Mobile Money Agent Competition, Inventory Management, and Pooling

A dissertation presented
by
Karthik Balasubramanian
to
the Technology and Operations Management Unit
in partial fulfillment of the requirements for the
degree of Doctor of Business Administration
in the subject of
Technology and Operations Management

Harvard Business School
Boston, Massachusetts

May 2018

Advisor: Professor David Drake Karthik Balasubramanian

Mobile Money Agent Competition, Inventory Management, and Pooling

Abstract

The use of electronic money transfer through cellular networks ("mobile money") is rapidly increasing in the developing world. The resulting electronic currency ecosystem could improve the lives of the estimated two billion people who live outside the formal financial infrastructure by facilitating more efficient, accessible, and reliable ways to store and transfer money than are currently available. Sustaining this ecosystem requires a healthy network of agents to conduct cash-for-electronic value transactions and vice versa. This thesis consists of three chapters formatted as stand-alone papers that explore how these agents compete, how they should manage their inventories, and how they might benefit from inventory pooling. Chapter I, *Service Quality and Competition: Empirical Analysis of Mobile Money Agents in Africa*, explores how service quality, competition, and poverty are related to demand and inventory. Chapter II, *Inventory Management for Mobile Money Agents in the Developing World*, describes various approaches to modeling the agent's inventory problem, including a simple analytical heuristic that performs well relative to actual agent decisions. Finally, Chapter III, *Inventory Pooling for Mobile Money Agents in the Developing World*, proposes an inventory pooling framework that has the potential to increase channel profitability by simultaneously reducing inventory requirements and increasing service levels system-wide.

Acknowledgements

The following people have been foundational to my development, and by extension, this thesis: Sen Balasubramanian, Saroja Balasubramanian, Malar Balasubramanian, Sumathi Balasubramanian, Ashton Rohmer, Kanimozhi Venkatesan, and Venkatesan Gounder. Guidance and support from my committee – Ananth Raman, David Drake, Jason Acimovic, and Doug Fearing – has been invaluable both to me personally and to this thesis. Additionally, the following people have played an important role for which I am deeply grateful: Nicole DeHoratius, Joel Goh, Kris Ferreria, Ryan Buell, and Mike Toffel. Finally, I also wish to thank the Harvard Business School Doctoral Programs Office for their generous financial support of this research. I am particularly grateful to DPO's Jennifer Mucciarone and Marais Young for their personal support and encouragement throughout the program.

Author list

The following additional author contributed to Chapter I: David Drake
The following additional authors contributed to Chapter II: David Drake and Doug Fearing
The following additional authors contributed to Chapter III: Jason Acimovic and David Drake

Contents

Introduction

Over two billion adults globally do not have access to a bank account (Demirguc-kunt 2012). This is a chilling statistic because not having a bank account severely hampers both quality of life and social mobility. Not having a bank account is harmful in several ways: moving money cannot be done inexpensively and easily (as it must travel in physical cash form); savings instruments are limited to physical assets such as gold, livestock and cash under the mattress (all of which are prone to theft and loss); credit is only available in physical form from one's immediate vicinity (often resulting in short repayment terms and extremely high interest rates); and insurance is also limited to what the local community is able to provide, which is useless in the face of a community-wide calamity (Morawczynski 2009). These challenges are particularly salient because those who are "unbanked" are also overwhelmingly likely to be poor (Helms 2006). The primary reason poor people are "unbanked" is that serving the poor, by definition people with little wealth, is not commercially viable for financial institutions. Given that the per-customer revenue potential is limited, even a highly efficient bank branch that transacts in person and in physical cash cannot serve poor customers profitably (Mas 2010).

The promise of delivering high-quality financial services on top of cellular networks – which have become nearly ubiquitous in the developing world – is stunning. When customers no longer need to interact with financial institutions in cash and in person, and instead use only electronic menus and electronic currency to transact, the cost structures of firms offering financial services are reduced so much that it is foreseeable that even the poorest customer

can be a profitable one (Kendall 2011). And given how important robust financial instruments are, one could imagine a world where the poor benefit substantially from a robust digital financial ecosystem of high-quality payment, savings, credit, and insurance products – all accessed through mobile phones.

This vision of digital financial inclusion, however, has been only partially realized – and that too only in a select few countries. Systems that allow people to send and receive money with their mobile phones – called "mobile money" platforms – have generated significant social value by allowing people to securely transfer and receive money. I first became acquainted with mobile money systems while living and working in Moshi, Tanzania in 2009 and 2010. I was astounded by both its impact and potential to generate real social value. In 2011, I was fortunate to join the Bill & Melinda Gates Foundation's Financial Services for the Poor team, where I was able to focus on the broadening and deepening of mobile money systems in East Africa and South Asia. In my role at the Gates Foundation, I met a large number of managers in the nascent space of mobile-enabled financial services. From these conversations, there emerged a consistent set of factors that seemed to be hampering the further progression of mobile money platforms towards a vision of robust digital financial inclusion. The factors hampering access, adoption, and usage of mobile money are wide and varied. First, an enabling regulatory environment that permits money to flow over mobile networks without unnecessary regulation is often missing (Evans and Pirchio 2014). Next, customer registration requirements frequently create sufficient friction to discourage adoption of digital financial services (Intermedia 2013). Additionally, many mobile money platforms are mired in "sub-scale traps" because their initial marketing roll-outs were under-funded (Mas and Radcliffe 2011). Pricing policies are also often not simultaneously conducive to the operator's business case and customer adoption and usage (Economides and Jeziorski 2017). Finally, the "bridges" that support the transition period between the current economy – in which cash is indispensable – and a future digital, "cash-lite" economy are often unreliable (Maurer et al. 2013): mobile money agents, the small shop owners that convert cash to elec-

tronic value and vice versa for a commission are often unable to complete these transactions (Eijkman et al. 2010).

While all of these factors are salient, worthy of academic research, and on the critical path to a fully inclusive digital financial ecosystem, the management of mobile money agencies is the piece of the puzzle that is most amenable to analysis with an operations management lens. This thesis, therefore, utilizes the operations management toolbox to analyze mobile money agencies and make recommendations about how they might be improved. This thesis consists of three chapters formatted as stand-alone papers that explore how these agents compete, how they should manage their inventories, and how they can benefit from inventory pooling. Chapter I empirically explores how service quality, competition, and poverty are related to demand and inventory. Chapter II describes various approaches to modeling the agent's fundamental inventory problem, including a simple analytical heuristic that performs well relative to actual agent decisions. Finally, Chapter III proposes the model for an inventory pooling framework that has the potential to increase channel profitability by simultaneously reducing inventory requirements and increasing service levels system-wide.

Chapter 1

Service Quality and Competition: Empirical Analysis of Mobile Money Agents in Africa

1 Abstract

The use of electronic money transfer through cellular networks ("mobile money") is rapidly increasing in the developing world. The resulting electronic currency ecosystem could improve the lives of the estimated two billion people who do not have access to a bank account by facilitating more secure, accessible, and reliable ways to store and transfer money than are currently available. The development of this ecosystem requires a network of agents to conduct cash-for-electronic value transactions and vice versa. This paper examines how service quality (consisting of pricing transparency and agent expertise in this setting), competition, and poverty are related to demand and inventory of electronic value and physical cash. We find that transaction demand increases with both pricing transparency and agent expertise, and that agent expertise interacts positively with competitive intensity. We also find that competition is associated with higher inventory holdings of both cash and elec-

tronic value. Furthermore, agents in high-poverty areas hold greater amounts of cash but do not carry a smaller amount of electronic value – suggesting that they incur greater working capital requirements than agents in lower-poverty areas. These results offer insight to mobile money operators with respect to monitoring, training, and the business case for their agents. This paper furthers our understanding of service quality, competition, and inventory, while developing a foundation for the exploration of mobile money by Operations Management scholars.

2 Introduction

In the past decade, "mobile money" platforms have experienced explosive growth in the developing world, with over 255 active mobile money systems in 89 countries (Groupe Speciale Mobile Association 2015). These platforms, primarily built and managed by mobile network operators, allow money to be stored in the form of digital currency (hereafter referred to as e-float). In much the same way that text messages can be sent quickly and cheaply, e-float can be securely and instantly transferred across long distances at a near-zero marginal transaction cost. Mobile money platforms are of particular interest to the base-of-the-pyramid (BoP) community—scholars and practitioners developing business models deliberately geared toward serving the population in poverty—because they have potential to connect millions of poor and "unbanked" people to the formal financial system. Mobile money has potential to provide several benefits: i) it can enable quicker recovery from economic shocks such as job loss or illness to the primary wage-earner (Jack and Suri 2014); ii) it can enable more efficient receipt of monetary transfers from non-governmental organizations (NGOs) after disasters (Aker et al. 2011); and iii) it can lay the foundation for access to formal savings, credit, and insurance opportunities for those who currently lack such access (Mas 2010).

A healthy network of cash-in/cash-out (CICO) agents to serve as the bridge between

physical cash and e-float is critical to the success of mobile money platforms. These agents, often small shop-owners, convert cash to e-float ("cash-in" transactions) and e-float back to cash ("cash-out" transactions) for a commission. Because these agents ensure the convenient convertibility of e-float, the quality with which they perform such transactions is crucial to consumer confidence in the platforms. Consequently, operations management scholars have a role to play in the development of robust mobile money ecosystems.

The contract and inventory theory tools developed over the past decades can apply to this fundamentally new context, but as a community we must first develop an understanding of the mobile money business, particularly the nature of demand and factors that influence agents' inventory decisions. Accordingly, we seek to provide an introduction to mobile money, explore the relationships between service quality, competition, and demand, as well as study the relationships between competition, poverty, and inventory. We explore four research questions. First, what is the relationship between an agent's service quality and their demand? We are interested in two dimensions of service quality pertinent in this context: i) pricing transparency (related to agent credibility); and ii) agent expertise (related to agent competence). Second, how is the relationship between service quality and demand moderated by competitive intensity? Third, what is the relationship between competitive intensity and inventory holdings of cash and e-float? And lastly, what is the relationship between the level of poverty in an agent's catchment area and agent inventory decisions?

