MONASH University

MONASH BUSINESS SCHOOL

Department of Economics
ISSN number 1441-5429

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

Discussion Paper no. 2021-15

Ayushi Bajaj and Nikhil Damodaran

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 analyze India's sudden demonetization of 86% of the cash in circulation with new notes gradually being replaced over the next several months. The welfare cost of this liquidity shock was equivalent to 1% of total consumption. Even though all consumers experienced a decline in welfare, its extent varied depending on the degree of cash dependence and the ability to switch to non-cash payments. The middle consumption deciles were disproportionately affected.

Keywords: Money, Payments, Shadow economy, Demonetization, Monetary policy

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

Ayushi Bajaj: Monash University (email: ayushi.bajaj@monash.edu); Nikhil Damodaran: O.P. Jindal Global University (email: ndamodaran@jgu.edu.in).

Consumer Payment Choice and the Heterogeneous Impact of India's Demonetization*

Ayushi Bajaj
Monash University

Nikhil Damodaran
O.P. Jindal Global University

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 analyze India's sudden demonetization of 86% of the cash in circulation with new notes gradually being replaced over the next several months. The welfare cost of this liquidity shock was equivalent to 1% of total consumption. Even though all consumers experienced a decline in welfare, its extent varied depending on the degree of cash dependence and the ability to switch to non-cash payments. The middle consumption deciles were disproportionately affected.

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

Keywords: Money, Payments, Shadow economy, Demonetization, Monetary policy

*We thank Guillaume Rocheteau, Pedro Gomis-Porqueras, Pushkar Maitra, Almuth Scholl, Mohammed Ait Lahcen, Randall Wright, Anirudh Tagat, Nikita Naik and discussants Makoto Watanabe and Raghbendra Jha for helpful discussions on earlier versions. We also thank participants at the Bank of Finland; Midwest Macro Workshop, Wisconsin; 5th HenU/INFER Macro Workshop, China; Deakin University, Monash University, University of Queensland; Summer School in Development Economics, Italy; Australasian Economic 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. The use of these instruments involves trade-offs which influence a consumer's payment choice. For consumers, carrying several small bills of cash is cumbersome while larger bills are more susceptible to counterfeiting. On the other hand, non-cash digital payments are usually devoid of such costs. However, using them requires a bank account in the least, which involves usage fees. Further, anonymous cash payments facilitate tax evasion whereas a shift towards digital payment methods generates a paper trail of tractable transactions. These factors together influence aggregate outcomes including the size of the shadow economy. It also creates a payments divide based on access and adoption, which has distinct distributional consequences.

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. This allows us to examine the consequences of this choice on consumption distribution, aggregate output, welfare and the shadow economy. If consumers transact in cash, they economize on their money holdings because they face a marginal carrying cost. However, the alternative of making non-cash payments is limited by usage costs, higher effective taxation and infrastructural constraints.

We apply this framework to analyze and quantify the heterogeneous impact of a unique monetary episode in India which led to a payments system shock. On November 8, 2016 the Government of India unexpectedly demonetized the two largest denomination bills comprising 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 over the next several months. This large liquidity shock occurred in an otherwise stable macroeconomic environment and led to an immediate fall in aggregate output and welfare, as consumers were unable to undertake routine cash transactions. This aggregate impact has been analyzed and quantified in recent studies of this monetary episode using alternative approaches including Chodorow-Reich et al. (2020). In this paper, we delve into the mechanisms behind its distributional consequences on heterogeneous consumers by

explicitly modeling their payment choices. We find that the aggregate welfare impact of the slow remonetization is comparable to that of 27.6% inflation. However, unlike inflation which most adversely impacts the top consumption deciles, this monetary shock is felt most severely by the middle deciles.

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. In addition, we assume that money can be held in two forms. The first form is cash, which itself is available in two denominations involving a trade-off between carrying cost and counterfeiting – each low denomination bill is less susceptible to be counterfeited but it is cumbersome to carry many small bills. The other is a non-cash digital form of payment which is easy to carry and cannot be counterfeited but it incurs a usage cost, independent of transaction size. For instance, consumers need a bank account with fixed operational fees to access instruments such as debit cards.

The second key feature of our model is heterogeneity on consumer preferences. We find that it is typically not worthwhile to invest in non-cash payments for consumers with a lower level of consumption and they end up using cash. For instance in India, rural monthly consumption averaged Rs. 1,430 as opposed to urban at Rs. 2,630 for 2011-12. Thus, the model would suggest that rural consumers undertake more cash transactions as their optimal response. This is corroborated by the fact that rural bank deposits comprise only about 10% of the total bank deposits in India, which indicates a significantly lower usage of non-cash payments in rural areas. This finding is further backed by empirical studies that document a predominance of cash payments for relatively smaller transactions such as Tagat et al. (2019) for India and Runnemark et al. (2015) for the US. Modeling consumption distribution via preference heterogeneity is a tractable way to capture such differences in payment methods and relative cash dependence, which helps analyze the differential impact of an aggregate demonetization shock.

Third, we assume that the sales tax levied by the government cannot be perfectly enforced. This allows us to analyze the link between payment choice, tax enforcement and the tax-evading shadow economy. Gomis-Porqueras et al. (2014) also presents a monetary model where a shadow economy emerges endogenously by assuming that cash transactions are not

subject to a sales tax. In contrast, we allow the level of enforcement to be endogenously determined as we explicitly model the effort to monitor tax collection. We find that the effective tax rate (i.e. after adjusting for probability of paying) on cash transactions is lower than on non-cash. This is because smaller transactions are often done in cash, so the tax enforcement authority exerts a lower level of effort to monitor them. These three features of the model as outlined above, make it well-suited to analyze this monetary episode, as the model explicitly addresses why cash matters and what determines its choice over non-cash payments.

We calibrate the model parameters to match key features of the Indian economy and determine the quantitative impact of consumers suddenly finding their high denomination bills to be unacceptable for transactions. We then ascertain the aggregate and disaggregate impact of the slow and costly remonetization that followed by adjusting the cost of payments to match the pace of currency replacement. For our exercise, we do not vary the rate of redistribution of new notes across households as in Chodorow-Reich et al. (2020). Instead we study the response of heterogeneous consumers to the same liquidity shock based on their ability to switch to alternative payment methods.

To measure its impact, we calculate the welfare cost of the policy shock. Our measure of the cost asks how much consumers would be willing to give up in terms of total consumption in order to go back to the ease of pre-demonetization payment systems. This cost was 1.3% of consumption owing to the slow and costly remonetization process. In contrast, the welfare cost of 10% inflation (as compared to 0% inflation) is 0.3% of consumption and that of 27.6% inflation is 1.3%. In terms of household level impact, we find that the payments system shock led to a decline in household welfare for every group in every region but the magnitude of this fall varied. The impact was felt most by consumers with a high cash dependence who were unable to switch to non-cash means i.e. those in the middle consumption deciles. Consumers in the top deciles were less affected because they either previously used non-cash means of payments or later found it worthwhile to transition away from cash following the shock. In contrast, inflation affects households carrying higher money balances the most i.e. ones in the top consumption deciles.

1.1 Related Literature

The monetary framework builds on Lagos and Wright (2005) and Rocheteau and Wright (2005). We extend the baseline setup to incorporate dual means of payments with a simplified denomination structure for cash using insights from Nosal and Rocheteau (2011) and Lee et al. (2005). Some other papers that model multiple means of payments include Li (2011) that finds that checks are used only in big transactions while cash is used in all transactions. 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 is borne by merchants. Kim and Lee (2010) presents a model of debit cards where sellers bear a fixed record-keeping cost regardless of transaction size. Zhu and Hendry (2019) and Williamson (2019) present models with multiple means of payment to analyze the effects of introducing central bank digital currency.

We also contribute to the literature on shadow economies by examining the link between tax enforcement and payment choice. As mentioned earlier, Gomis-Porqueras et al. (2014) also explores the endogenous emergence of a shadow economy but with exogenous tax enforcement. 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 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) 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 literature. 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 deployment of new notes to find that districts experiencing slower remonetization had relatively larger reductions in economic activity, faster adoption of alternative payment technologies, and lower bank credit growth. Karmakar and Narayanan (2019) uses a panel dataset on Indian households to find that the 17 percent of households who did not have bank accounts experienced 2 to 7 percent lower consumption than the control group of households with bank accounts. Agarwal et al. (2019) finds evidence for increased use digital payments following the shock which can induce over-spending.

Crouzet et al. (2019) finds evidence for network effects in adoption of electronic payments by retailers which reduced the costs of the demonetization shock especially for groups that had higher adoption rates prior to the shock. We do not consider firm's/retailer's decision to accept different payments, but in our analysis of demonetization we allow for increased technological development which could reduce the exogenous costs of using non-cash/digital means. Chanda and Cook (2019) uses geographical variation in deposits to identify the effects of the shock. Wadhwa (2019) uses consumer pyramids data to find empirical effect of the shock on consumption and find a higher decline in consumption for richer than poorer households. Some other papers that analyze this episode from different angles include Waknis (2017), Agrawal (2018) and Tagat and Trivedi (2020).

2 Model

The economy is populated by households, firms and a government/central bank with a consolidated budget constraint. Time is discrete and continues forever, each time period is denoted by $t \in \{0, 1, 2, ...\}$. Households supply labor l_t to firms and consume two goods: a general good q_t consumed at the end of every time period and a special good y_t to be consumed earlier with probability α. The general good will serve as the numéraire. This consumption sequence is based on the centralized and decentralized market structure in Lagos and Wright (2005) and Rocheteau and Wright (2005).

