactions are harder to track, they facilitate tax evasion, which increases the size of a parallel shadow economy. However, the alternative of switching to digital payments involves an entry cost which restricts its usage to consumers above a certain threshold level of consumption. This is under the assumption that there are no infrastructural constraints on access to a digital payments system, which may exist in particular for rural regions in India. This leads to a payment divide – an economic divide emanating from divergence in payment choices – which affects aggregate welfare.

We employ this framework on payment methods with preference heterogeneity and an endogenous shadow economy to understand the impact of unexpected demonetization of India’s two large denomination bills. Since this episode acts as a case of a liquidity and payments system shock, it enables us to draw conclusions of its effect on aggregate output, welfare and the size of the shadow economy. Our calibration captures key features of India’s payment system to conclude that this shock led to an immediate fall in aggregate output by 20%. We disaggregate the effects of this shock based on sectors and consumption deciles to find that the regional divergence in means of payments led to a fall in rural output by twice more than its urban counterpart in the first quarter following the shock.

The one-time demonetization shock not only caused temporary inconveniences but also some more medium term aggregate and distributional consequences. The slow and costly replacement of demonetized bills made the fall in real economic outcomes persistent as well as worsened distributional outcomes contingent on the payments divide. Urban households in the upper deciles could switch to non cash payments more easily while their rural counterparts could not. This is because they could bear the cost of switching to digital means by taking a reduction in their surpluses. We also show that the effects of demonetization and the subsequent slow remonetization created a greater regional divide and higher consumption inequality emanating from a shock to the payments system.
References


A Online Appendix: Proofs

A.1 Proof of Lemma 1

From (11), \( \frac{\partial m_i}{\partial \epsilon_i} > 0 \). To see that, fully differentiate (11) with respect to \( \epsilon_i \) (ignore the time subscript and assume that \( \phi_t^{-1}/\phi_t \) is not affected by \( \epsilon_i \) which will be true in a stationary monetary equilibrium),

\[
u'(\phi_t m_i^t) + \epsilon_i u''(\phi_t m_i^t) \frac{\partial m_i}{\partial \epsilon_i} \phi_t = 0.
\]

Since \( u'(\cdot) > 0 \) and \( u''(\cdot) < 0 \), we get that \( \frac{\partial m_i}{\partial \epsilon_i} > 0 \). \hfill \blacksquare

A.2 Proof of Lemma 2

From (11), \( \frac{\partial m_i}{\partial \gamma} < 0 \) since \( u'(\cdot) > 0 \) and \( u''(\cdot) < 0 \). Setting \( \gamma = 0 \) in (11) we get (17), thus \( m_i^d > m_i^c \). \hfill \blacksquare

A.3 Proof of Lemma 3

Since \( V_i^t = \max\{V_i^t c, V_i^t d\} \) as given by (14), we need to compare (5) and (16). We’ll get \( V_i = \max\{V_i^c, V_i^d\} = V_i^c \) if and only if,

\[
- \phi_{t-1} m_i^c + \beta \{ -\gamma \phi_t m_i^c + \alpha s(\phi_t m_i^c) + \phi_t m_i^c \} > - \phi_{t-1} m_i^d + \beta \{ -\kappa + \alpha s(\phi_t m_i^d) + \phi_t m_i^d \},
\]

where \( m_i^c \) is given by (18) and \( m_i^d \) is given by (19). Note that we defined \( s(x^i) \equiv \epsilon^i u(x^i) - x^i \).

Dividing both sides by \( \beta \) and simplifying,

\[
+ \phi_t m_i^c - \phi_{t-1}(1 + \rho)m_i^c - \gamma \phi_t m_i^c + \alpha s(\phi_t m_i^c) > + \phi_t m_i^d - \phi_{t-1}(1 + \rho)m_i^d - \kappa + \alpha s(\phi_t m_i^d).
\]

Using \( \phi_{t-1}/\phi_t = (1 + \iota)/(1 + \rho) \),

\[
- \iota \phi m_i^c - \gamma \phi m_i^c + \alpha s(\phi m_i^c) > - \iota \phi m_i^d - \kappa + \alpha s(\phi m_i^d).
\]
Thus, for household-\(i\) to be indifferent, the net cost of carrying cash (costs include interest foregone and carrying cost, and benefits include early consumption surplus) has to equal the net cost of carrying digital means (costs include interest foregone and fixed cost, and benefits include early consumption surplus), or \(\bar{\kappa}^i\) will be given by:

\[
-\iota \phi m^{ic} - \gamma \phi m^{ic} + \alpha s(\phi m^{ic}) = -\iota \phi m^{id} - \bar{\kappa}^i + \alpha s(\phi m^{id}).
\]