To address these questions, we use a combination of agent network and demographic data sources from Kenya and Uganda, two East African countries at different stages of mobile money market development. We use an in-person survey of over 3,000 mobile money agents that operate in the two countries. We then combine the locations of the surveyed agents with the precise locations of over 68,000 bank branches, bus stands, and mobile money agents, as well as population and poverty estimates for each square kilometer of the countries. While this paper makes no causal claims, we elucidate associations that are of both academic and practitioner interest. Specifically, we find that agents who are more transparent with

transaction pricing and agents who are more knowledgeable experience significantly (both statistically and economically) greater demand. Agents' pricing transparency does not interact with competitive intensity in a statistically significant way. Additionally, agents that provide more knowledgeable service, on the other hand, seem to reap greater rewards from their performance in the face of greater competition. This can be related to Hill's (1993) operations strategy framework of "order-qualifiers" and "order-winners": pricing transparency may act as an "order-qualifier" among a portion of the consumer-base, while expertise may act as an "order-winner"— a dimension along which agents compete. We also find that agents who face more competitive intensity stock more inventory of both cash and e-float. Lastly, we find that agents in high-poverty areas stock more cash, but do not stock a correspondingly smaller amount of e-float. This indicates that agents in high-poverty areas incur higher working capital costs than agents in lower-poverty areas.

3 Literature review

Our work relates to literature focused on service quality, competition, inventory management, and operations at the BoP.

Service quality has attracted academic interest in the past three decades. While quality has many definitions and dimensions (Reeves and Bednar 1994), we focus on two widely-recognized dimensions of service quality: credibility and competence (Parasuraman and Zeithaml 1988). We measure credibility and competence, respectively, through each agent's pricing transparency and their expertise with respect to transaction policies and procedures. While (to the best of our knowledge) there are no empirical studies that directly explore the effect of pricing transparency and a provider's expertise on demand, the framework of trust developed in the literature is relevant. A common conceptualization of trust is two-dimensional: trust is composed of "benevolence trust" and "competence trust" (Singh and Sirdeshmukh 2000). Benevolence trust (the faith customers have in firms not to cheat them)

is related to our measure of credibility (i.e., pricing transparency) and competence trust (the faith customers have in firms to be able to competently fulfill demand) is related to agent expertise. Both of these concepts and their applications are discussed further in Section 5 where we develop our hypotheses.

In addition to literature exploring service quality, there has also been interest in the effect of competition on inventory. Olivares and Cachon (2009) argue that the theory on this relationship is mixed. On one hand, greater competitive intensity will drive down price, predicting lower optimal inventory holdings. On the other hand, greater competitive intensity will force firms to compete on service level, driving inventory holding up. In the context of the US auto market, Olivares and Cachon (2009) find that greater competitive intensity among dealers results in greater inventory holdings. Our work explores this relationship, but in the context of an emergent financial service in the developing world.

Finally, because mobile money has potential to dramatically lower the cost structure of providing financial services to those living in poverty, mobile money is fundamentally related to the growing literature on serving the BoP (Prahalad and Hammond 2002). Research in the operations management community focused on the BoP is quite nascent (e.g., Sodhi and Tang 2011; Gold et al. 2013). Our work here contributes to this emerging stream by examining the relationship between poverty in catchment areas and agent inventory decisions.

Our work differs from existing literature in several notable ways. First, our context is very different from most studies examining service quality, competition, and/or inventory holding. Rather than analyzing competition and inventory in traditional sectors in the developed world, such as the US auto market, we focus on an innovative financial service in East Africa. Second, we are able to analyze how competition moderates the relationship between service quality and demand, rather than only looking at service quality or competition individually. Third, our unique combination of rich survey and spatial data allows us to examine the relationship between poverty and inventory decisions which, to the best of our knowledge, has not been studied in any context.

4 Context

In this section, we provide an overview of mobile money's history, impact, and implications. We also describe the mechanics of a mobile money transaction, as well as challenges and opportunities confronting mobile money systems.

4.1 Mobile money motivation and history

The poor comprise the vast majority of the "unbanked", the two billion people globally who do not have an account at a formal financial institution (Demirguc-kunt 2012). The poor and unbanked—roughly one-third of the world's population—rely mostly on physical cash when transferring money. Thus, the velocity of money is limited by how quickly cash can be physically transported, by foot or by bus in most circumstances (Batista and Vicente 2013). This limitation is a critical disadvantage to the poor when money is needed most, such as in the aftermath of a negative economic shock (e.g., sickness or job loss) or when a rare opportunity to climb out of poverty through investment emerges (e.g., fertilizer or improved seed purchases) (Helms 2006). At these decisive moments, friends and family willing and able to transfer money must rely on expensive and/or unreliable methods such as bus money transfer services (Morawczynski 2009). Furthermore, saving for these pivotal moments is more challenging with inferior savings tools; to store and save money, most either hide cash in their homes (at risk of theft and ineligible for interest), or purchase relatively illiquid assets like gold or livestock (that are often sold at a loss in times of need) (Collins et al. 2009). One study found that among a large sample of the poor in Uganda, 75% had lost some portion of their cash or physical asset savings in the previous year (Wright and Mutesasira 2001). Similarly, informal credit for investment opportunities and insurance options for risk mitigation are substandard among the unbanked; credit is often only available from moneylenders at usurious rates, and formal insurance is generally inaccessible, if it exists at all (Collins et al. 2009). The poor, especially those living in rural areas, remain unserved by formal

financial institutions because their low balances and transaction sizes yield little revenue for banks (Mas 2010). Furthermore, because the rural poor live in definitionally low density and remote areas, these regions lack the scale required to make the provision of traditional financial services an attractive proposition. Consequently, financial institutions have largely found it impractical to profitably serve the poor in the developing world, particularly those residing in rural areas (Kendall 2011).

The rapid growth of cellular networks in the developing world in the past decade lays the groundwork for a potential paradigm shift in financial services for the poor. According to the Boston Consulting Group (BCG), the proportion of unbanked people in the developing world with access to mobile phones was estimated to be 80% in 2011, and that number is likely to continue to grow (BCG 2011). Recognizing an opportunity in 2003, the UK Department for International Development approached Vodafone, a major mobile network operator, and its Kenyan affiliate, Safaricom, about developing and piloting a new service in Kenya that would allow micro-credit institutions to disburse loans and receive repayments electronically. After the pilot project, Safaricom noticed that many people were using the service to repay the loans of others (generally as a means of settling a secondary transactional obligation). Safaricom quickly realized the potential of its service as a tool for domestic money transfer. They branded the product "M-Pesa" ("pesa" means "money" in Swahili) and launched the service in March of 2007 (Buku and Meredith 2013). The uptake was rapid – in less than five years, M-Pesa amassed over 12 million customers from among Kenya's population of 45 million (Figure 1.1 depicts M-Pesa's growth). Several years later, neighboring Uganda also experienced rapid growth in mobile money adoption with its own mobile money platforms.

4.2 Transaction mechanics and inventory challenges

As M-Pesa's tagline "send money home" would suggest, the use of mobile money to remit money to a family member or friend from an urban area to a rural area grew quickly. The following is a common use case: the head of a household in a village outside of Kisumu, Kenya

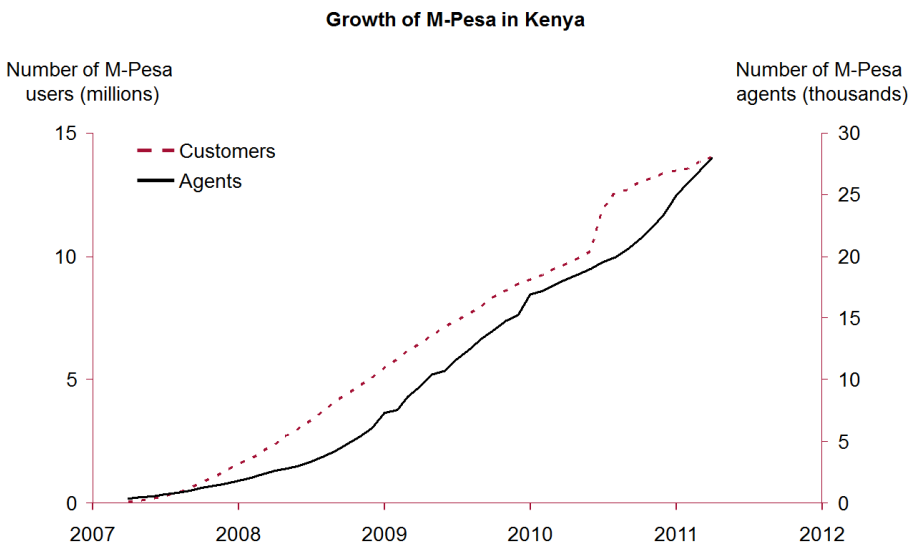


Figure 1.1: Growth of M-Pesa customers and agents in Kenya, adapted from Jack and Suri (2014)

travels to Kenya's major urban center, Nairobi, in order to gain higher-wage employment. Because urban laborers in the informal economy typically get paid in physical cash, it is a non-trivial challenge to send this physical cash to his family still living in a rural area. With mobile money, he can first conduct a "cash-in" transaction with an urban agent in Nairobi – a transaction that involves the agent crediting the laborer's mobile money account with e-float in exchange for physical cash (Stage 1 in Figure A1). In return, the agent is compensated by the mobile money system operator for her services with a commission. The laborer is now able to execute a person-to-person transfer of his new e-float balance to his rural family with his phone (not necessarily a smartphone), in a process similar to sending an SMS (Stage 2 in Figure A1). For this transaction, the laborer pays a transfer fee to the mobile money system operator. After receiving this e-float (less the transfer fee), the laborer's wife can go to a nearby agent to conduct a "cash-out" transaction (Stage 3 in Figure A1) – a transaction where the agent deducts e-float value from a mobile money account and gives the customer cash. For this transaction, the laborer's wife pays a cash-out fee, which is split between the

cash-out agent and the operator.[1]

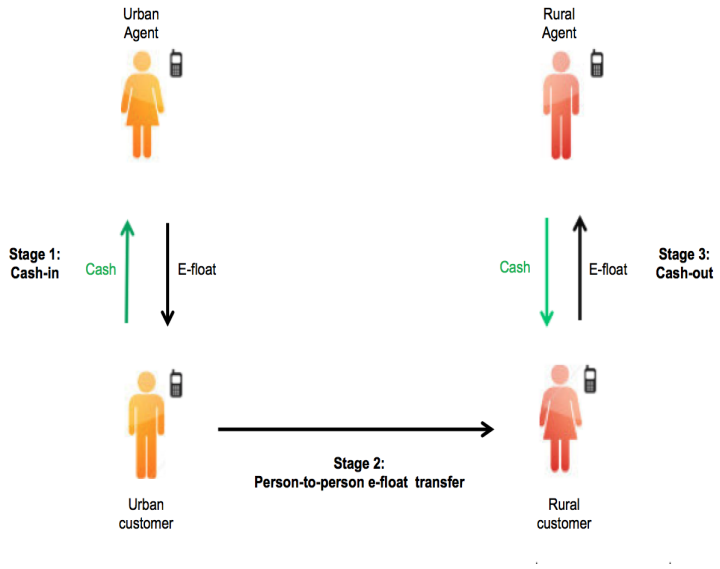


Figure 1.2: Schematic of a typical urban-to-rural person-to-person transfer, adapted from Agrawal (2009)

In order to conduct a cash-in or cash-out transaction, the agent must have inventory of e-float or cash, respectively. Unfortunately, stockouts are an acute problem in mobile money networks (Intermedia 2013). Because stockouts of cash and e-float make it harder for customers to easily convert between the two forms of money, they degrade consumer confidence in the convenient convertibility of e-float.