Households differ from each other based on the utility they receive from consumption of the special good. Household type-i's utility is given by $\epsilon^i u(y_t^i)$, where $i \in \mathbb{I}$. We assume that with probability α, $\epsilon^i > 0$ for household-i and $\epsilon^i = 0$ with complementary probability $1 - \alpha$. The proportion of household type-i is given by π^i. The lifetime discounted expected utility of household type-i is:

$$\mathbb{E} \sum_{t=0}^{\infty} \beta^t [\alpha \epsilon^i u(y_t^i) + U(q_t^i) - l_t^i], \tag{1}$$

where $\beta \equiv (1 + \rho)^{-1} \in (0, 1)$ is the discount factor between periods, $U(q_t^i)$ is the utility from consumption of the general good and the dis-utility from labor l_t^i 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 $U(\cdot)$.

Firms operate in a perfectly competitive market. 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 production of the special good, then their 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 along with profits, Δ_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 households/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.

The central bank controls the total (fiat) money supply as given by M_t which grows at a constant rate, Π. The price of money in terms of the numéraire is denoted by ϕ. The fiat money is available as cash which incurs an additional carrying cost of γ per bill.[1] There is an alternative non-cash or digital/electronic means of payment, like a debit card which can be used as a medium of exchange just as cash except that it will not incur the carrying cost and cannot be counterfeited. But, there is a fixed (i.e. independent of the amount) usage/access fee for holding the digital means κ which falls on households (similar to fees charged for maintaining a bank account/debit card). So, households decide how much money balances to hold each period denoted by m_t and then decide whether to carry it as cash, m_t^c or as non-cash/ digital, m_t^d.[2]

[1] In Appendix A, we also consider a simple denomination structure for cash. The low is fully divisible, and the high denomination is available in $k > 1$ units of the low denomination. The central bank adjusts the supply of each denomination as per demand. Besides, these bills can also be counterfeited at nominal cost $\delta > 1$ per bill during the late consumption period.

[2] This payments structure captures the two major components of the money supply measure, M1 - currency or cash and demand deposits or non-cash payments. Both forms of payments face the opportunity cost of holding money as compared to other interest bearing assets. However, only cash usage incurs further carrying costs and counterfeits depending on the number and type of bills held, while operating a bank account has some associated usage costs regardless of the size of deposits. Since demand deposits are increasingly used to make digital payments via debit cards and mobile wallets, we refer to non-cash payments as digital. Note however that in India the use of paper checks is still quite prevalent (RBI, 2019).

The government levies sales tax τ on goods consumed. In the early consumption period since transactions are anonymous, the government cannot perfectly enforce taxes on special goods, and has to exert effort ω to increase enforcement. This affects the probability of successfully enforcing tax payments or the probability that taxes are paid $P(\omega)$. We assume that $P'(\omega) > 0$, $P''(\omega) < 0$, $P(0) = 0$, $P(\infty) = 1$. An example of a function which satisfies the above properties is: $P(\omega) = (1 - \exp^{-\lambda\omega})$ with $\lambda \geq 0$ sufficiently large. In the late consumption period there are no such frictions so tax collection can be perfectly enforced without additional effort. The consolidated government and central bank budget constraint in period-t is given by

$$\mathbb{T}_t = \phi_t M_{t+1} - \phi_t M_t + \tau \sum_i \pi^i [\alpha P(\omega_t^i) p_t y_t^i + q_t^i], \quad \text{or} \quad \mathbb{T}_t = \Pi \phi_t M_t + \tau \sum_i \pi^i [\alpha P(\omega_t^i) p_t y_t^i + q_t^i], \tag{2}$$

where $\mathbb{T}_t$ is the real value of lump-sum transfer from the government to households which adjusts every period to maintain equality.

3 Equilibrium

Households

Let $W_t(\cdot)$ denote the value function of households at the beginning of the late consumption sub-period in period t. In each such sub-period, the state variable is the current money balance held by household-i m_t^i. Household-i chooses its level of general goods consumption q_t^i and labor supply l_t^i. Money balance m_{t+1}^i is carried to the next early consumption sub-period. There is discounting between these two sub-periods so the continuation value of early consumption denoted by V_{t+1} is multiplied by β. This gives us the following problem for household-i at the beginning of the late consumption sub-period:

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

s.t.

$$q_t^i = l_t^i w_t - \tau q_t^i + \Delta_t + \mathbb{T}_t^i + \phi_t m_t^i - \phi_t m_{t+1}^i,$$

where w_t is wages per unit of labor supplied, τ is tax rate on goods consumed, Δ_t is firm profits transferred to households, $\mathbb{T}_t$ is lumpsum transfer by government and ϕ_t is the value

of money in terms of the general good. Substitute for l_t^i and normalize $w_t = 1$ (equivalently one can adjust the disutility from labor in the value function) to get the following problem:

$$W_t(m_t^i) = \Delta_t + \mathbb{T}_t + \phi_t m_t^i + \max_{q_t^i}\{U(q_t^i) - q_t^i(1+\tau)\} + \max_{m_{t+1}^i}\{-\phi_t m_{t+1}^i + \beta V_{t+1}(m_{t+1}^i)\}. \quad (3)$$

The above simplification makes the problem for money holdings tractable, which follows from the quasi-linear structure of preferences. To solve the portfolio decision problem for m_{t+1}^i that determines the total money balances held at the end of every period we need to know $V_{t+1}(m_{t+1}^i)$ which depends on the choice of means of payment.[3] We use superscripts c and d to denote money held as cash and digital means respectively, so each household type-i chooses her money balance for next period as cash denoted by m_{t+1}^{ic} or non-cash/digital, m_{t+1}^{id}.

If household-i carries cash then we get the value of early consumption as,

$$V_t^c(m_t^{ic}) = -\gamma\phi_t m_t^{ic} + \alpha \max_{y_t^{ic}} \left\{\epsilon^i u(y_t^{ic}) + W_t\left(\phi_t m_t^{ic} - (p_t + P_t^i\tau)y_t^{ic}\right)\right\} + (1-\alpha)W_t\left(\phi_t m_t^{ic}\right), \quad (4)$$

s.t.

$$(p_t + P_t^i\tau)y_t^{ic} \leq \phi_t m_t^{ic}.$$

The carrying cost γ of each bill held has to be subtracted from this value, and if household-i receives an early consumption shock with probability α, they consume y^{ic}. They pay cash to firms at price p and pay taxes at rate $P_t^i\tau$. With complementary probability $(1-\alpha)$ i.e. if the household does not receive the shock then, the bills are carried forward to the late consumption period. Here p_t is the price of the special good in terms of the numéraire, and probability P is the likelihood of being caught not paying taxes which is a function of the enforcement effort ω_t^i of the tax authority. We will assume that buyers simply face an effective tax rate of $P(\omega_t^i)\tau$.

We simplify the above value $V_t(\cdot)$ as follows using $W_t(\phi_t m_t^{ic}) = W_t(0) + \phi_t m_t^{ic}$:

[3] As shown in Appendix A we can replace the portfolio choice problem for cash with two denominations with a problem that only includes the low denomination. Thus, if the household decides to use cash it is sufficient to solve the problem with the low denomination only. For our quantitative results in Section 4, the two denominations will feature which helps calibrate the cost parameter γ by using available data on counterfeits and also for quantifying the impact of demonetizing only the large denomination bills.

$$V_t^c(m_t^{ic}) = -\gamma\phi_t m_t^{ic} + \alpha \max_{y_t^{ic}}[\epsilon^i u(y_t^{ic}) - (p_t + P_t^i\tau)y_t^{ic}] + \phi_t m_t^{ic} + W_t(0), \tag{5}$$

s.t.

$$(p_t + P_t^i\tau)y_t^{ic} \leq \phi_t m_t^{ic}.$$

Now to get the households choice of money holdings (given that they carry it as cash), we plug $V_t^c(m_t^{ic})$ from above in to (3). We get the following portfolio choice maximization problem for a household-i carrying cash:

$$\max_{m_t^{ic}} \left\{ -\phi_{t-1} m_t^{ic} + \beta \left\{ -\gamma\phi_t m_t^{ic} + \alpha \max_{y_t^{ic} \leq \mathbb{C}_t} [\epsilon^i u(y_t^{ic}) - (p_t + P_t^i\tau)y_t^{ic}] + \phi_t m_t^{ic} \right\} \right\}, \tag{6}$$

where $\mathbb{C}_t \equiv [\phi_t/(p_t + P_t^i\tau)]m_t^{ic}$ is the constraint on early consumption. Each household takes the prices p_t, ϕ_t as given and maximizes the above. We first obtain the optimal early consumption demand by households y_t^{ic} given ϕ_t and p_t by solving $\max_{y_t^{ic}}[\epsilon^i u(y_t^{ic}) - (p_t + P_t^i\tau)y_t^{ic}]$ subject to $(p_t + P_t^i\tau)y_t^{ic} \leq \phi_t m_t^{ic}$. We get that,

$$y_t^{ic} = \min \left\{ \frac{\phi_t m_t^{ic}}{(p_t + P_t^i\tau)}, u'^{-1}((p_t + P_t^i\tau)/\epsilon^i) \right\}. \tag{7}$$

We now solve the money demand problem (6) for household-i carrying cash m_t^{ic} (assuming an interior solution) to get:

$$-\phi_{t-1} + \beta\phi_t \left\{ -\gamma + \alpha \left[\frac{\epsilon^i}{(p_t + P_t^i\tau)} u' \left(\frac{\phi_t m_t^{ic}}{(p_t + P_t^i\tau)} \right) - 1 \right] + 1 \right\} = 0, \tag{8}$$

where we use $y_t^{ic} = \phi_t m_t^{ic}/(p_t + P_t^i\tau)$ from (7) i.e. households do not bring more real balances than what they need in trade as money is costly to hold. Given prices $\phi_t > 0$, $p_t > 0$ and tax enforcement effort $\omega_t \geq 0$, we get that money holding for any household-i is increasing in ϵ^i and decreasing in γ. Later we will see that this result holds true when prices and effort levels are endogenous as well.