Using \(\phi m^{ij} = y^{ij}\) for \(j = \{c, d\}\) and substitute back for \(s(x^i)\),

\[
\bar{\kappa}^i = \gamma y^{ic} + \alpha \epsilon^i[u(y^{id}) - u(y^{ic})] - (\iota + \alpha)(y^{id} - y^{ic}).
\]

(ii) To show that \(\bar{\kappa}^i\) is increasing in \(i\), take the derivative of the above with \(\epsilon^i\),

\[
\frac{\partial \bar{\kappa}^i}{\partial \epsilon^i} = \gamma \frac{\partial y^{ic}}{\partial \epsilon^i} + \alpha [u(y^{id}) - u(y^{ic})] + \alpha \epsilon^i \left[ u'(y^{id}) \frac{\partial y^{id}}{\partial \epsilon^i} - u'(y^{ic}) \frac{\partial y^{ic}}{\partial \epsilon^i} \right] - (\iota + \alpha) \left[ \frac{\partial y^{id}}{\partial \epsilon^i} - \frac{\partial y^{ic}}{\partial \epsilon^i} \right],
\]

\[
\frac{\partial \bar{\kappa}^i}{\partial \epsilon^i} = \alpha [u(y^{id}) - u(y^{ic})] + [\alpha \epsilon^i u'(y^{id}) - \iota - \alpha - \gamma] \frac{\partial y^{id}}{\partial \epsilon^i} - [-\gamma + \alpha \epsilon^i u'(y^{ic}) - \alpha - \iota] \frac{\partial y^{ic}}{\partial \epsilon^i} > 0
\]

The last two terms are zero from the first order conditions (19) and (18), and the first term is positive, since \(y^{id} > y^{ic}\) from Lemma 2.

\[\blacksquare\]

**A.4 Proof of Proposition 1**

We prove existence and uniqueness for (iii) partial digital payments, others will follow similarly.

We need to show that if \(\bar{\kappa}^l < \kappa < \bar{\kappa}^m\), where

\[
\bar{\kappa}^i = \gamma y^{ic} + \alpha [s(y^{ld}) - s(y^{lc})] - \iota (y^{ld} - y^{lc}),
\]

\[
\bar{\kappa}^m = \gamma y^{mc} + \alpha [s(y^{md}) - s(y^{mc})] - \iota (y^{md} - y^{mc}),
\]

then \(l\) uses cash and \(m, h\) digital payments where we used we defined \(s(x^i) \equiv \epsilon^i u(x^i) - x^i\).
From Lemma 3 (ii), \( \bar{\kappa}^l < \kappa^m < \bar{\kappa}^h \), so we get \( \bar{\kappa}^l < \kappa < \bar{\kappa}^m < \bar{\kappa}^h \) and,

\[
\alpha[\epsilon^l u'(y^{lc}) - 1] = \iota + \gamma,
\]

\[
\alpha[\epsilon^m u'(y^{md}) - 1] = \iota,
\]

\[
\alpha[\epsilon^h u'(y^{hd}) - 1] = \iota,
\]

\[
\pi^l y^{lc} + \pi^m y^{md} + \pi^h y^{hd} = \phi M.
\]

We need to show that the above four equations solve for \((\phi, y^{lc}, y^{md}, y^{hd}) \in \mathbb{R}_+^4\) and the solution is unique. This is straightforward to see given that \(u''(\cdot) < 0\).

\[\blacksquare\]

### A.5 Proof of Lemma 4

Consider \(j = c\), substitute for \(P(\omega^{ic})\) in (28) to get,

\[
\alpha \left[ \epsilon^{ic} u' \left( y^{ic} (1 - \tau) + \frac{1}{\lambda} \right) - 1 \right] \left[ 1 - \tau + \frac{1}{y^{ic} \lambda} \right] = \iota + \gamma,
\]

and we have \(\frac{\partial y^{ic}}{\partial \epsilon^{ic}} > 0\). Fully differentiate (38) with respect to \(\epsilon^{ic}\),