When an agent stocks out of cash, the customer desiring to cash-out has two options: she can return to the same agent later with the hope that the agent has replenished his inventory of cash, or she can travel to a different agent within the same operator network to cash-out (assuming that this second agent has cash available). When the customer desiring to cash-in experiences a stockout of e-float, the customer has the above two options, as well as a third option: to travel to an agent serving a different operator, assuming that the person he is sending money to also has an account with the competing operator. Note that because

[1]Cash-out commissions are generally larger than cash-in commissions, typically by 50% or more. These commissions are generally determined as an increasing step function of transaction value.

e-float is actual currency, it cannot be "created" on the spot by either the agent or the operator. Each unit of e-float an operator issues must be backed by traditional deposits at a prudentially regulated financial institution. Though moving e-float once it has been issued is clearly easier than moving cash, agents can, and do, still stock out of e-float if they have not been able to procure enough e-float to satisfy demand. This procurement process is not trivial, as it generally requires the agent visiting a bank or other financial Intermediary to "purchase" additional e-float.

5 Hypothesis development

Here we develop hypotheses related to service quality—in terms of pricing transparency and expertise—and competitive intensity, including its interaction with service quality. We also describe hypotheses related to inventory management with respect to competitive intensity and poverty.

5.1 Service quality and demand

We posit that agent pricing transparency and agent expertise, two dimensions of service quality, are important influencers of demand among mobile money agents. In this case, both of these quality elements can be examined through a customer trust lens. As mentioned previously, a common conceptualization of trust is two-dimensional: trust is composed of "benevolence trust" and "competence trust" (Singh and Sirdeshmukh 2000).

Benevolence trust is commonly defined as the "perceived willingness of the trustee to behave in a way that benefits the interests of both parties, with a genuine concern for the partner, even at the expense of profit" (Garbarino and Lee 2003). CICO transaction prices, in general, are set by the operator and are standardized across all agents serving that operator's platform. Furthermore, posting of CICO pricing is both mandated by operators and expected by consumers. The absence of pricing transparency, therefore, may

serve as a warning sign to customers and may degrade benevolence trust. This relates to emerging "disclosure" literature. The key difference is that, in the disclosure literature, the established norm is non-transparency—organizations have the decision whether or not to reveal traditionally unobservable information such as their environmental performance (e.g., Toffel and Reid 2009; Kalkanci and Plambeck 2012), corporate social responsibility (e.g., Dhaliwal et al. 2011; Gamerschlag et al. 2010), or operational processes (e.g., Buell and Norton 2011). In our setting, however, the established norm with respect to pricing is one of transparency—operators mandate that the tariff be posted, and the majority (over 90%) of agents comply with this mandate. We posit that violating the established transparency norm erodes benevolence trust and that non-compliance (i.e., non-transparency) will therefore be related to attenuated demand.

The second component of trust, competence trust, is generally defined as the perceived ability of the firm to deliver services reliably and without flaws (Garbarino and Lee 2003). A customer has competence trust in an agent if the customer believes that the agent has the knowledge and ability to properly conduct CICO transactions (e.g., the agent knows the correct daily transaction limits, identification requirements, and other operator policies regarding the use of mobile money). Though expertise cannot be observed as easily as pricing transparency (presence of a posted tariff sheet), agent expertise (or lack thereof) can be assessed by customers if guidance from agents is either confirmed or discovered to be incorrect. Perceptions of expertise can also be shaped by the opinions and experiences of those in the customer's social network. Greater expertise may thus lead to greater competence trust.

Sun and Lin (2010) find that department store benevolence trust and competence trust both increase customer loyalty (measured on a scale that includes future repeat purchase intent). Because transparency likely engenders benevolence trust in agents, and expertise likely engenders competence trust in agents, we posit that customers reward agents for pricing transparency and expertise, respectively, with higher demand.

Hypothesis 1 *A) Customers reward agents for pricing transparency; demand increases with*

pricing transparency. B) Customers reward agents for expertise; demand increases with agent expertise.

5.2 Service quality, competition, and demand

Intuition suggests that in most settings competitive intensity would attenuate demand, as consumers have a choice between a firm and its competitor(s). Indeed, in their study of auto dealerships, Olivares and Cachon (2009) label this the "sales effect": increased competition decreases a dealer's sales. We posit that this effect is present in the mobile money context as well: increasing competitive intensity (increasing the number of proximate competitors) decreases each agent's demand.

Furthermore, we posit that agents who engender benevolence trust by being more transparent will be able to attract demand away from agents who do not engender such trust. We therefore hypothesize that the interaction between competitive intensity and pricing transparency is positive (i.e., the rewards for transparency increase with competitive intensity). Similarly, we posit that agents who engender greater competence trust by demonstrating expertise will be able to attract demand away from agents who do not. As with pricing transparency, we hypothesize that the interaction between competitive intensity and expertise is positive (i.e., the rewards for expertise increase with competitive intensity).

Hypothesis 2 *A) Agent demand decreases in competitive intensity. B) Competitive intensity increases the rewards from pricing transparency. C) Competitive intensity increases the rewards from expertise.*

5.3 Competition and inventory levels

Olivares and Cachon (2009) decompose the relationship between competition and inventory into a "sales effect" and a "service-level effect". The "sales effect" refers to the notion that proximate competition depresses demand and sales (analogous to our hypothesis H2A)

which in turn reduces desired inventory holdings. What they term the "service-level effect" refers to how competition affects service-level (the proportion of demand that is fulfilled with available inventory). They argue that the theory on this question is mixed. On one hand, greater competitive intensity will drive down price, predicting lower optimal inventory holdings. On the other hand, greater competitive intensity will force firms to compete on service level, driving inventory holding up. In the mobile money context, agents do not compete on price, because price is fixed by the mobile money operator. Thus theory would suggest that controlling for sales, competitive intensity should be positively associated with service-level (and thus, inventory holdings of cash and e-float), as agents could gain (lose) demand from less (more) reliable competitors. In the context of the US auto market, Olivares and Cachon (2009) find that greater competitive intensity among dealers results in greater inventory holdings. However, research across many disciplines has noted that the developed world is vastly different from the developing world: consumers have far less purchasing power (Demirguc-kunt 2012), margins are generally far lower (Prahalad and Hammond 2002), and legal systems are less robust (Helms 2006), among many other differences. Even though the operating environment for mobile money agents in East Africa is fundamentally different from auto dealers in the United States, we hypothesize that the finding of Olivares and Cachon (2009) that controlling for sales, inventory holdings increase with competition, is generalizable to this different context.

Hypothesis 3 *Agent inventory holdings of cash and e-float increase with competitive intensity.*

5.4 Poverty and inventory levels

The relationship between the level of poverty in an agent's catchment area and inventory holdings of cash is complicated. On one hand, the poor have definitionally lower net worth and cash-flow, which might suggest that their transaction sizes are smaller than the average mobile money customer. Controlling for the number of transactions, then, smaller transac-

tion sizes would likely result in a decreased need for inventory of cash. On the other hand, given that the initial "killer app" of mobile money was domestic remittance ("send money home"), most often from areas of comparative wealth (largely urban areas) to areas of relative poverty (rural areas, in many cases), we might expect that there would be significantly more cash-out demand than cash-in demand in high-poverty areas. Indeed, Eijkman et al. (2009) observe that for five representative agents in rural areas in 2009, total cash-out transaction value was much higher than total cash-in transaction value. This imbalanced demand suggests that agents in high-poverty[2] areas may need to carry more cash inventory, because they do not have as much cash-in transaction demand that would otherwise contribute to their inventory of cash. We thus posit that, controlling for other factors, agents serving high-poverty areas carry more cash than agents serving lower-poverty areas.

Theory is mixed on the relationship between e-float inventory and poverty levels. On one hand, agents serving high-poverty areas sometimes experience high-magnitude demand for e-float (cash-in).[3] These large cash-in demand arrivals would necessitate large e-float holdings to satisfy this demand. These arrivals may be attributable to the surprisingly complex financial lives of the poor. Collins et al. (2009) note that many poor people have income that does not come in a steady stream but is "lumpy" (e.g., farmers have no income until they are able to harvest and sell their crops). This "lumpiness" necessitates savings in some form to smooth consumption over periods of no income. Economides (2015) finds evidence that Tanzanian mobile money accounts are sometimes used as a secure savings alternative to keeping cash at home. This could possibly explain some high-value cash-in behavior in high-poverty areas, and thus drive up the need for greater e-float holdings by agents. On the other hand, however, Intermedia (2013) notes that poor households are less likely to own a phone, a SIM card, and use mobile money, suggesting that the poor lag in adoption

[2]We note that high-poverty areas and rural areas are not always equivalent (there are urban areas with high-poverty, such as slums, and there are also rural areas that are relatively wealthy, such as vacation communities), but it has been noted that over 70% of the world's poor live in rural areas (International Fund For Agricultural Development 2011). As will be seen later, our data will allow us to tease apart population density and poverty level.