Next, if instead of cash, household-i carries money in digital means, then

$$V_t^d(m_t^{id}) = -\kappa + \alpha \max_{y_t^{id}} \{\epsilon^i u(y_t^{id}) + W_t(\phi_t m^{id} - (p_t + P_t^i \tau) y_t^{id})\} + (1-\alpha) W_t(\phi_t m_t^{id}), \quad (9)$$

s.t.

$$(p_t + P_t^i \tau) y_t^{id} \leq \phi_t m_t^{id}.$$

The fixed usage cost of digital means κ has to be subtracted from this value and the net utility from early consumption is the same as in (4). We use the same intermediate steps used to obtain (6) to derive household-i's portfolio choice problem if she carries digital means, which becomes:

$$\max_{m_t^{id}} \{-\phi_{t-1} m_t^{id} + \beta\{-\kappa + \alpha \max_{y_t^{id} \leq \mathbb{C}_t} [\epsilon^i u(y_t^{id}) - (p_t + P_t^i \tau) y_t^{id}] + \phi_t m_t^{id}\}\}. \quad (10)$$

where $\mathbb{C}_t \equiv \phi_t m_t^{id} / (p_t + P_t^i \tau)$. The optimal early consumption demand by households remains the same as before and is given by (7). We can now solve the portfolio choice problem (10) for a household-i carrying digital means m_t^{id} (assuming an interior solution) to get:

$$-\phi_{t-1} + \beta \phi_t \left\{ \alpha \left[\frac{\epsilon^i}{(p_t + P_t^i \tau)} u' \left(\frac{\phi_t m_t^{id}}{(p_t + P_t^i \tau)} \right) - 1 \right] + 1 \right\} = 0, \quad (11)$$

where we use $y_t^{id} = \phi_t m_t^{id} / (p_t + P_t^i \tau)$ as before.

The key difference between (8) and (11) is the presence of carrying cost of cash, γ in the former. Thus, cash households carry less money balances and consume less goods in the early consumption period. But they also do not have to pay the fixed cost κ (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.

Finally, the choice between carrying money as cash or digital means depends on which gives the maximum value so we get,

$$V_t = \max\{V_t^c(m_t^{ic}), V_t^d(m_t^{id})\}. \quad (12)$$

Firms

We now solve the problem of each perfectly competitive firm in the early consumption period. Firm's expected revenue in terms of the numéraire (i.e. the late consumption good in period-t) is,

$$z_t = \bar{z} + \max_{y_t^s}[-c(y_t^s) + p_t y_t^s]. \tag{13}$$

Each firm produces $\bar{z}$ units of general (or late consumption) good at the end of each period, and it can also speed up production at cost $c(\cdot)$ to produce the special goods y_t^s for early consumption by households. The firm maximization problem for y_t^s gives $p_t = c'(y_t^s)$, and under linear cost i.e. $c(y) = y$, we get $p_t = 1$ which implies $z_t = \bar{z}$.

Tax enforcement effort

We will now derive the optimal level of tax enforcement agent's effort ω_t which will give us the effective tax rate (i.e. after adjusting for the probability of paying taxes), $P(\omega_t)\tau$. The tax enforcement agent takes the tax rate τ and the early consumption output of household-i, y_t^i as given and maximizes tax revenue net of (linear) cost of effort ω^i, to solve the following problem (suppress the time subscript),

$$\max_{\omega^i}\{P(\omega^i)\tau y^i - \omega^i\}. \tag{14}$$

The first-order condition gives,

$$\omega^{i\star} = P'^{-1}\left(\frac{1}{\tau y^i}\right). \tag{15}$$

The optimal enforcement effort level $\omega^{i\star}$ thus depends on the tax rate τ and output y^i of household-i. A higher tax rate would lead to higher enforcement from (15) but it also leads to lower early consumption output from (8) which lowers the incentives to enforce taxes. Thus, it is unclear whether enforcement is higher or not with a higher tax rate. However, as will be seen below any other parameter that affects output y^i of household-i will also affect enforcement effort level ω^i for the household, hence we add superscript-i to ω.

Market clearing

The money market clearing condition in the late consumption period implies that real money demand and supply are equal. Suppressing the time subscript this implies,

$$\sum_{i\in\mathbb{I}} \left(\pi^{ic}\phi m^{ic} + \pi^{id}\phi m^{id}\right) = \phi M. \tag{16}$$

3.1 Steady state equilibrium

We will now focus on the stationary monetary equilibria where $1+\Pi \equiv M_t/M_{t-1} = \phi_{t-1}/\phi_t$ 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.

In the steady state, the first-order condition that determines the choice of real balances held by households carrying cash (8) can be simplified as follows:

$$\frac{\epsilon^i}{1+P(\omega^i)\tau}u'\left(y^{ic}\right) - 1 = \frac{\iota+\gamma}{\alpha}, \tag{17}$$

where $y^{ic} = \phi m^{ic}/[1+P(\omega^i)\tau]$. For households carrying digital means of payment in the steady state, (11) becomes:

$$\frac{\epsilon^i}{1+P(\omega^i)\tau}u'\left(y^{id}\right) - 1 = \frac{\iota}{\alpha}, \tag{18}$$

where $y^{id} = \phi m^{id}/[1+P(\omega^i)\tau]$. The above two first-order conditions equate the households marginal utility of consumption to its cost. The monetary wedge between the households marginal utility of consumption and price of the good is equal to $(\iota+\gamma)/\alpha$ in (17) and ι/α in (18) .

The probability of paying taxes $P(\omega^i)$ is given by (15) and the market clearing condition by (16). And, whether household-i carries cash or digital means is determined as in (12), which depends on a number of factors. We will characterize this choice based on the cost of holding digital means κ by obtaining a threshold on this cost, $\bar{\kappa}^i$ above which household-i carries cash as described in Lemma 1. If $\kappa = \bar{\kappa}^i$, then household-i is indifferent between holding cash and digital, and if $\kappa > \bar{\kappa}^i$, then prefers cash. The threshold cost for using non-cash payments will be strictly greater than the carrying cost of cash because there has to be an adjustment term for the net benefit of additional output obtained from using non-cash means. Hence, $\bar{\kappa}^i > \gamma\phi m^{ic}$ as shown in the lemma below.

Lemma 1 (Thresholds)**.** *(i) Household-i holds cash if and only if the cost of using non-cash means of payment $\kappa > \bar{\kappa}^i$, where $\bar{\kappa}^i$ is given by:*

$$\bar{\kappa}^i = \overbrace{\gamma y^{ic}(1+P^i\tau)}^{\text{cash carrying cost}} + \overbrace{\alpha\epsilon^i[u(y^{id})-u(y^{ic})]}^{\text{benefit from additonal output}} - \overbrace{(\iota+\alpha)[y^{id}(1+P^i\tau)-y^{ic}(1+P^i\tau)]}^{\text{cost of additional output}}. \quad (19)$$

To see how $\bar{\kappa}^i$ varies across household types i, first consider a special case where the probability of paying taxes P^i is exogenous. In this case, it can be easily verified that $\bar{\kappa}^i$ is increasing in ϵ^i. This implies that if type l with preference ϵ^l prefers to hold digital means of payment, then type h with preference $\epsilon^h > \epsilon^l$ also prefers the same i.e. when $\kappa < \bar{\kappa}^l < \bar{\kappa}^h$. Intuitively, if the probability of paying tax is the same irrespective of transaction size, then households with higher preference for early consumption also do not find it optimal to use cash if the ones with lower preference do not. This follows from the fixed cost structure of digital payments.

However, when tax enforcement varies with transaction size, as given by (15), the higher types might find it optimal to hold cash as tax enforcement is lower when undertaking the smaller cash transactions. To illustrate this, consider types l, h with $\epsilon^l < \epsilon^h$ as before. Assume that l prefers to hold non cash means of payment or $\kappa < \bar{\kappa}^l$. Now we need to see if this necessarily implies that h will prefer that as well i.e. is $\kappa < \bar{\kappa}^h$ given $\kappa < \bar{\kappa}^l$. If yes then, $y^{ld} < y^{hd}$ which implies that enforcement for type-h's transaction will also be higher i.e. $\omega^{ld} < \omega^{hd}$. But, it is possible that if type-h holds cash then the tax enforcement effort is lower for her or $\omega^{hc} < \omega^{ld} < \omega^{hd}$. If it is substantially lower, then at the current κ, household h might prefer the smaller cash transaction with a lower effective tax, while l prefers to pay the fixed cost of using digital means of payment for the larger digital transaction. If this is the case then, $\bar{\kappa}^h < \kappa < \bar{\kappa}^l$ given $\epsilon^l < \epsilon^h$.

For the analysis that follows, we will assume that there are three types of households, l, m, h 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. There will be four possible equilibrium regimes as defined below.