\[
u' \left( y^{ic} (1 - \tau) + \frac{1}{\lambda} \right) \left[ 1 - \tau + \frac{1}{y^{ic} \lambda} \right] = \frac{\partial y^{ic}}{\partial \epsilon^{ic}} \left\{ -\epsilon^{ic} u'' \left( y^{ic} (1 - \tau) + \frac{1}{\lambda} \right) (1 - \tau) + \frac{\left[ \epsilon^{ic} u' \left( y^{ic} (1 - \tau) + \frac{1}{\lambda} \right) - 1 \right]}{y^{ic} \lambda} \right\},
\]

and it can be seen that the term on the left hand side above is positive and so is the term in curly braces on the right hand side given that \(u''(\cdot) < 0\). Hence, \(\frac{\partial y^{ic}}{\partial \epsilon^{ic}} > 0\). From (31), we get that \(\frac{\partial \bar{\tau}^{ij}}{\partial y^{ij}} > 0\) so \(\bar{\tau}^{ij} > \bar{\tau}^{kj}, j \in \{c,d\}\) if \(\epsilon^i > \epsilon^k\).

(ii) Similarly, from (38), we also get \(\frac{\partial y^{ic}}{\partial \gamma} < 0\) and from (31) \(\frac{\partial \bar{\tau}^{ic}}{\partial \gamma} < 0\). This implies \(\bar{\tau}^{id} > \bar{\tau}^{ic}\) (using \(\gamma = 0\)).

\[\blacksquare\]
A.6 Proof of Lemma 5

(i) As before we get, $V^i = \max\{V^{ic}, V^{id}\} = V^{ic}$ if and only if,

$$-\phi_{t-1}m^{ic} + \beta\{-\gamma\phi_t m^{ic} + \alpha s (\phi_t m^{ic}[1 - P(\omega^{ic})\tau]) + \phi_t m^{ic}\} >$$

$$-\phi_{t-1}m^{id} + \beta\{-\kappa + \alpha s (\phi_t m^{id}[1 - P(\omega^{id})\tau]) + \phi_t m^{id}\},$$

where $m^{ic}$ is given by (18) and $m^{id}$ is given by (19). We also defined $s_i(x) \equiv \epsilon_i(u(x) - x)$. Rest of the steps follow from Lemma 3 (i).

(ii) The proof is the same as Lemma 3. ■

A.7 Proof of Proposition 2

(i) It is straightforward to see that $y^l$, $y^m$ will fall and $y^h$ will remain unchanged. Hence, $W^l$ and $W^m$ fall and so will $\mathbb{W}$.

(ii) From Proposition 5 and Proposition 1, if $\tilde{\kappa}^l(\gamma_1) < \kappa_0 < \tilde{\kappa}^m(\gamma_1)$ then a partial digital payments equilibrium exists with $\pi_d = \pi^h + \pi^m$. So, from Lemma 2 $y^m$ increases, and $y^l$ falls as $\frac{\partial y^l}{\partial \gamma} < 0$. Welfare, $W^m$ falls as compared to before the shock as shown next. Denote $W^m_0 = U(q^m) - l^m - \gamma_0 y^m_{r,0} + \alpha[u(y^m_{r,0}) - y^m_{r,0}]$ as the pre-shock welfare. We know that $W^m_0 > W^md = U(q^m) - l^m - \kappa + \alpha[u(y^md) - y^md]$ as type-$m$ used cash instead of digital means. Given that $\tilde{\kappa}^l(\gamma_1) < \kappa_0 < \tilde{\kappa}^m(\gamma_1)$, we also have that $W^md > W^m_1 = U(q^m) - l^m - \gamma_0 y^mc_{r,1} + \alpha[u(y^mc_{r,1}) - y^mc_{r,1}]$ from Lemma 5. The fall in $W^l$ is straightforward to see. And, that $W^h$ does not change.

(iii) (a) is straightforward. (b) If $\gamma = \gamma_0$ and $\kappa_1 < \kappa_0$ such that $\tilde{\kappa}^l(\gamma_0) < \kappa_1 < \tilde{\kappa}^m(\gamma_0)$ then Proposition 1 implies that a partial digital payments equilibrium exists with $\pi_d = \pi^h + \pi^m$. So, from Lemma 2 $y^m$ increases, welfare, $W^m$ rises as compared to before the shock as cost of payments falls as well in this case. Similarly, $W^h$ rises and aggregate welfare rises as well as tax revenues increase as output is higher and compliance is higher. Finally, size of the shadow $S$ as given by the simplified version in (34) falls as $\sum_{i \in I} \pi^i y^i$ rises. ■