[3] Insight from informal interviews by the authors in Kenya and Uganda in 2014 and 2015.

of mobile technology. It is therefore likely that adoption of e-float for usage as a means for purposes other than domestic remittance, such as utility payments, tax payments, and depositing into bank accounts is lower among the poor than the general population. This decreased demand would lead to agents in poor areas carrying less e-float. Furthermore, because agents in high-poverty areas likely have very limited budgets to invest in inventory of cash and e-float, and because we posit that agents in high-poverty areas carry relatively more cash, agents might be forced to carry less e-float. The notion that cash inventory might take priority over e-float inventory is supported by the fact that cash-out commissions are generally larger than cash-in commissions, typically by 50% or more. Because agent budget constraints are likely to be very salient in this context, we posit that agents in high-poverty areas carry less e-float inventory than their peers who serve lower-poverty areas.

Hypothesis 4 *Agent inventory holdings of cash increase with the proportion of population in poverty in the catchment area, and inventory holdings of e-float decrease with the proportion of population in poverty in the catchment area.*

6 Data and empirical specification

To test these hypotheses, we combine three data sources, each sampling Kenya and Uganda. First, we use data from an in-person, cross-sectional survey of over 3,000 mobile money agents. We combine this data with the precise locations of over 68,000 financial access and transportation points (including mobile money agents, banks, and bus stands) in the two countries. Finally, we integrate granular (per square kilometer) spatial estimates of population and poverty (with the latter defined as the number of people living on less than $2 per day) in each given square kilometer.

6.1 Agent network survey

We use data from large surveys conducted by the Helix Institute of Digital Finance, an organization that provides training and data to digital financial service providers in the developing world. We use two of their recent surveys conducted in Kenya and Uganda throughout 2013 (McCaffrey et al. 2014, Githachuri et al. 2014). Between 1,500 and 2,000 mobile money agents were surveyed in-person in each country by a team of professional surveyors. The surveys were designed to be nationally representative through three steps. First, the total number of agents to be surveyed within a given country was determined based on financial constraints. Next, the number of surveys were apportioned to each operator in each district (analogous to a county in the United States) by taking the product of the total number of surveys, the ratio of district population to national population and the operator's national market share. Finally, in each district, a local team lead identified a representative enumeration area (EA). Once the EA was identified, surveyors were given mutually exclusive routes to walk, applying a "left-hand-rule" walk pattern, skipping a pre-determined number of agents on the left-hand side before attempting to interview the next agent.[4] This pre-determined number was based on the number of agents the given operator had in the district, which was derived from the spatial census data described below. Each survey lasted between 30-60 minutes. As with many in-person surveys, there was a very high response rate (roughly 95%). All data were point estimates of steady-state levels and were self-reported by the agents or observed directly by the surveyor. The interview covered a wide array of topics, including demographics and location (latitude and longitude), products and services offered, inventory management, revenue and commission structure, platform performance, training, monitoring, and support. From these data we glean independent variables $PricingTransparency$ and $Expertise$, as well as controls $Male$, $Tills$, $Sunday$, $Dedicated$, and $OperatorG$. Each of these variables is described in detail in the econometric

[4]The "left (right)-hand-rule" is a common methodology for selecting stationary interviewees for in-person data collection.

specification subsection. From the 3,831 survey responses of mobile network operators' CICO agents, we obtain our final cross-sectional survey dataset for testing our hypotheses relating to demand of 3,215 observations by excluding observations with missing values of the control variable *Sunday* (179), missing estimates of transactions denied due to system failures (266), missing estimates of transactions denied due to stockouts (148), and missing estimates of successfully completed transactions (23). We obtain our final cross-sectional survey dataset for testing our hypotheses relating to inventory of cash and e-float of 2,770 and 2,844, respectively, by dropping observations with missing values of the variable *Sunday* (179), missing estimates of the average number of successful transactions (23), and missing estimates of cash inventory (982) and e-float (909), respectively. We test (and find support for) the robustness of our results to the inclusion of records with missing values through multiple imputation. We discuss and provide the results of these robustness tests in Section 7 and the Appendix.

6.2 Spatial census of financial access points

We use data from fspmaps.com, a financial access mapping effort funded by the Bill & Melinda Gates Foundation (BMGF). In 2012 and 2013, Brand Fusion, a research company based in South Africa, employed teams in Kenya and Uganda to collect geographic data. These teams canvassed both countries for financial service access points, recording their locations with GPS-enabled smartphones. The raw data behind fspmaps.com consists of the exact latitude and longitude coordinates of over 68,000 mobile money agents (with operator details), bank branches, and bus stands in Kenya and Uganda. Table 1.1 shows the collected geographic data breakdown by type and country. The teams also supplemented every geographic data point with pictures of the financial access point for validation. By combining the surveyed agent locations with the spatial census data through a process known as buffer analysis, we generate the independent variable $DirectCompetition$, as well as controls $IndirectCompetition$, $Bank1km$, and $Bus1km$. Buffer analysis is a spatial analysis tech-

nique that generates circles of specified radii around points of interest (such as mobile money agents) and then calculates the number of other features of interest (such as the number of other agents and population counts) that fall within these circles.

	Kenya	**Uganda**	**Total**
Mobile money agents	48,524	17,889	66,413
Bank branches	1,221	477	1,698
Bus stands	613	121	734

Table 1.1: Table of geotagged points by country

6.3 High-resolution spatial demographics

Lastly, we draw upon spatial population and income data generated by WorldPop, an organization focused on creating high-quality maps for the humanitarian sector. In order to create these maps, WorldPop combines three sources of data: satellite imagery (Radarsat-1 country mosaics and Landsat Enhanced Thematic Maps); the Africover database containing geographic data on roads, land cover, and bodies of water; and country-level census data. The resulting integrated model generates precise population estimates for every square kilometer of Africa (Tatem et al. 2007). WorldPop used a similar method to develop a high-resolution spatial data layer of the population in poverty—those living on less than $2 per day. To generate its spatial poverty data, WorldPop employs a process known as "Bayesian geostatistics" to integrate geocoded well-being surveys conducted by USAID (Demographic and Health Survey) and the World Bank (Living Standards Measurement Survey). The resulting data layer, like the general WorldPop data layer, has a resolution of 1 square kilometer (Tatem 2013). Figure 1.3 depicts a spatial data layer of the poor population in Kenya's Nyanza Province. The green circles depict buffers of 5, 10, and 15 kilometers (for ease of viewing) around bank branches. By combining the surveyed agent locations with the spatial demographic data, we generate control variable $PopK1km$ (population in thousands within a 1km radius of the agent) and independent variable $PovRatio1km$ (the fraction of the

population within 1km who live on less than $2 a day) through buffer analysis.

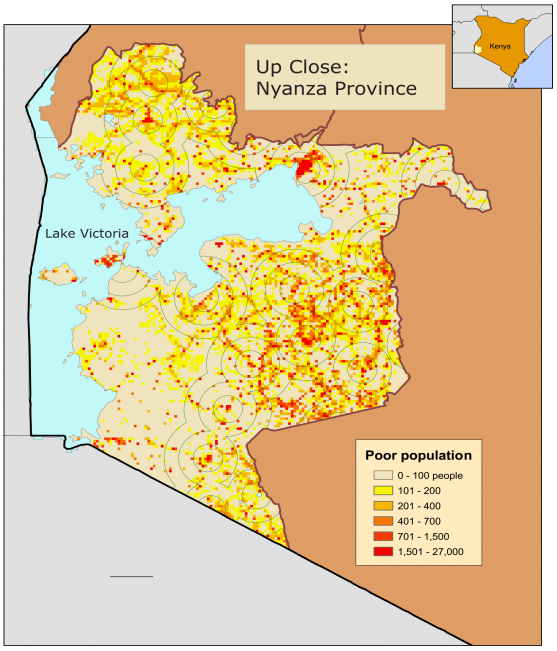


Figure 1.3: Spatial poor population and bank buffers in Kenya's Nyanza Province, adapted from BMGF (2012)

6.4 Econometric specifications

We use two different sets of OLS regressions to test our hypotheses: one to test our hypotheses related to demand, and the other to test our hypotheses related to inventory holding.

Dependent variables

From the survey data, for each agent we have point estimates of the number of successful transactions on an average day (T), the number of transactions denied due to stockouts (S), and the number of transactions denied due to system failure (F). We define demand as $D = T + S + F$. To test our hypotheses related to demand (Hypotheses 1 and 2), we conduct an OLS regression on the natural logarithm of demand ($log(Demand)$).

Also from the survey data, we have agent-level point estimates for each agent's cash and e-float inventory in Ugandan Shillings and Kenyan Shillings, respectively. For comparability, we convert these estimates to United States Dollars (USD) using year-end 2013 exchange rates. To test our hypotheses related to inventory (Hypotheses 3 and 4), we conduct an OLS regression on the natural logarithm of cash inventory ($log(AvgCashUSD)$) and e-float inventory ($log(AvgEFloatUSD)$), respectively.

We log demand and inventory for two principle reasons. First, demand and inventory cannot be negative in this context (to be deemed active, all agents must have at least one transaction per day, and agents generally cannot "borrow" cash or e-float). Second, using the log of demand and inventory simplifies the interpretation of estimated coefficients: a one unit change in an independent variable corresponds to a β*100% change in demand and inventory, respectively.

Independent and interaction variables

Our independent variables for the first model examing demand are $PricingTransparency$, $Expertise$, and $DirectCompetition$. $PricingTransparency$ is an indicator variable that captures whether an agent has posted the tariff sheet (listing of transaction prices) prominently in the shop, with a 1 indicating that the agent posted the tariff sheet and a 0 indicating that the agent did not post the tariff sheet prominently in the shop. $Expertise$ is an indicator variable that captures whether the agent correctly answered a difficult question about mobile money policy, with a 1 indicating that the agent correctly responded to the question (greater expertise) and a 0 indicating that they did not respond correctly to the question (lesser expertise).[5] $DirectCompetition$, our measure for competitive intensity, indicates the number of other agents primarily serving the same operator as a particular agent within 1 kilometer of that agent.[6] To estimate the impact of competition on service quality, we

[5]In both Kenya and Uganda, this difficult question was to identify the exact maximum amount of e-float a customer could keep in his/her account.

[6]The distance threshold of 1km was based on informal interviews with mobile money customers in Uganda and Kenya, who indicated they, in general, would walk no more than 10-15 minutes (roughly 1km) after

include interactions between $DirectCompetition$, $PricingTransparency$, and $Expertise$.

For the second set of models examining inventory (one for cash holdings and the other for e-float holdings), we again use $DirectCompetition$ as an independent variable. We also use $PovRatio1km$ as an independent variable, where $PovRatio1km$ is the fraction of the total population within 1km of the agent who live on less than $2 a day.