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

(i) only cash payments, $\pi_d = 0$ *as a tuple* $(\phi, y^{lc}, y^{mc}, y^{hc}, \omega^l, \omega^m, \omega^h) \in \mathbb{R}^7_+$ *where* y^{lc}, y^{mc}, y^{hc} *solve* (17)*,* $\omega^l, \omega^m, \omega^h$ *solve* (15) *and* ϕ *is derived from* (16)*,*

(ii) partial cash payments, $\pi_d = \pi^h$ *as a tuple* $(\phi, y^{lc}, y^{mc}, y^{hd}, \omega^l, \omega^m, \omega^h) \in \mathbb{R}^7_+$ *where* y^{lc}, y^{mc} *solve* (17)*,* y^{hd} *solves* (18)*,* $\omega^l, \omega^m, \omega^h$ *solve* (15) *and* ϕ *is derived from* (16)*,*

(iii) partial cash payments, $\pi_d = \pi^h + \pi^m$ *as a tuple* $(\phi, y^{lc}, y^{md}, y^{hd}, \omega^l, \omega^m, \omega^h) \in \mathbb{R}^7_+$ *where* y^{lc} *solves* (17)*,* y^{md}, y^{hd} *solve* (18)*,* $\omega^l, \omega^m, \omega^h$ *solve* (15) *and* ϕ *is derived from* (16)*,*

(iv) no cash payments, $\pi_d = 1$ *as a tuple* $(\phi, y^{ld}, y^{md}, y^{hd}, \omega^l, \omega^m, \omega^h) \in \mathbb{R}^7_+$ *where* y^{lc}, y^{md}, y^{hd} *solve* (18)*,* $\omega^l, \omega^m, \omega^h$ *solve* (15) *and* ϕ *is derived from* (16)*.*

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

We will now present some results for our model starting with comparative statics. As discussed earlier, early consumption output and enforcement effort level are decreasing in the cost of carrying cash. And, they are both higher for types with a higher preference shock ϵ^i.

Lemma 2 (Comparative Statics)**.** *(a) Early consumption output* y^i *for any household type-i varies with the cost of carrying cash* γ *and preference* ϵ^i *as follows: (i)* $\partial y^i/\partial\gamma < 0$ *and, (ii)* $\partial y^i/\partial\epsilon^i > 0$*. (b) The enforcement effort* $\omega^i \geq 0$ *varies as follows: (i)* $\partial\omega^i/\partial\gamma < 0$ *and, (ii)* $\partial\omega^i/\partial\epsilon^i > 0$*.*

Due to the difference in tax enforcement agent's efforts ω^{ij}, the effective tax rate $P^{ij}\tau$ will vary across households, where superscript i denotes their preference and $j \in \{c, d\}$ their choice of means of payments. The effective tax rate is progressive in nature as it is higher for households who consume more. Also note that $\gamma = 0$ when households do not use cash, so the effective tax rate is higher on non-cash payments. Alternatively, we could also explicitly assume that non-cash transactions are easier to enforce by letting the probability of enforcement vary explicitly by type of payment. Instead the probabilities are just taken to be functions of the level of effort which turns out to be higher for digital transactions.

3.2 Demonetization, slow remonetization and inflation

In an overnight surprise move in November 2016, the two highest denomination bills in India were demonetized, as a result of which 86% of currency in circulation ceased to be

legal tender. Following the sudden unanticipated shock, the demonetized bills had to be replaced with new ones. But, this remonetization process was slow and it took several months to get the money supply back to its trend growth path, as shown in Figure 1 below. The institutional realities of the time meant that cash, as a means payment, 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. And, people who used non-cash means of payments also got affected as it became cumbersome and costlier to access banks and ATMs.

We implement this policy shock as follows: at the beginning of the early consumption period-t, the central bank/government announces that the current cash in circulation will cease to be legal tender. In the model this implies that cash will no longer be readily redeemable for general goods in the following late consumption sub-period. The old bills can however be exchanged for new ones in that period, and money will continue to grow at the rate Π. But to capture the process of slow and costly remonetization in the model, we let the carrying cost of cash γ increase in the following period-$t+1$. Denomination specific demonetization is considered for the quantitative exercise as discussed in Appendix A.

In essence, this monetary intervention makes money costly to hold and it will be useful to compare its impact with that of another common monetary policy tool - changes to the nominal interest rate ι or inflation rate Π. We will discuss their effect on aggregate and household level outcomes including output, welfare and the size of the shadow economy. Aggregate early consumption output is $\alpha \sum_i \pi^i y^i$. From Lemma 2 (a), this output is maximized in the no cash payments equilibrium and is decreasing in inflation Π and cost of using cash γ.

In response to the demonetization policy announcement in period-t, firms no longer accept cash in exchange for special goods, as then they will end up with worthless pieces of paper in the following sub-period. This leads to a fall in output y_t^i for households who carried cash. Households carrying unused cash bills can still exchange them for new ones in the late consumption period-t which they potentially value for early consumption in the following sub-period $t+1$. Slow and costly remonetization in the model captured by an increase in γ, leads to a fall in money holdings and consumption for households carrying cash in $t+1$ as shown in Lemma 2 (a). But, the rise in γ also leads to an increase in $\bar{\kappa}$ as defined in (19). With κ unchanged, which we will assume to be the case for now (we let it adjust in the quantitative section), if it becomes relatively cheaper for some types to use non-cash means

of payment, then their consumption will increase.

However, a more useful measure to capture the policy impact is welfare. Define aggregate welfare $\mathbb{W}$ as the sum of utilities within each period composed of the two stages,

$$\mathbb{W} \equiv \sum_{i \in \mathbb{I}} \pi^i \left\{ [U(q^i) - q^i] + \alpha[\epsilon^i u(y^i) - y^i] - \gamma y^i (1 + P^i \tau) \mathcal{I}^{ic} - \kappa \mathcal{I}^{id} \right\}, \tag{20}$$

where $\mathcal{I}^{ic} = 1$ if type-i carries cash and is zero otherwise and $\mathcal{I}^{id} = 1$ if type-i carries digital and is zero otherwise. A higher level of inflation, Π (or an increase in the nominal interest rate, ι) will lead to a fall in output under any payments regime and hence reduce aggregate welfare. But, the effect varies across households given their preference for early consumption. To look at this differential effect define welfare of household-i as its utility within a period or $(1 - \beta)V^i$,

$$\mathbb{W}^i \equiv (1-\beta)V^i = [U(q^i) - q^i(1+\tau)] + \alpha[\epsilon^i u(y^i) - y^i(1+P^i\tau)] - \gamma y^i(1+P^i\tau)\mathcal{I}^{ic} - \kappa \mathcal{I}^{id} + \mathbb{T}^i - \Pi \phi m^i, \tag{21}$$

where $\mathbb{T}^i$ is transfer to household-i. Total lump-sum transfer by the government, $\mathbb{T}$ is given by (2). We will assume that household transfers are proportional to each household's tax and money holdings.[4] So, $\mathbb{W}^i$ will simply become the sum of trade surpluses generated by each household across the two sub-periods net of cost of payments (or, the term inside the summation in (20)). Inflation leads to a fall in welfare for each household. But, the extent of fall will be greater for households with higher money balances or the richer ones, implying that inflation is re-distributive.[5]

In response to the demonetization policy announcement, since firms no longer accept cash in exchange for special goods, there is an unambiguous fall in aggregate welfare $\mathbb{W}_t$ that period, as output y_t^i for households who carried cash falls. There will be a decline in welfare for cash households with no change for non-cash ones. And, given that households with

[4] Alternatively, we can also assume the transfers to be equal across households. In this case, we get that household welfare $\mathbb{W}^{ie}$ is the surplus from trade for each household net of taxes and costs. Aggregate welfare can be derived by summing across household welfares, and it can be easily verified that it is the same under both cases and is equal to (20).

[5] If instead we assume that the lump-sum transfers are equal across households then if inflation leads to a substantial decrease in the probability of tax enforcement for household-i, P^i leading to a sufficiently large fall in tax payments then welfare of that household, $\mathbb{W}^{ie}$ can also increase in response to inflation.

lower money balances typically carry cash, the effect of this shock falls disproportionately on the poorer households or the ones with fewer real balances.

The adverse impact of slow remonetization (captured by an increase in γ) will be felt directly by households whose consumption y^i_{t+1} falls. But, even households who consume more after switching to non-cash payments in $t+1$ see a decline in their utility net of payments cost as the increased cost of cash payments triggered the switch to non-cash means. Thus, aggregate welfare $\mathbb{W}_{t+1}$ will be lower as compared to $\mathbb{W}_{t-1}$. Household welfare, $\mathbb{W}^i_{t+1}$ will also be lower for all i except for households who used non-cash means in $t-1$.[6]

It is worth noting that even though both inflation and slow remonetization make money costlier to hold leading to a fall in aggregate welfare, their impacts on individual households are quite different. While inflation tends to be re-distributive as it affects the richer households (ones with larger money holdings) more adversely, generally speaking demonetization followed by slow remonetization affects them less severely than others, if at all. We will present quantitative results for all these different cases in the next section. For now, 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 $\pi^l + \pi^m + \pi^h = 1$.

Proposition 2 (Policy)**.** *If the cost of carrying cash is γ_0 and the cost of using non-cash means of payment is κ_0 with $\bar{\kappa}^l_0 < \bar{\kappa}^m_0 < \kappa_0 < \bar{\kappa}^h_0$ where $\bar{\kappa}^i_0$ is given by (19), we get that,*
(i) Sudden demonetization of currency as described above leads to a fall in early consumption outputs y^l, y^m with y^h unchanged, and a fall in welfares, $\mathbb{W}^l, \mathbb{W}^m$ with $\mathbb{W}^h$ unchanged.
(ii) If the increase in the carrying cost of cash $\gamma_1 > \gamma_0$ implies that $\bar{\kappa}^l_1 < \kappa_0 < \bar{\kappa}^m_1 < \bar{\kappa}^h_1$, where $\bar{\kappa}^i_1$ is given by (19) by setting $\gamma = \gamma_1$, then y^m increases, y^l falls, y^h is unchanged, W^l and W^m fall, W^h is unchanged.

Finally, since tax enforcement effort (15) is increasing in output there is less overall tax evasion when output is higher. To measure the degree of tax evasion we define the size of the shadow economy $\mathbb{S}$ 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) as,

[6] Owing to fewer transactions, government sales tax revenue also falls in response to demonetization followed by slow remonetization. So, if instead we assume equal lump-sum transfers across households, then $\mathbb{W}^{ie}$ falls for all households.

$$\mathbb{S} \equiv \frac{\alpha \sum_{i \in \mathbb{I}} \pi^i [1 - P(\omega^i)] y^i}{\sum_{i \in \mathbb{I}} \pi^i q^i + \alpha \sum_{i \in \mathbb{I}} \pi^i P(\omega^i) y^i}. \tag{22}$$

A higher level of inflation, Π (or an increase in the nominal interest rate, ι) will reduce output and hence lower tax enforcement under any payments regime. The lower enforcement means that the degree of tax evasion is higher. However, the relative size of the shadow sector can still fall because the output generated for early consumption (on which tax is evaded) is lower. The effect of slow remonetization on the size of the shadow economy also remains unclear, because even if enforcement probability P goes up for some transactions that switch away from cash, it decreases for others due to the lower output generated.

4 Quantitative Results

4.1 Parameter Calibration

We will now quantify the above results using a calibrated version of the model where we mimic the policy shock to compute its welfare cost. To aid comparison we will also present the welfare cost of inflation for our calibrated model. We match key features of the Indian economy at quarterly frequency. We first pin down parameters in the model that are common across household types to match aggregate data. We then use consumption expenditure data to determine household specific parameters.

For the early consumption goods market – the 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 utility function for the late consumption goods market – the centralized market – is assumed to be $U^i(q) = \epsilon^i A log(q)$, which implies that the quantity of late consumption goods consumed in equilibrium is $q^{i*} = \epsilon^i A/(1+\tau)$. As is standard in the literature, the parameters of the utility functions (A, σ) are chosen to match the relationship between $M1$ as a fraction of the nominal GDP and nominal interest rate (on 91-day Treasury Bill). Data is obtained from Reserve Bank of India and Federal Reserve Economic Data for the period Q2:1996 to Q3:2016. Real money demand as a fraction of GDP in the model with a representative

household is given by:

$$\frac{\phi M}{Y} = \frac{y(1+P\tau)}{\alpha y + q}, \tag{23}$$

which is a function of nominal interest rate ι through y as given in (17). Note that we set $\epsilon^i = 1$ for the representative household. Simplify (17) using $u(y) = y^{(1-\sigma)}/(1-\sigma)$ to get,

$$\frac{y^{-\sigma}}{1+P\tau} = \frac{\iota+\gamma}{\alpha} + 1. \tag{24}$$

The tax rate τ is taken to be the typical sales tax rate in India for the calibration period at 12.5%. Note that the sales tax rate was not uniform and it varied across commodities and states ranging from 0% on essential commodities to 20% on liquor.

The probability of paying taxes P is a function of effort ω which we will assume to be the cumulative density function of exponential distribution. It satisfies the assumptions on $P(\omega)$ laid out earlier and is given by $P(\omega) = (1 - \exp^{-\lambda\omega})$ with $\lambda > 0$ sufficiently large, else $P(\omega) = 0$. Using the first order condition for the optimal level of effort given by (15), we get that $P(\tau, y, \lambda) = \max\{0, 1 - 1/y\tau\lambda\}$. Thus, the key parameter value that will help pin this probability is λ. We obtain this probability by using data from the Economic Survey 2017-18 on the percent of firms in the informal sector which is 87%.

We calibrate α along with A and σ to match the sales tax to GDP ratio along with average money demand and its interest elasticity (as discussed above). For the fiscal year 2014-15 the sales tax to GDP was 10.6% as reported from the central government budget estimates data. Sales tax to GDP in the model is $\tau(\alpha P y + q)/(\alpha y + q)$. The remaining parameter is the cost of carrying cash γ which is obtained by using data on counterfeiting from the Reserve Bank of India (RBI)'s Fake Indian Currency Notes data. We elaborate on the calibration strategy for γ in Appendix A where we discuss the model with counterfeiting and denominations. A summary of the calibration strategy with targets and parameter values is given in Table 1.

The preference heterogeneity parameters (ϵ^i, π^i) are chosen to match per capita consumption expenditures from the Consumer Expenditure Survey (CES) 2011-2012.[7] We group households into rural and urban deciles based on their consumption levels since we are inter-

[7]This is the latest government-run large-scale consumption data survey available prior to the demonetization episode.

Table 1: Key Calibration targets and Parameter values

Calibration targets	**Value**
average money demand $\phi M/Y$ (annual)	0.184
nominal interest rate ι (annual)	0.071
CPI inflation rate π (annual)	0.065
elasticity of $\phi M/Y$ with ι (negative)	0.186
sales tax/Y	0.106
fraction of counterfeits (%)	0.002
Parameters	**Value**
early consumption utility elasticity σ	0.28
late consumption utility weight A	4.56
early consumption probability α	0.96
carrying cost of cash γ	3e-5
sales tax rate τ	0.125
probability of paying tax, rate parameter λ	10.4

ested in analyzing the impact of the policy shock on these different population groups. We obtain ϵ^i by matching the ratio of decile i's consumption to the population weighted average consumption with the equivalent ratio of early consumption in the model. This gives the range of ϵ^i as $[0.8, 1.5]$ for urban deciles and for rural it lies in $[0.7, 1.2]$.

To derive the cost of using digital or non-cash payments κ we make a distinction between the two major components of M_1 i.e. currency in circulation with the public and demand deposits. The former is the source of cash payments while the latter encompasses non-cash monetary payments. For instance, households use demand deposits directly for payments through writing checks, bank transfers, debit cards and increasingly through mobile payments.[8] 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 and top rural decile in our model is 0.38 of total money holdings. We pin down κ to be the threshold cost for making digital payments $\bar{\kappa}$ as given in (19) for the top rural decile. As a check, the model implied size of the shadow economy as a percentage of GDP, as defined in (22) is 14.9% which is comparable to the estimate for India in Schneider and Enste (2000) at 22.4%.

Finally note that we could assume the cost of making non-cash or digital transactions to be different for rural and urban areas, which a uniform κ would not capture. However, the

[8]The value of transactions made using debit cards (as reported in RBI's payment system indicator database) as a fraction of demand deposits fluctuated between 20-25%.

low usage of non-cash means of payment in rural regions does not seem to be on account of lack of availability or physical presence of bank branches. For instance, RBI banking statistics shows that even though there are 35% of all banks branches in rural areas, rural bank deposits comprise only about 10% of the total bank deposits indicating presence but rather low usage. Thus, setting a uniform cost is more general as the lack of non-cash transactions for rural regions stems from other factors such as a high fee per transaction due to smaller transactions or a lack of trust. These features are captured by our preference parameter.

4.2 Welfare cost

We quantify the impact of a sudden demonetization of high denomination bills (such as the one that occurred on November 8, 2016 in India) and its subsequent slow and costly remonetization by computing its welfare cost. We ask how much consumers would be willing to give up in terms of total consumption to attain the ease of pre-demonetization payment systems. We will then compare it to the welfare cost of inflation i.e. percent consumption sacrifice to attain 0% inflation. As discussed earlier, both policies make money costlier to hold and have a similar impact at the aggregate level but their disaggregated effects are quite different.

Let aggregate welfare in economy E be given by $\mathbb{W}_E$ as defined in (20). If consumption is reduced by Δ for all households in economy E, then welfare is given by:

$$\mathbb{W}_{E,\Delta} = \sum_{i \in \mathbb{I}} \pi^i \left\{ U(\Delta q^i) - q^i + \alpha[\epsilon^i u(\Delta y^i) - y^i] - \gamma y^i (1 + P^i \tau) \mathcal{I}^{ic} - \kappa \mathcal{I}^{id} \right\} \tag{25}$$

where $\mathcal{I}^{ic} = 1$ if type-i carries cash and zero otherwise; similarly for $\mathcal{I}^{id}$. We measure the welfare cost of moving from economy E (pre-shock) to economy E' (post-shock) by the share of consumption that consumers are willing to give up in order to go from economy E' to E. That is, the cost is $1 - \Delta$ where $\Delta \in [0, 1]$ satisfies $\mathbb{W}_{E,\Delta} = \mathbb{W}_{E',\Delta=1}$.

10% inflation: We first compute the aggregate welfare cost of inflation and find that consumers are willing to sacrifice 0.33% of total consumption in order to go from an economy with 10% inflation to one with 0%. We also measure the welfare cost for different consumer

Table 2: Aggregate welfare cost

Policy	Welfare cost[1]
a) demonetization	1.92%
b) slow remonetization	1.27%
c) 10% inflation	0.33%
d) 27.6% inflation	1.27%

[1] % of total consumption consumers are willing to give up in order to go to the pre-intervention economy (for a,b) or 0% inflation (for c,d)

groups, first by grouping them by regions (rural versus urban) and then by ranking them according to their consumption deciles. As discussed, we would expect the richer households i.e. ones with higher money balances to be more adversely affected by inflation than the poorer. The richer groups would include the higher consumption deciles in both regions. We find that the welfare cost of 10% inflation for the highest urban decile is 1.13% of consumption and for the lowest is 0.15%. The corresponding numbers for the top rural decile is 0.55% and bottom is 0.13%.[9]

Demonetization: Now consider a sudden policy announcement to demonetize the high denomination bills in the early consumption period. First, we need to derive the denomination wise cash holdings for households. We assume a simplified two denomination structure with $x_1 = Rs.100$ and $x_2 = Rs.500$, $k = 5$. We re-scale the real money demand per period for each decile by the ratio of observed quarterly (nominal) consumption expenditure to model's per period (real) consumption. Once we obtain the decile-wise nominal money demand, we divide it into our two denominations by holding as much of the larger denomination as possible in order to optimize carrying costs. The value of high denomination bills held will be given by $m_2^i = k[m^i/k]$, where $[m^i/k]$ is the closest integer value that could be held in high denominations. The remainder is allocated to holding in lower denomination m_1^i for household-i. We find that the fraction of cash held in Rs. 500 denominated bills in the model is 70%.