Control variables

We include several control variables that could plausibly affect demand and inventory holdings. In the first set of models, we include the following control variables: $Male$ indicates agent gender, with a 1 indicating male and a 0 indicating female. $Dedicated$ indicates whether the agent is solely dedicated to the mobile money business or whether they operate another retail business on the side, with a 1 indicating dedication solely to mobile money transactions and a 0 indicating that the agent operates other businesses in parallel. $Sunday$ indicates whether the agent is open for business on Sunday, with a 1 indicating that the agent operates on Sunday and a 0 indicating that the agent does not operate on Sunday. $Tills$ is a variable that takes integer values greater than 0 to indicate the number of agent tills (i.e., "virtual cash registers") that the agent operates. $PopK1km$ is the population (in thousands) within a 1 kilometer radius of an agent. We include $PovRatio1km$, described above, as a control variable in the first set of models. Additionally, $IndirectCompetition$ is the number of agents primarily serving other operators in a 1 kilometer radius of an agent. $Bank1km$ is the number of bank branches in a 1 kilometer radius of an agent; banks are a major resource in helping agents rebalance inventory (so agents can get cash and/or e-float). $Bus1km$ is the number of bus stops in a 1 kilometer radius of an agent; this is a proxy for the difficulty of transit in the vicinity of the agent. $Uganda$ is a binary variable that takes a value of 1 if the surveyed agent operates in Uganda, and a 0 if the agent operates in Kenya. Finally, $OperatorG$ represents a set of 5 (+1) indicator variables for each operator/brand

being unsatisfied with one agent to access a different agent. Also, population and poverty estimates used in this analysis have a granularity of 1 square km.

represented in the sample, which captures differences across operators not accounted for by our independent variables or the control variables above.

The second set of models, examining inventory holdings of cash and e-float, control for the average number of successful transactions T that an agent conducts in a day, as well as many of the control variables used in the first model that also could plausibly affect inventory holdings: $PopK1km$, $IndirectCompetition$, $Bank1km$, $Busk1km$, $Dedicated$, $OperatorG$, and $Uganda$[7].

7 Results

Summary statistics are presented in the appendix. Table A1 includes means, standard deviations, and differences in means. Table A2 presents pair-wise correlation coefficients. Generally speaking, Kenya leads Uganda in mobile money market development. Though agents' estimated demand (the sum of successful transactions, transactions denied due to stockouts, and transactions denied due to system failure) is greater in Uganda than Kenya, Kenyan agents conduct more successful transactions. Ugandan agents thus experience more stockouts and system failures than Kenyan agents. With regard to $PricingTransparency$, the vast majority of agents are transparent with pricing, led by Kenyan agents. With regard to $Expertise$, approximately 81% of agents passed the expertise test, with Kenyan agents lagging slightly. Direct competition among agents, as measured by the number of other agents that directly compete with each agent, is lower in Uganda than in Kenya. This is reversed for the number of indirectly competing agents (agents serving other operators), where Ugandan agents have more indirectly competing agents within one kilometer on average than their Kenyan counterparts.

The sample contained slightly more female agents than male agents, and this gender ratio does not differ significantly across the two countries. Additionally, just under half of

[7]For robustness, we tested for non-linear effects of direct competition on agents' demand through the addition of $DirectCompetition^2$. No non-linear effects were observed; $DirectCompetition^2$ proved insignificant in all cases, and its inclusion did not meaningfully affect the direction or magnitude of our results.

the sampled agents operated on Sunday, with Ugandan agents more likely to operate on Sunday than Kenyan agents. Just under half of the agents in the sample run a dedicated mobile money business. We also note that Kenyan agents in the sample locate on average in more densely populated areas than Ugandan agents, while Ugandan agents are much more likely to be located in areas of high poverty concentration than Kenyan agents. Finally, we see that Kenyan agents in the sample on average have at least twice as many bank branches and bus stands within a one kilometer radius than Ugandan agents. This is reflective of the fact that Kenya is more economically developed than Uganda, and thus its financial and transportation infrastructure is more developed than its neighbor.

We present the results from four OLS regressions relating to our transaction demand hypotheses in Table A3. We include the set of control indicator variables representing operator brands ($OperatorG$) in regressions but exclude them from the tables to preserve operator confidentiality. The first model excludes interactions with $DirectCompetition$. The next two models each include a single interaction term, and the final model includes both interaction terms.[8] The discussion that follows is based on the full model unless otherwise noted. Robust standard errors are reported, as the Breusch-Pagan test indicated the presence of heteroskedasticity.

Examining our hypotheses relating to inventory holdings of cash and e-float, the results from two OLS regressions are presented in Table A4. Again, robust standard errors are reported, as the Breusch-Pagan test indicated the presence of heteroskedasticity.

7.1 Service quality, competition, and demand

Our first set of hypotheses focused on relationships between service quality (pricing transparency and expertise) and demand. The data support hypothesis 1A. Pricing transparency (non-transparency) is associated with an increase (decrease) in the agent's demand. The

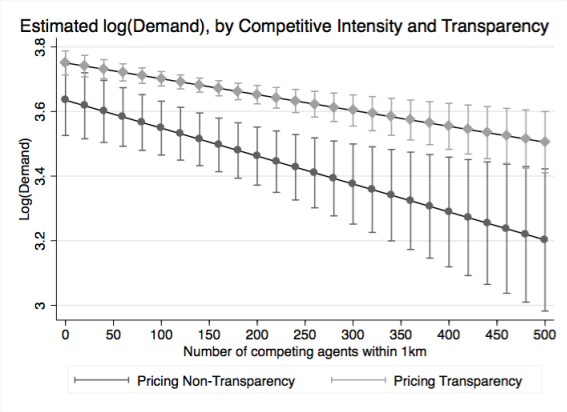
[8]Models 2 and 3 are provided so that estimated coefficients for the interaction terms in the full model can be compared to those in the models that isolate a single interaction. This is done to ensure that results in the full model are not skewed by potential multi-colinearity, given that both interaction terms include competitive intensity.

data also support hypothesis 1B. Agent expertise is associated with greater agent demand. The magnitudes of these relationships are also notable; the presence of a tariff sheet is associated with a 12% increase in demand and the ability to answer a difficult question about mobile money policy is associated with an 8% increase in demand.

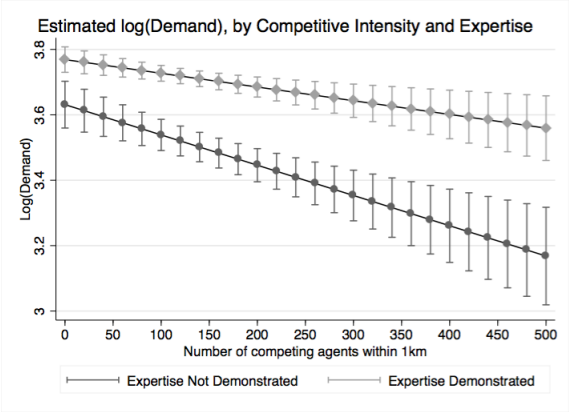
Regarding competitive intensity, we note that the coefficient on *DirectCompetition* is negative and statistically significant. The data support hypothesis 2A, that demand decreases in competitive intensity. The data indicate that competitive intensity also interacts with expertise but not with pricing transparency, as the coefficient on *DirectCompetition* x *PricingTransparency* is not statistically significant. Hypothesis 2B is therefore not supported. Finally, the coefficient on *DirectCompetition* x *Expertise* is significant and positive, supporting hypothesis 2C: competitive intensity enhances the relationship between expertise and demand. These latter two relationships are illustrated in Figures 1.4a and 1.4b. In Figure 1.4a, we see the estimated $log(Demand)$ values as *PricingTransparency* and *DirectCompetition* are varied. The main relationship between pricing transparency and demand, the gap between the two lines, is clearly visible, but there is not a statistically significant difference in the slopes of the two lines (which would be evidence of an interaction between competitive intensity and pricing transparency). Figure 1.4b illustrates estimates of *logDemand* as *DirectCompetition* and *Expertise* are varied. Analogously to the relationship between pricing transparency and demand, we clearly see the direct relationship between expertise and demand in the gap between the two lines. However in Figure 1.4b a statistically significant difference in the slopes of the two lines is also visible, evidence of an interaction between competitive intensity and expertise.

7.2 Competition, poverty, and inventory levels

We now turn to our hypotheses relating to inventory holding. We note that the coefficients on *DirectCompetition* in both cash and e-float inventory models are statistically significant.



(a) Pricing transparency



(b) Expertise

Figure 1.4: Plots of predicted demand as a function of competitive intensity and service quality measures

Hypothesis 3 is thus supported: agent inventory holdings of both cash and e-float increase with competition.

With respect to poverty level in agent catchment areas, we note that the coefficient on $PovRatio1km$ is positive and significant in the model for cash inventory, while the coefficient of $PovRatio1km$ in the e-float inventory model is not significant. We thus find partial support for hypothesis 4: agent inventory holdings of cash increase with the poverty rate in the catchment area. However, we do not observe evidence of a negative relationship between e-float inventory holding and poverty.

7.3 Robustness checks

To test whether our results are robust to the inclusion of observations with missing values, we use multiple imputation (predictive mean matching) to predict missing values of specific variables within observations. We then estimate $log(demand)$, $log(CashUSD)$, and $log(EFloatUSD)$ using these augmented sets of survey observations. Results for these tests are presented in Tables A5 and A6 in the Appendix. While estimated coefficient values vary slightly relative to the base models, the results presented above largely hold (both directionally and in significance). The only exception is our finding that cash inventory

holding increases with competitive intensity became slightly weaker in significance level in the multiple imputation model, though the relationship is still significant at the $p < 0.1$ level.

8 Discussion and Conclusion

Our results provide insight on the relationships between service quality, competition, inventory, and demand in a new and important operations context that is particularly relevant to the BoP community.

While our analyses do not result in causal claims, this paper does document positive relationships between pricing transparency and demand as well as expertise and demand. In the mobile money context, agents who are transparent with their prices and agents who are more knowledgeable about mobile money policies experience greater demand. We also find evidence that expertise interacts with competition in the mobile money context: the rewards for expertise increase when an agent faces greater competitive intensity. However, our data suggest that pricing transparency does not interact in a statistically significant way with competitive intensity. This suggests that pricing transparency may not be a dimension of competition in mobile money; rather, some fraction of the (potential) customer base may simply choose not to do business with non-transparent agents, whether or not there are other competitors around. Using the operations strategy terminology coined by Hill (1993), pricing transparency in this context may act as an "order-qualifier" for a segment of the market—these customers may not consider transacting with agents who are not transparent with prices. Our data suggest that expertise, on the other hand, may act as an "order-winner" for a segment of customers—all else equal, these customers may choose more expert agents among their set of "qualified" agents. We do note that there may be reverse causality at play here: greater demand may naturally lead to greater expertise because the agent can reap the rewards of "learning by doing." Similarly, an agent may try to compensate for low

demand by attempting to overcharge customers (which is more possible when the tariff sheet is not displayed). However, in both cases, the association itself is informative, regardless of causality.

This paper also documents positive relationships between competitive intensity and inventory of both cash and e-float in the mobile money context. This observation supports the finding of Olivares and Cachon (2009), who show that inventory of US auto dealers increases as proximate competition increases. They postulate that there is a"service-level" effect — firms compete on service-level to attract business away from competitors. Our results provide support for such a "service-level effect" in the mobile money context as well, with agents carrying more inventory to increase their service level — perhaps as another "order-winner".

We also find that an agent's cash inventory increases with the proportion of the population in their catchment area living below the poverty level. While there may be many reasons for increased cash-holding, the most likely in this context is that remittances to higher-poverty areas are a key driver of mobile money demand; this cash-out demand thus requires agents in high-poverty areas to carry sufficient cash inventory. Given that agents have limited budgets and that cash inventory increases with poverty level, we might expect that agents in high-poverty areas would carry less e-float. We do not find support for this notion, however. This is possibly because agents in high-poverty areas might experience significant demand (at least at certain times) for e-float that is not always off-set by cash-out demand (which would generate e-float inventory). The consequence of this finding is important to agents, operators, and policymakers interested in the poor: because agents in high-poverty areas carry more cash and not a correspondingly lower level of e-float than agents not in high-poverty areas, controlling for other factors (such as the number of successful transactions), these agents have a less compelling business case. There may be some efficiency gains to be realized with better agent inventory control policies — which is an area for future work where the Operations Management community is uniquely positioned to contribute.

Our analyses also provide additional encouragement for operators exploring policies de-

signed to enhance the customer experience, such as mandatory and strictly enforced tariff sheet posting and rigorous training to ensure knowledge of policies. Finally, this work may spur operators to consider adjusting incentives for agents based on their location. For example, an operator might consider increasing the commissions for cash-in and cash-out transactions for agents in higher poverty areas. This is because the business case appears worse in high-poverty areas due to the tendency of agents in these areas to carry greater levels of inventory. This could be the basis for future work that explores operator-agent contracting and incentives.

This analysis, like other analyses using survey data, suffers from potential bias: demand and inventory estimates were self-reported estimates by the agent. To the extent that agents thought their operators might see the results (even though they were promised confidentiality by the third party research firm, and this confidentiality was honored), agents' estimates of demand and inventory might be biased. However, there is no evidence to suggest that results here would be affected by self-reporting; i.e., there is no evidence that some agents might be more biased than others. Additionally, our dataset limited our measure of demand to an estimate of the raw count of transactions. An interesting additional dimension of analysis would consider the value of these transactions as well.

8.1 Conclusions

Mobile money is a rapidly growing industry that has potential to dramatically improve the lives of the poor in many ways. We begin to explore this industry by examining mobile money demand drivers. We find that agents who are transparent with transaction pricing experience relatively greater demand. Agents who are relatively more knowledgeable not only experience greater demand, but also seem to reap greater benefit from their expertise in the face of competition. We also explore the relationships between competition, poverty, and inventory holdings of cash and e-float. We find that agents who face more competitive intensity carry more inventory of both cash and e-float, while agents who serve high-poverty

areas hold more cash but not a smaller amount of e-float.

Acknowledgments

Sincerest thanks to Mike McCaffrey and Leena Anthony from MicroSave, a founding partner of The Helix Institute of Digital Finance, for being so helpful and enthusiastic about this research. Thanks also to the Research Solutions Africa surveyor team. Many thanks to Eric KramakSemp and Todd Slind of SpatialDev for facilitating access to the raw spatial data behind fspmaps.com, as well as Andy Tatem and the WorldPop team for the spatial population and poverty data. The authors are also very grateful to Jake Kendall and Daniel Radcliffe of the Bill & Melinda Gates Foundation for connecting us with MicroSave and Karina Nielsen of CGAP for connecting us with SpatialDev. Thanks also to Giovanni Zambotti and Stacy Bogan of Harvard University's Center for Geographic Analysis, as well as Xiang Ao and Jonathon Polit of Harvard Business School's Research Computing Services for their invaluable help in preparing and analyzing data. The authors are grateful to the Harvard Business School Doctoral Programs Office for generous research support.

9 Appendix

Table A1: Summary table of means and differences (standard devisations in brackets)

	Overall	**Kenya**	**Uganda**	**Difference**
log(Demand)	3.68	3.65	3.71	-0.06**
	[0.70]	[0.71]	[0.68]	
log(CashUSD)	5.66	5.66	5.66	0
	[0.92]	[0.95]	[0.89]	
log(EFloatUSD)	5.9	5.87	5.92	-0.05+
	[0.85]	[0.93]	[0.76]	
T	35.17	39.32	29.61	9.71**
	[25.59]	[28.47]	[19.81]	
Expertise	0.81	0.77	0.85	-0.08**
	[0.40]	[0.42]	[0.36]	
PT	0.93	0.95	0.89	0.06**
	[0.26]	[0.22]	[0.31]	
DirectCompetition	134.09	138.23	128.53	9.7+
	[167.71]	[153.04]	[185.52]	
Dedicated	0.45	0.44	0.45	-0.01
	[0.50]	[0.50]	[0.50]	
Sunday	0.46	0.43	0.49	-0.06**
	[0.50]	[0.50]	[0.50]	
PovRatio1km	0.19	0.08	0.34	-0.26**
	[0.18]	[0.11]	[0.15]	
PopK1km	26.87	31.05	21.25	9.8**
	[33.88]	[40.94]	[19.57]	
IndirectCompetition	74.35	68.68	81.99	-13.31*
	[194.68]	[231.73]	[128.78]	
Bank1km	8.05	10.34	4.96	5.38**
	[16.06]	[19.67]	[8.26]	
Bus1km	4.3	6.54	1.3	5.24**
	[15.22]	[19.72]	[2.20]	
Male	0.43	0.44	0.43	0.01
	[0.50]	[0.50]	[0.50]	
Tills	1.12	1.04	1.23	-0.19**
	[0.59]	[0.33]	[0.81]	

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A2: Cross-correlation table

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. log(Demand)	1.00															
2. log(CashUSD)	0.46	1.00														
3. log(EFloatUSD)	0.45	0.64	1.00													
4. TxAvg	0.74	0.41	0.38	1.00												
5. Expertise	0.11	0.11	0.10	0.10	1.00											
6. PT	0.06	0.10	0.10	0.06	0.05	1.00										
7. DirectComp	0.16	0.20	0.20	0.16	-0.06	0.01	1.00									
8. Dedicated	0.12	0.11	0.09	0.08	-0.03	0.02	0.08	1.00								
9. Sunday	0.15	0.07	0.05	0.13	-0.00	0.01	0.01	-0.05	1.00							
10. PovRatio1km	0.04	0.03	0.03	-0.12	0.11	0.00	-0.32	0.03	0.05	1.00						
11. PopK1km	0.03	0.04	0.02	0.06	-0.08	-0.05	0.47	0.03	0.06	-0.39	1.00					
12. IndirectComp	-0.19	-0.11	-0.13	-0.14	0.01	-0.01	-0.07	-0.05	-0.06	-0.10	0.10	1.00				
13. Bank1km	-0.05	0.01	0.00	0.00	0.01	0.01	0.33	-0.01	-0.06	-0.27	0.19	0.75	1.00			
14. Bus1km	-0.10	-0.03	-0.05	-0.04	-0.01	0.05	0.09	-0.02	-0.04	-0.16	0.07	0.79	0.80	1.00		
15. Male	0.05	0.07	0.05	0.08	0.07	-0.02	0.02	-0.03	0.13	0.02	0.04	0.01	0.01	0.00	1.00	
16. Tills	0.12	0.11	0.11	0.09	0.04	0.01	0.01	0.02	0.05	0.09	0.06	0.01	-0.00	-0.01	0.00	1.00

Table A3: Relationships between service quality, competition, and demand: OLS regression results

	log(Demand)	log(Demand)	log(Demand)	log(Demand)
Expertise	0.148**	0.149**	0.0848*	0.0863*
	(0.0263)	(0.0263)	(0.0395)	(0.0395)
PT	0.163**	0.123*	0.166**	0.127*
	(0.0443)	(0.0584)	(0.0443)	(0.0583)
DirectComp	-0.000530**	-0.000866**	-0.000931**	-0.00125**
	(0.000121)	(0.000279)	(0.000201)	(0.000324)
Dedicated	0.0810**	0.0806**	0.0833**	0.0829**
	(0.0213)	(0.0213)	(0.0213)	(0.0213)
Sunday	0.171**	0.170**	0.172**	0.171**
	(0.0214)	(0.0214)	(0.0214)	(0.0214)
PovRatio1km	-0.0935	-0.0928	-0.0882	-0.0875
	(0.0958)	(0.0959)	(0.0960)	(0.0960)
PopK1km	0.00126**	0.00126**	0.00132**	0.00131**
	(0.000342)	(0.000342)	(0.000343)	(0.000343)
IndirectComp	0.0000843	0.0000864	0.0000851	0.0000871
	(0.000125)	(0.000125)	(0.000125)	(0.000125)
Bank1km	0.00920**	0.00918**	0.00904**	0.00902**
	(0.00169)	(0.00169)	(0.00170)	(0.00170)
Bus1km	-0.00635**	-0.00638**	-0.00618**	-0.00622**
	(0.00167)	(0.00167)	(0.00168)	(0.00168)
Male	0.0812**	0.0803**	0.0818**	0.0809**
	(0.0215)	(0.0215)	(0.0215)	(0.0215)
Tills	0.135**	0.135**	0.135**	0.136**
	(0.0262)	(0.0263)	(0.0264)	(0.0264)
Uganda	0.624**	0.620**	0.630**	0.626**
	(0.0672)	(0.0672)	(0.0672)	(0.0672)
PT*DirectComp		0.000369		0.000356
		(0.000281)		(0.000285)
Expertise*DirectComp			0.000505*	0.000501*
			(0.000204)	(0.000205)
_cons	2.135**	2.173**	2.173**	2.210**
	(0.0763)	(0.0847)	(0.0781)	(0.0862)
N	3215	3215	3215	3215
adj. R^2	0.281	0.281	0.282	0.282