[9] If instead we assumed that the lump-sum transfers are equal across households then if tax payment falls sufficiently for some households then their welfare could also increase in response to inflation. Since the tax enforcement effort depends on output, its inflation led fall might lead to a sufficient decline in tax payments for some. We find that this is true for households in the top 10-20th rural decile who are better off when inflation is 10% as compared to 0%. Recall that under proportional transfer, household welfare does not depend on tax payments, so there is no such trade-off.

As discussed earlier, a sudden demonetization policy announcement would imply that households find their high denomination bills to be no longer acceptable for transactions. Firms refuse to accept these bills, as otherwise they would end up with worthless pieces of paper. The immediate aggregate welfare cost of the policy shock on account of a drop in consumption is 1.92% of total consumption. But the impact of the shock is felt differently across the different consumer groups, depending on the proportion of cash used for transactions by each group as well as the composition of higher denominations in their cash portfolio. For example, in our calibrated model households in the top urban decile do not hold cash and the bottom most carry only low denomination bills and hence they both remain unaffected. The worst affected group is the top 20-30th urban consumption decile whose welfare cost of the sudden demonetization of high denomination bills is 7.41%.[10]

Slow Remonetization: The overnight demonetization shock was not immediately reversed as it took several months for money supply to go back to its pre-shock level. As shown in Figure 1, M_1 at the end of November 2016 was 24% below its level in October 2016 and 18% lower in February 2017. Furthermore, the composition of M_1 also changed. In October 2016 demand deposits/M_1 stood at 0.4 (shaded in dark blue), while in November 2016 this ratio increased to 0.6. The rise in the share of demand deposits in the measure of M_1 can be attributed to the old demonetized bills that were deposited in banks. However, they were increasingly used to make non-cash payments as the rise in point-of-sale debit card transactions by value as well as volumes indicates (reported in RBI's payment system indicator database). There was also an increase in the use of checks following demonetization (RBI, 2019).

The process of this slow remonetization decelerated recovery 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 non/cash or digital means 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.

[10]If instead we assume that transfers are equal across households then the adverse impact of demonetization is felt by everyone. And, the magnitude of impact for the worst affected households is cushioned. For example, households in the top 20-30th urban decile have a welfare cost of 6.47% when transfers are shared equally.

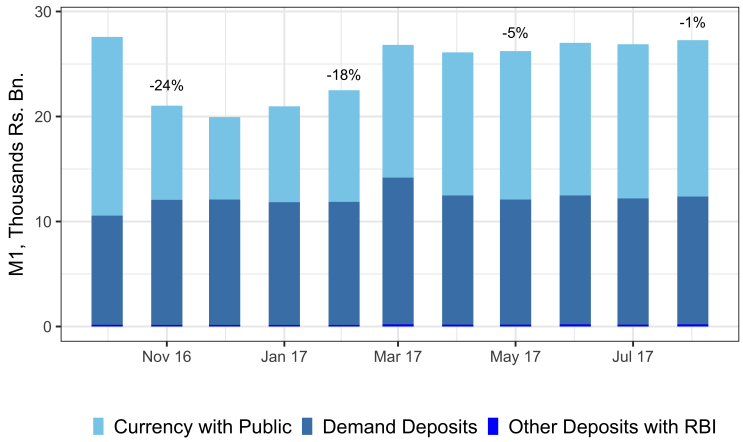
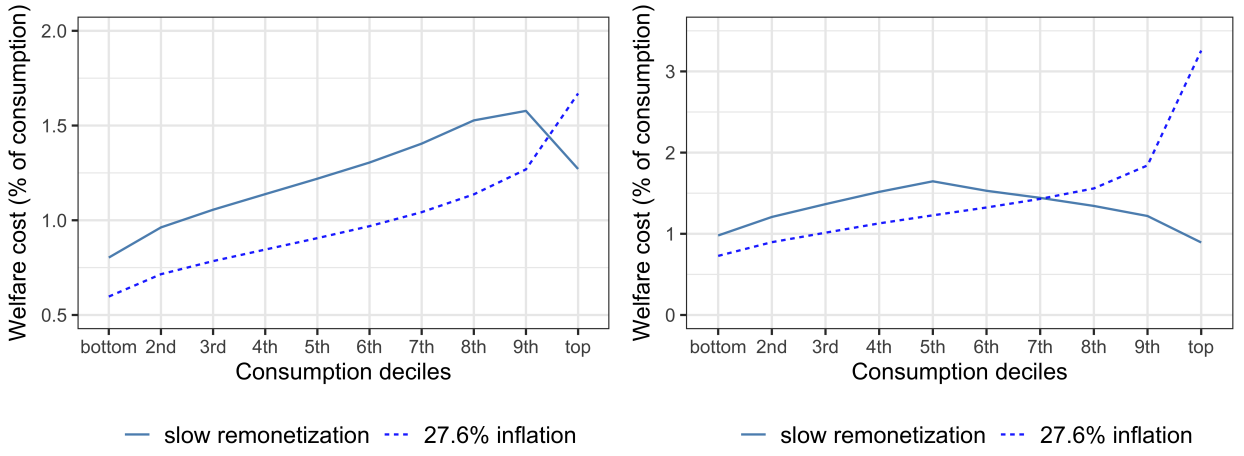


Figure 1: Money supply M_1 and its components
Numbers on the bars show the % declines relative to Oct '16.

We model this slow remonetization by adjusting the cost of using cash, γ to match the remonetization rate from Figure 1 following the shock. We also adjust the cost of using non-cash/digital payments, κ by incorporating the temporary substitution of currency with demand deposits over this period as follows. In the model, households in a particular decile either use digital or cash payments but not both. Since the ratio of demand deposits to M_1 increased to 0.6 in November 2016, we conclude that the top five urban deciles and the top two rural deciles used non cash payments as the ratio of their combined money holdings to total was 0.6. We imply the cost of digital payments, κ to be such that it induces these groups to hold money as deposits i.e. $\kappa = \bar{\kappa}^{r9}$.



(a) Rural (b) Urban

Figure 2: Welfare cost by consumption deciles

Welfare cost is % of total consumption consumers are willing to give up in order to go to the pre-intervention economy (for slow remonetization) or 0% inflation (for 27.6% inflation)

The aggregate welfare cost of slow remonetization owing to the higher payments costs is 1.27% of total consumption. But, the impact is highest for the consumption deciles that have high cash dependence and are unable to switch to non-cash means. The recovery in output is faster for groups that are better equipped to switch to digital payments, that is if their transactions are large enough to justify the cost of making this change. Since the capacity to switch away from cash is higher for the richer households or ones with higher money balances, they are less affected by the slowdown in remonetization. The welfare costs for the different consumption deciles are given in Figure 2. All households experience a fall in welfare in response to the costly and slow remonetization process, as they are willing to give up a positive fraction of their total consumption to attain the ease of pre-demonetization payment systems. But, those in the top consumption deciles for both rural and urban regions are less affected, with ones in the middle the most.

In comparison to the welfare cost of 10% inflation, the cost of remonetization is significantly higher because the effect of increased cost of payment as captured by γ and κ feature directly in the welfare, and not just through their effect on output. Another way to measure the impact of slow and costly remonetization is to find the level of inflation for which the aggregate welfare cost (of going to 0% inflation) is the same. We find that the welfare cost of 27.6% inflation is also 1.27% of total consumption i.e. equivalent to that of slow remone-

tization. But at the disaggregate level, slow remonetization and inflation have the opposite effects. Slow remonetization imposes a larger welfare cost on consumers in the lowest rural decile at 0.8% of consumption versus 0.6% for 27.6% inflation. The corresponding numbers for the top urban decile are 0.9% versus 3.3%.

Inflation makes the richest groups the worst-off, as can be seen by their high welfare costs in Figure 2, while the effect of slow remonetization is felt most by households in the middle consumption deciles. The effect of inflation on households ranked by their consumption size is monotonic because when money gets costly to hold, those carrying higher money balances are the most affected. The effect of slow and costly remonetization is however hump-shaped. This is because when cash becomes costlier than digital means, consumers who were already using the latter or are able to make the switch are not as worse off as those who could not. But, among households who could not switch to non-cash means of payments, the ones with the higher dependence on cash are more adversely affected as they have to pay the higher carrying cost. For them cash gets costlier than before but not costly enough to push them to switch away from cash. This results in a hump-shaped impact of the policy with the worst affected groups being in the middle deciles.[11]

Finally, as discussed previously, the effect on the size of the shadow economy from either monetary intervention is ambiguous. Increasing inflation reduces output and lowers tax enforcement effort. Thus, the degree of tax evasion on this output is higher, pushing the size of the shadow economy up. But, if the share of early consumption output (on which tax is evaded) falls significantly, then the size of the shadow economy can fall even though tax evasion on this output is higher. We find that when the rate of inflation is 10%, the size of the shadow economy falls to 13% as compared to 15% under 0% inflation. Similarly, for slow and costly remonetization captured by an increase in payments costs, tax enforcement effort increases for some transactions that switch to digital payments but falls for others. So, the overall effect on enforcement and the shadow economy remain ambiguous. For our calibrated model, we find that the size of the shadow economy falls to 12% from 15% in response to slow remonetization.