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A4: Relationships between competition, poverty, and inventory: OLS regression results

	log(CashUSD)	log(EFloatUSD)
DirectComp	0.000420*	0.000513**
	(0.000178)	(0.000160)
PovRatio1km	0.419**	0.133
	(0.123)	(0.119)
T	0.0125**	0.0107**
	(0.00110)	(0.00101)
PopK1km	0.00100+	-0.000373
	(0.000568)	(0.000476)
IndirectComp	0.0000200	-0.000180
	(0.000179)	(0.000162)
Dedicated	0.0597+	0.0732*
	(0.0311)	(0.0286)
Bank1km	0.00580*	0.00439*
	(0.00235)	(0.00223)
Bus1km	-0.00216	-0.000898
	(0.00247)	(0.00242)
Uganda	0.421**	0.561**
	(0.0841)	(0.0825)
_cons	4.415**	4.866**
	(0.0713)	(0.0743)
N	2770	2844
adj. R^2	0.238	0.216

Standard errors in parentheses

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A5: Relationships between service quality, competition, and demand: OLS regression results (multiple imputation model)

	log(Demand)	log(Demand)	log(Demand)	log(Demand)
Expertise	0.161**	0.162**	0.101**	0.102**
	(0.0246)	(0.0246)	(0.0366)	(0.0366)
PT	0.209**	0.171**	0.212**	0.175**
	(0.0398)	(0.0525)	(0.0397)	(0.0524)
DirectComp	-0.000417**	-0.000743**	-0.000819**	-0.00113**
	(0.000112)	(0.000257)	(0.000190)	(0.000304)
Dedicated	0.0899**	0.0898**	0.0914**	0.0914**
	(0.0197)	(0.0197)	(0.0197)	(0.0197)
Sunday	0.179**	0.178**	0.180**	0.179**
	(0.0208)	(0.0208)	(0.0208)	(0.0208)
PovRatio1km	-0.136	-0.135	-0.129	-0.128
	(0.0875)	(0.0875)	(0.0876)	(0.0876)
PopK1km	0.00108**	0.00109**	0.00114**	0.00114**
	(0.000324)	(0.000325)	(0.000324)	(0.000325)
IndirectComp	0.000184+	0.000184+	0.000185+	0.000185+
	(0.000110)	(0.000110)	(0.000110)	(0.000110)
Bank1km	0.00808**	0.00810**	0.00795**	0.00797**
	(0.00149)	(0.00149)	(0.00149)	(0.00149)
Bus1km	-0.00753**	-0.00756**	-0.00739**	-0.00742**
	(0.00150)	(0.00150)	(0.00150)	(0.00150)
Male	0.0754**	0.0748**	0.0760**	0.0753**
	(0.0200)	(0.0200)	(0.0200)	(0.0200)
Tills	0.137**	0.137**	0.138**	0.138**
	(0.0253)	(0.0254)	(0.0254)	(0.0255)
Uganda	0.590**	0.587**	0.595**	0.592**
	(0.0587)	(0.0587)	(0.0586)	(0.0587)
PT*DirectComp		0.000356		0.000347
		(0.000258)		(0.000261)
Expertise*DirectComp			0.000502*	0.000499*
			(0.000195)	(0.000195)
_cons	2.050**	2.086**	2.087**	2.122**
	(0.0691)	(0.0764)	(0.0713)	(0.0784)
N	3831	3831	3831	3831

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A6: Relationships between competition, poverty, and inventory: OLS regression results (multiple imputation model)

	log(CashUSD)	log(EFloatUSD)
DirectComp	0.000342^{+}	0.000481**
	(0.000186)	(0.000171)
PovRatio1km	0.425**	0.156
	(0.131)	(0.114)
T	0.0113**	0.0101**
	(0.00103)	(0.000836)
PopK1km	0.000986	-0.000271
	(0.000626)	(0.000600)
IndirectComp	0.0000731	-0.000184
	(0.000187)	(0.000167)
Dedicated	0.0656*	0.0848**
	(0.0282)	(0.0287)
Bank1km	0.00594*	0.00480*
	(0.00236)	(0.00220)
Bus1km	-0.00295	-0.00129
	(0.00234)	(0.00230)
Uganda	0.418**	0.548**
	(0.0830)	(0.0836)
_cons	4.425**	4.875**
	(0.0706)	(0.0724)
N	3831	3831

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Chapter 2

Inventory Management for Mobile Money Agents in the Developing World

1 Abstract

Mobile money systems, platforms built and managed by mobile network operators to allow money to be stored as digital currency, have burgeoned in the developing world as a mechanism to transfer money electronically. Mobile money agents exchange cash for electronic value and vice versa, forming the backbone of an emerging electronic currency ecosystem that has potential to connect millions of poor and "unbanked" people to the formal financial system. Unfortunately, low inventory service levels are a major impediment to the further development of these ecosystems. This paper describes models for the agent's inventory problem, unique in that sales of electronic value (cash) correspond to an equivalent increase in inventory of cash (electronic value). This paper presents a "brute force" Markov inventory model and an analytical heuristic that are used to determine optimal stocking levels for cash and electronic value given an agent's historical demand. These models are tested in a variety

of simulated scenarios as well as with a large sample of transaction-level data provided by an East African mobile operator. While both the Markov model and the heuristic outperform historical actual agent decisions by reducing the sum of stockout losses and cost of capital associated with holding inventory, the heuristic matches or outperforms the Markov model both in simulated settings and also in a real, historical setting. The simple and intuitive heuristic increased estimated agent profits by 15% relative to profits realized through agents' actual decisions.

2 Introduction

The rapid growth of cellular networks in the developing world in the past decade has laid the groundwork for a potential paradigm shift in financial services for the poor. Traditionally, the poor and unbanked—the roughly one-third of the world's population who do not have an account at a formal financial institution and live on less than $2 a day—have relied primarily on physical cash when transferring money (Mas 2010). Thus, the velocity of money has been limited by how fast cash can be physically transported, by foot or by bus in most circumstances (Batista and Vicente 2013). This limitation is a critical disadvantage to the poor when money is needed most, such as in the aftermath of a negative economic shock (e.g., sickness or job loss) or a rare opportunity to climb out of poverty through investment (e.g., fertilizer or improved seed purchases) (Helms 2006). At these decisive moments, friends and family willing and able to transfer money have traditionally relied on expensive and/or unreliable cash transfer methods (Morawczynski 2009).

According to the Boston Consulting Group (BCG), approximately 80% of the unbanked population in the developing world had access to mobile phones in 2011 (BCG 2011). Recognizing an opportunity, mobile network operators in the developing world began launching money transfer platforms—known colloquially as "mobile money"—with each platform allowing money to be stored and transferred in the form of digital currency (hereafter referred

to as e-float). In much the same way that text messages can be sent quickly and cheaply, e-float can be instantly transferred across long distances at a near-zero marginal cost. By the end of 2015, mobile money systems hosted roughly 410 million mobile money accounts in the developing world, a 31% year-over-year increase (Groupe Speciale Mobile Association 2015).

By connecting the poor and unbanked in the developing world to the formal financial system, mobile money has provided several benefits. For example, mobile money has been shown to: enable quicker recovery from economic shocks such as job loss or illness to the primary wage-earner (Jack and Suri 2014); enable more efficient receipt of monetary transfers from non-governmental organizations (NGOs) after disasters (Aker et al. 2011); and lay the foundation for access to formal savings, credit, and insurance opportunities for those who currently lack such access (Mas 2010).

2.1 Transaction mechanics and inventory challenges

Cash-in/cash-out (CICO) agents serve as the backbone of mobile money networks, providing a bridge between physical cash and e-float. These agents, often small shop owners, invest in inventories of cash and e-float, and then convert cash to e-float ("cash-in" transactions) and e-float back to cash ("cash-out" transactions) for a commission. The following is a typical use case: An urban laborer in Dar es Salaam, Tanzania gets paid in cash. He conducts a cash-in transaction with an urban agent in which he gives the agent cash and the agent credits the laborer's mobile money account with e-float (Stage 1 in figure A1). For her role in executing the transaction, the agent receives a cash-in commission from the mobile money operator (notably, customers generally do not pay for cash-in transactions). The laborer, now with a balance of mobile money, uses his phone to send this e-float to his family outside of Mwanza, Tanzania in much the same way he might send an SMS message (Stage 2 in figure A1). The operator collects a fee from the laborer for executing this person-to-person (P2P) transfer. Having instantaneously received the e-float on her phone, the

laborer's wife goes to the local agent outside of Mwanza to conduct a cash-out transaction. She gives the agent e-float in exchange for cash (Stage 3 in figure A1). Like the cash-in agent, the cash-out agent is also compensated with a commission from the operator for her role in executing the transaction. Cash-out commissions are generally larger than cash-in commissions, typically by 50% or more. These commissions are most often determined as an increasing step function of transaction value. Commissions for both cash-in and cash-out transactions are generally paid out from the operator to the agent in e-float monthly (rather than immediately after each transaction). In order to conduct a cash-in or cash-out

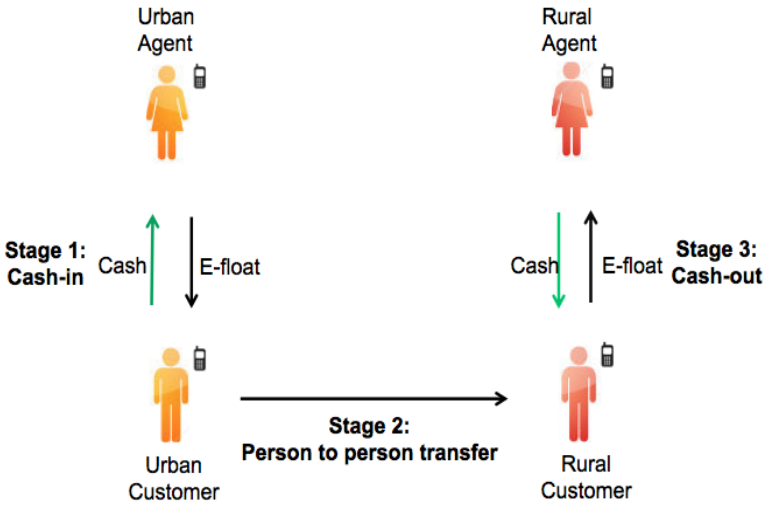


Figure A1: Schematic of a typical urban-to-rural person-to-person transfer, adapted from Agrawal (2009)

transaction, the agent must have sufficient inventory of e-float or cash, respectively. However, stockouts are an acute problem in mobile money networks; customers are often unable to complete CICO transactions because agents run out of cash and/or e-float (Intermedia 2013). Because stockouts of cash and e-float make it harder for customers to easily convert between the two forms of money, they degrade consumer confidence in the convenient convertibility of e-float. Note that because e-float is actual currency, it cannot be "created" on the spot by

either the agent or the operator. Each unit of e-float an operator issues must be backed by traditional deposits at a prudentially regulated financial institution. Though moving e-float once it has been issued is clearly easier than moving cash, agents can, and do, stock out of e-float nonetheless.