[11] On the other hand, if transfers are equal then the adverse impact of slow remonetization is spread across households on account of the lower overall tax revenue generated. This cushions the large welfare impact for the middle deciles but at the expense of the lower deciles.

5 Conclusion

This paper examines the relationship between the choice of payment methods based on preferences and macroeconomic outcomes. We model this choice using a tractable monetary framework by building on Lagos and Wright (2005) and Rocheteau and Wright (2005). 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. However, the alternative of switching to digital payments involves a fixed usage cost which restricts it to consumers above a certain threshold level of consumption. Second, since cash led transactions are harder to track, they facilitate tax evasion, which increases the size of a parallel shadow economy. 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, our monetary framework enables us to draw conclusions of its impact on aggregate output, welfare and the size of the shadow economy. Our calibration captures key features of India's payment system to conclude that the welfare cost of the policy shock was equivalent to 1.3% of consumption owing to slow and costly remonetization. We disaggregate the effects of this shock based on regions and consumption deciles to find that the divergence in means of payments usage is reflected in the high welfare cost numbers for some groups. The impact of the policy shock is highest for the consumption deciles that have high cash dependence and are unable to switch to non-cash means i.e. those in the middle consumption deciles. The recovery in output is faster for groups that are better equipped to switch to digital payments, that is if their transactions are large enough to justify the cost of making this change. Since the capacity to switch away from cash is higher for the richer households or ones with higher money balances, they are less affected by the slowdown in remonetization.

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.

References

Agarwal, S., Ghosh, P., Li, J., and Ruan, T. (2019). Digital Payments Induce Over-Spending : Evidence from the 2016 Demonetization in India. *Working paper*.

Agrawal, S. (2018). Black Economy and Demonetisation. *Working Paper*.

Ait Lahcen, M. and Gomis-Porqueras, P. (2021). A Model of Endogenous Financial Inclusion: Implications for Inequality and Monetary Policy. *Journal of Money, Credit and Banking*, 53(5):1175–1209.

Camera, G. (2001). Dirty money. *Journal of Monetary Economics*, (47):377–415.

Chanda, A. and Cook, C. J. (2019). Who Gained from India's Demonetization? Insights from Satellites and Surveys. *SSRN Electronic Journal*.

Chodorow-Reich, G., Gopinath, G., Mishra, P., and Narayan, A. (2020). Cash and the Economy: Evidence from India's Demonetization. *Quarterly Journal of Economics*, pages 57–103.

Crouzet, N., Gupta, A., and Mezzanotti, F. (2019). Shocks and Technology Adoption : Evidence from Electronic Payment Systems. *Working Paper*.

Di Nola, A., Kocharkov, G., Scholl, A., and Anna-Mariia Tkhir (2018). The Aggregate Consequences of Tax Evasion. *Working Paper*.

Gomis-Porqueras, P., Peralta-Alva, A., and Waller, C. J. (2014). The shadow economy as an equilibrium outcome. *Journal of Economic Dynamics and Control*, 41:1–19.

Karmakar, S. and Narayanan, A. (2019). Do Households Care About Cash? Exploring the Heterogeneous Effects of India's Demonetization. *SSRN Electronic Journal*.

Kim, Y. S. and Lee, M. (2010). A model of debit card as a means of payment. *Journal of Economic Dynamics and Control*, 34(8):1359–1368.

Koreshkova, T. A. (2006). A quantitative analysis of inflation as a tax on the underground economy. *Journal of Monetary Economics*, 53(4):773–796.

Lagos, R. and Wright, R. (2005). A Unified Framework for Monetary Theory and Policy Analysis. *Journal of Political Economy*, 113(3):463–484.

Lahiri, A. (2020). The Great Indian Demonetization. *Journal of Economic Perspectives*, 34(1):55–74.

Lee, M., Wallace, N., and Zhu, T. (2005). Modeling Denomination Structures. *Econometrica*, 73(3):949–960.

Li, Y. (2011). Currency and checking deposits as means of payment. *Review of Economic Dynamics*, 14(2):403–417.

Lotz, S. and Vasselin, F. (2019). A New Monetarist Model of Fiat and E-Money. *Economic Inquiry*, 57(1):498–514.

Nosal, E. and Rocheteau, G. (2011). *Money, Payments and Liquidity*. MIT Press.

RBI (2019). Benchmarking India's Payment Systems. *Reserve Bank of India Report*, (June).

Rocheteau, G. and Wright, R. (2005). Money in search equilibrium, in competitive equilibrium, and in competitive search equilibrium. *Econometrica*, 73(1):175–202.

Rogoff, K. S. (2017). *The curse of cash: How large-denomination bills aid crime and tax evasion and constrain monetary policy*. Princeton University Press.

Runnemark, E., Hedman, J., and Xiao, X. (2015). Do consumers pay more using debit cards than cash? *Electronic Commerce Research and Applications*, 14(5):285–291.

Schneider, F. and Enste, D. H. (2000). Shadow Economies: Size, Causes, and Consequences. *Journal of Economic Literature*, 38(1):77–114.

Tagat, A., Ozmen, M., and Trivedi, P. L. (2019). Consumer Payments Survey of India: A Closer Look at Household Finances and Payment.

Tagat, A. and Trivedi, P. L. (2020). Demand for cash: An econometric model of currency demand in India. *Macroeconomics and Finance in Emerging Market Economies*, 00(00):1–18.

Wadhwa, S. (2019). Impact of Demonetization on Household Consumption in India. *Working Paper*.

Waknis, P. (2017). Demonetisation through Segmented Markets: Some Theoretical Perspectives. *Economic and Political Weekly*, 52(9).

Williamson, S. D. (2019). Central Bank Digital Currency: Welfare and Policy Implications. *Working Paper*, pages 1–20.

Zhu, Y. and Hendry, S. (2019). A Framework for Analyzing Monetary Policy in an Economy with E-Money. *SSRN Electronic Journal*.

A Appendix: Denominations

In this section we first show why we can replace the portfolio choice problem when households carry cash in both high and low denominations with one in only low denominations as in (6). Next, we give details on how the carrying cost of cash γ is calibrated using data on the percentage of counterfeit bills as a fraction of total currency in circulation.

The full portfolio choice problem when households carry cash will also include the choice of denominations which in turn depends on the proportion of counterfeit bills in circulation, η. This proportion or the degree of counterfeiting in the economy, depends on the decision of potential counterfeiters. The nominal cost of counterfeiting $\delta > 1$ per bill implies that low denomination bills will not be counterfeited. If $\delta < k$ which we will assume, then any number of the high-denomination bills can be counterfeited. Recall that the low denomination is fully divisible, and the high is available in $k > 1$ units of the low denomination. So, in real terms, we get $\phi_t < \delta\phi_t < k\phi_t$, i.e. the counterfeiter would like to counterfeit as many high-bills as possible in any period. But, there will be an upper limit on counterfeiting of these bills because if η is too high, households will not demand any high-bills. In fact, η will be such that households are indifferent between holding their portfolio m_t fully in low denomination bills or in a mixed form. If households carry m_t money balances in mixed form then they hold the maximum possible in high denominations i.e. $[m_t/k]$ bills are held in the high-denomination and $m_t - k[m_t/k]$ in low.

If a representative household carries cash in either denomination then its value from early consumption becomes,

$$V_t(m_t) = -\phi_t\gamma m_t + \phi_t\gamma(k-1)\left[\frac{m_t}{k}\right] + \tag{26}$$

$$+\alpha \max_{y_t}\left\{\epsilon u(y_t) + W_t\left(\phi_t m_t - k\eta\phi_t\left[\frac{m_t}{k}\right] - p_t y_t\right)\right\} + (1-\alpha)W_t\left(\phi_t m_t - k\eta\phi_t\left[\frac{m_t}{k}\right]\right),$$

s.t.

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

Note that now the carrying cost of each bill held depends on the denominations held. 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: $-\gamma[m_t/k] - \gamma k\left(m_t/k - [m_t/k]\right) = -\gamma m_t + \gamma(k-1)[m_t/k]$.

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 α. In this case, they consume y_t and if they do not receive this shock (with probability $1-\alpha$), they carry 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 their portfolio as a high-denomination bill is a counterfeit with probability η which is detected in the next stage. Thus, the household gets $1-\eta$ times the value held in high denomination bills, i.e. consumption of special goods is constrained by: $k(1-\eta)\phi_t[m_t/k] + k\phi_t\left(m_t/k - [m_t/k]\right) = \phi_t m_t - k\eta\phi_t[m_t/k]$.

We simplify the above value $V_t(\cdot)$ as follows:

$$V_t(m_t) = -\phi_t\gamma m_t + \phi_t\gamma(k-1)\left[\frac{m_t}{k}\right] + \alpha\max_{y_t}[\epsilon u(y_t) - p_t y_t] + \phi_t m_t - k\eta\phi_t\left[\frac{m_t}{k}\right] + W(0), \quad (27)$$

s.t.

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

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 (3). Ignore constants to get the following portfolio choice maximization problem for a household each period:

$$\max_{m_t}\left\{-\phi_{t-1}m_t + \beta\left\{-\phi_t\gamma m_t + \phi_t\gamma(k-1)\left[\frac{m_t}{k}\right] + \alpha\max_{y_t\leq\mathbb{C}_t}[\epsilon u(y_t) - p_t y_t] + \phi_t m_t - k\eta\phi_t\left[\frac{m_t}{k}\right]\right\}\right\}, \quad (28)$$

where $\mathbb{C}_t \equiv (\phi_t/p_t)m_t - k\eta(\phi_t/p_t)\left[m_t/k\right]$ is the constraint on early consumption. Each household takes the prices p_t, ϕ_t and the degree of counterfeiting η as given and maximizes the above.

But as discussed, we can replace the portfolio choice problem in (28) with high and low denominations given above with the problem with only low denominations as given in (6). In addition, due to counterfeiting, the supply of money, M in the market clearing condition (16) will need to be augmented to reflect the counterfeits on the high-denomination bills as well. The augmented supply of money M_t^s is given by:

$$M_t^s = M_t + \frac{\eta}{(1-\eta)}k\left[\frac{M_t}{k}\right], \quad (29)$$

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

Now we show how we calibrate the cost of carrying cash γ using data on the percentage of counterfeiting η 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 denominations. For the model we collapse the denominations into two: high being the Rs. 500 denoted as x_2 and low as the Rs. 100 bill represented by x_1. Hence, the value of the high denomination bill in terms of the low given by $k \equiv x_2/x_1 = 5$.

Equating demand and supply for each denomination we get, $\phi m_2 = M_2/P(1-\eta)$ and $\phi m_1 = M_1/P$, where M_1, M_2 stand for supplies of the two denominations. Note that we augment the supply of M_2 given the proportion of counterfeits η. Define surplus from early consumption as $s(y) \equiv \epsilon u(y) - py$. And as discussed, since households are indifferent between holding their portfolio m_t fully in low denomination bills or in a mixed form we equate (28) and (6) to re-write it as,

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

This gives us a relation between γ and η. Since the two surpluses are approximately equal, we set $\gamma = \eta k/(k-1)$.

B Appendix: Proofs

B.1 Proof of Lemma 1

Since $V_t^i = \max\{V_t^{ic}, V_t^{id}\}$ as given by (12), we need to compare (6) and (10). We'll get $V^i = \max\{V^{ic}, V^{id}\} = V^{ic}$ if and only if,

$$-\phi_{t-1}m_t^{ic} + \beta\{-\gamma\phi_t m^{ic} + \alpha[\epsilon^i u(y_t^{ic}) - y_t^{ic}(1+P_t^i\tau)] + \phi_t m^{ic}\} >$$

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

where $y_t^{ic} = \phi m_t^{ic}/(1+P_t^i\tau)$ is given by (17) and $y_t^{id} = \phi m_t^{id}/(1+P_t^i\tau)$ is given by (18).

Divide both sides by β and use $\phi_{t-1}/\phi_t = (1+\iota)/(1+\rho)$ (also ignore time subscript, t),

$$-\iota y^{ic}(1+P^i\tau) - \gamma y^{ic}(1+P^i\tau) + \alpha[\epsilon^i u(y^{ic}) - y^{ic}(1+P^i\tau)] > -\iota\phi y^{id}(1+P^i\tau) - \kappa + \alpha[\epsilon^i u(y^{id}) - y^{id}(1+P^i\tau)].$$

Thus, $V^i = \max\{V^{ic}, V^{id}\} = V^{ic}$ if and only if,

$$\kappa > \bar{\kappa}^i \equiv \gamma y^{ic}(1+P^i\tau) + \alpha\epsilon^i[u(y^{id}) - u(y^{ic})] - (\iota+\alpha)[y^{id}(1+P^i\tau) - y^{ic}(1+P^i\tau)].$$

B.2 Proof of Proposition 1

We will prove existence and uniqueness for (iii) partial cash payments, others will follow similarly. If $\bar{\kappa}^l < \kappa$, $\kappa < \bar{\kappa}^m$ and $\kappa < \bar{\kappa}^h$, then type l uses cash, and m and h non-cash/digital from Lemma 1. So, we need to find $\{y^{lc}, y^{md}, y^{hd}, \omega^{lc}, \omega^{md}, \omega^{hd}, \phi\}$ by solving for:

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

$$P'(\omega^{lc}) = \frac{1}{\tau y^{lc}}.$$

$$\alpha\left[\frac{\epsilon^m}{(1+P(\omega^{md})\tau)}u'(y^{md}) - 1\right] = \iota,$$

$$\omega^{md} = P'^{-1}\left(\frac{1}{\tau y^{md}}\right).$$

$$\alpha\left[\frac{\epsilon^h}{(1+P(\omega^{hd})\tau)}u'(y^{hd}) - 1\right] = \iota,$$

$$\omega^{hd} = P'^{-1}\left(\frac{1}{\tau y^{hd}}\right),$$

$$\pi^l y^{lc}(1+P(\omega^{lc})\tau) + \pi^m y^{md}(1+P(\omega^{md})\tau) + \pi^h y^{hd}(1+P(\omega^{hd})\tau) = \phi M.$$

Consider the first equation for y^{lc}, ignore constants and take the derivative of the left hand side with respect to y^{lc} to get:

$$\frac{u''\left(y^{lc}\right)}{[1+P(\omega^{lc})\tau]} - \frac{u'\left(y^{lc}\right)}{[1+P(\omega^{lc})\tau]^2}\tau P'(\omega^{lc})\frac{\partial \omega^{lc}}{\partial y^{lc}} < 0.$$

From the second equation for ω^{lc}, we get $\frac{\partial \omega^{lc}}{\partial y^{lc}} > 0$ since $P''(\cdot) < 0$. Thus, the left hand side of the first equation is decreasing in y^{lc} since $u''(\cdot) < 0$, hence we get a unique solution for y^{lc} and ω^{lc}. Similarly, we can get a unique $\{y^{md}, y^{hd}, \omega^{md}, \omega^{hd}\}$, and finally we get ϕ from the last equation. ■

B.3 Proof of Lemma 2

(a) (i) Fully differentiate (17) with respect to γ:

$$\left\{\frac{u''(y)}{[1+P(\omega)\tau]} - \frac{u'(y)}{[1+P(\omega)\tau]^2}\tau P'(\omega)\frac{\partial \omega}{\partial y}\right\}\frac{\partial y}{\partial \gamma} = \frac{1}{\epsilon\alpha}.$$

Thus $\frac{\partial y}{\partial \gamma} < 0$ since $u'(\cdot) > 0$, $u''(\cdot) < 0$ and $P'\omega) > 0$.

(b)(i) And, $\frac{\partial \omega}{\partial \gamma} = \frac{\partial \omega}{\partial y}\frac{\partial y}{\partial \gamma} < 0$.

(a) (ii) Take the derivative of (15) with respect to y:

$$P''(\omega)\frac{\partial \omega}{\partial y} = -\frac{1}{\tau y^2}.$$

We get that $\frac{\partial \omega}{\partial y} > 0$, since $P''(\omega) < 0$. We've suppressed the i superscript. Next, fully differentiate (17) with respect to ϵ:

$$\frac{u'(y)}{[1+P(\omega)\tau]} + \left\{\epsilon\frac{u''(y)}{[1+P(\omega)\tau]} - \epsilon\frac{u'(y)}{[1+P(\omega)\tau]^2}\tau P'(\omega)\frac{\partial \omega}{\partial y}\right\}\frac{\partial y}{\partial \epsilon} = 0.$$

Thus $\frac{\partial y}{\partial \epsilon} > 0$ since $u'(\cdot) > 0$, $u''(\cdot) < 0$ and $P'\omega) > 0$.

(b)(ii) And, $\frac{\partial \omega}{\partial \epsilon} = \frac{\partial \omega}{\partial y}\frac{\partial y}{\partial \epsilon} > 0$.

■

B.4 Proof of Proposition 2

(i) Given that $\bar{\kappa}_0^l < \bar{\kappa}_0^m < \kappa_0 < \bar{\kappa}_0^h$, types l and m hold cash and h non-cash in period 0. Thus, y^l, y^m will fall and y^h will remain unchanged in response to demonetization. Thus, W^l and W^m fall while W_h is unaffected.

(ii) If $\bar{\kappa}_1^l < \kappa_0 < \bar{\kappa}_1^m < \bar{\kappa}_1^h$ then only l holds cash in period 1, m moves to non-cash (as discussed below) and h continues to carry non-cash. So, from Lemma 2 *(a)(i)* y^m increases, and y^l falls as $\frac{\partial y^l}{\partial \gamma} < 0$.

Consider welfare for type-m under proportional transfer as described in the text:

$$\mathbb{W}_0^m = [U(q^m) - q^m] + \alpha[\epsilon^m u(y^{mc}) - y^{mc}] - \gamma_0 y^{mc}(1 + P^{mc}\tau),$$

$$\mathbb{W}_1^m = [U(q^m) - q^m] + \alpha[\epsilon^m u(y^{md}) - y^{md}] - \kappa_0.$$

At $\gamma = \gamma_0$ since type-m uses cash it is true that $W_1^m < W_0^m$. If γ increases or at $\gamma = \gamma_1 > \gamma_0$,

$$[U(q^m) - q^m] + \alpha[\epsilon^m u(y_1^{mc}) - y_1^{mc}] - \gamma_1 y_1^{mc}(1 + P_1^{mc}\tau) < \mathbb{W}_0^m,$$

and from Lemma 1 at γ_1, if it is the case that $\kappa_0 < \bar{\kappa}_1^m$, then type-$m$ uses digital means and $W_1^m < W_0^m$.

The fall in W^l and no change in W^h can be verified easily. ■