In managing CICO transactions, the agent's fundamental challenge is an inventory problem: determining how much cash and e-float to carry in order to most-profitably support their mobile money business. This is a non-trivial challenge. In this setting, the agent not only serves uncertain demand for cash and e-float, but each sale of cash (e-float) also generates equivalent inventory of e-float (cash)—i.e., agents not only face stochastic demand, as is typical in many inventory settings, they also face stochastic replenishment for each good through sales of the other good.

2.2 Preview of results

This paper first presents a "true" model of the agent's inventory evolution and costs. However, this "true" model is intractable, necessitating alternative approaches to solve for the optimal starting values of cash and e-float with which each agent should begin the day. The first approach makes two assumptions about the demand process that allow for solving the problem to optimality with brute force. This first approach is referred to as the "Markov model" and is developed in §4.5 of this paper. The second approach uses an approximation of the agent's underage cost to generate a simple analytical heuristic to solve the problem. This second approach is referred to as the "net demand heuristic," and is developed in §4.6.

The models are then evaluated in §5. In §5.1, simulated settings for evaluation are created using two simplifying assumptions that make the Markov model optimal. In these scenarios, the net demand heuristic does not underperform the Markov model in an economically significant way. Explorations of the reasons for this result are presented, which include analyses relating to the rarity of "double-stockouts," the correlation of minimum and maximum cumulative demand, and the flatness of the cost function around the optimal budget and

the optimal split of the budget between cash and e-float. Next, the models are evaluated in scenarios where the simplifying assumptions are partially relaxed in §5.2. In these scenarios, the heuristic outperforms the Markov model. Finally, the models are evaluated under no simplifying assumptions by employing a large dataset of mobile money agent transactions provided by an East African mobile network operator in §5.3. While both the Markov model and the heuristic significantly increase agent revenue net of cost of capital by reducing the sum of estimated stockout losses and the cost of capital associated with holding inventory, the heuristic performs best in the "real world" evaluation. The heuristic improves estimated aggregate agent net revenue by 15% relative to actual performance.

3 Relation to the Literature

This work relates theoretically to inventory management literature as well as contextually to mobile money literature.

3.1 Inventory management

The mobile money agent's fundamental inventory challenge is informed by decades of work focused on inventory management under demand uncertainty. While some settings discussed in the literature bear resemblance to the problem we study, to the best of our knowledge, there has not been analysis on scenarios where satisfaction of demand for one good generates inventory of another — and vice versa. One such setting relates to "remnant inventory systems." Adelman and Nemhauser (1999) study made-to-order cable manufacturing, where the satisfaction of demand for a cable of certain length generates "remnants" of inventory that can be used to satisfy future orders of shorter cable, but this relationship does not go in both directions. For example, cutting a cable with length 20 from an inventory of cable with length 30 yields "remnant" inventory of length 10, but satisfaction this remnant inventory can never satisfy demand for cable of length 20 again. The critical difference

between remnant inventory systems and mobile money is that satisfaction of demand yields an equivalent amount of inventory of the other good – and vice versa. Another setting that bears resemblance to the mobile money setting is the "stochastic cash balance problem," which has been the subject of significant research by the operations management community beginning in the 1960s (e.g., Girgis 1968, Neave et al. 1970, Chen and Simchi-Levi 2009). The problem is so-named because a bank (or any general firm) has a challenge in managing its inventory of cash: too much cash results in excessive cost of capital, while too little cash incurs some penalty cost, for example the cost of not meeting a banking reserve ratio requirement. The stochastic cash balance problem is different from the standard stochastic inventory problem because demand can be both positive (withdrawals decrease inventory) or negative (deposits increase inventory). Given this fact, the stochastic cash balance problem has been used to study products that can be returned, contributing to the body of research on reverse logistics (e.g., Fleischmann et al. 1997). As will be shown, the mobile money agent's problem shares this feature of demand spanning both negative and positive values. However, while stochastic cash balance problems focus on the single trade-off between holding too much and too little cash, the mobile money problem deals with two separate but linked sets of trade-offs: both too much versus too little cash as well as too much versus too little e-float. Furthermore, while the stochastic cash balance literature largely focuses on developing continuous review policies, such as a two-sided (s, S) policy (e.g., Porteus and Neave 1972, Hausman and Sanchez-bell 1975) that allow for mid-period inventory adjustments, most mobile money agents do not typically have the opportunity to "re-balance" multiple times daily. This allows us to focus on two key values for each agent: the optimal amount of cash and the optimal amount of e-float with which the agent should begin each day.

3.2 Mobile money

The rapid emergence of mobile money has attracted the interest of fields ranging from economics to sociology to public policy. Jack and Suri (2014) study mobile money's social

welfare impacts, finding that households using mobile money were significantly more able to smooth consumption after a negative economic shock (e.g., sickness or job loss) than comparable households not using mobile money. Jack and Suri explain this disparity by demonstrating that mobile money users who experienced shocks received more numerous and larger remittances from farther away than their non-user counterparts. Mobile money had the effect of significantly widening and enhancing informal insurance networks; family and friends were able to send users more money more efficiently in times of crisis. Suri and Jack (2016) also find evidence that access to M-Pesa, the dominant mobile money system in Kenya, has lifted 194,000 households (2% of all Kenyan households) out of poverty since its inception in 2007. Mbiti (2011) demonstrates that mobile money's introduction as a tool for sending money was so disruptive to the markets for remitting money that, in many cases, remittance prices fell by over 50% over a six-year period. Mbiti (2011) also shows that, like the adoption of mobile telephony, early adopters tended to be wealthy, urban, and educated. However, as with mobile telephony, mobile money has progressed down-market and geographically widened its reach very quickly. This is particularly relevant in the developing world where over 80% of adults do not have a bank account (Kendall 2011). Indeed, Mbiti and Weil (2013) note that the number of mobile money agents in Kenya exceeded 25 times the combined total of bank branches and ATMs in the country. Morawczynski (2009) pursued an ethnographic approach to study mobile money's role in empowering women, finding that many of the women interviewed reported that using mobile money to store savings significantly reduced the risk of their husbands appropriating their money, thus increasing their financial autonomy. Balasubramanian and Drake (2015) study how service quality and competition are related to demand, finding that average demand increases with both pricing transparency and agent expertise. That study also finds that agent expertise interacts positively with competitive intensity, suggesting that expertise is a significant dimension of competition between agents. Finally, Aker et al. (2011) study Concern Worldwide's (CW) response to the 2010 drought crisis in Niger. Instead of distributing physical relief items, CW

distributed money. Each month for five months, some beneficiaries received physical cash transfers and others received money transfers via mobile money. Aker et al. (2011) show that the cost of distributing monetary assistance via mobile money, as well as the cost to the beneficiaries of receiving mobile money, was significantly lower than the cost associated with physical cash distribution. As more governments and NGOs shift emergency relief from the distribution of goods to the distribution of money, mobile money's importance post-disaster is expected to increase (BTCA 2014). Accordingly, improving mobile money agents' service reliability will also become increasingly important. To the best of our knowledge, this paper is the first to address this challenge. We do so by applying an operations management lens to this context, focused on improving mobile money agents' inventory management.

4 Inventory Models

Mobile money agents face stochastic demand and interrelated stochastic replenishment for two goods—i.e., sales of cash (cash-out transactions) generate inventory of e-float, and sales of e-float (cash-in transactions) generate inventory of cash. As a consequence, traditional inventory models cannot be applied to this context in a straightforward manner. Solving the agent's "true" model of inventory evolution and costs is intractable, so two alternative approaches are developed to solve for the optimal starting values of cash and e-float with which each agent should begin the day. This first approach, the "Markov model," solves the problem to optimality with brute force by making two assumptions about the demand process. The second approach, the "net demand heuristic," uses an approximation of the agent's underage cost to generate a simple analytical heuristic to solve the problem. This section begins with a description of the parameters used in the modeling process and then develops each of the approaches in turn.

4.1 Description of parameters

At the start of each day, an agent facing daily per-unit cost of capital γ chooses a budget b, and then splits this budget between her daily starting cash quantity q_1 and her starting e-float quantity $f_1 = b - q_1$. The agent then experiences cash and e-float demand arrivals over the course of N time periods of equal duration, indexed by τ. Let D^c_τ represent the magnitude of cash demand at time-period τ. Analogously, the magnitude of e-float demand at period τ is represented as D^e_τ. For all τ, D^c_τ and D^e_τ are non-negative and in reality are neither independent nor identically distributed. The duration of each time period is chosen to be arbitrarily small. Thus, the probability of more than one positive transaction is small enough that we assume no more than one transaction occurs each time period. Most time-periods thus will have D^c_τ and D^e_τ both equal to zero. q_τ and f_τ represent the inventories of cash and e-float at the start of period τ, prior to the realization (if any) of a positive demand D^c_τ and D^e_τ. The quantities of cash and e-float, q_τ and f_τ, fluctuate throughout the day as demand arrives. If the agent is presented with positive cash demand (a customer wants to withdraw cash by surrendering e-float) in period τ, the agent earns a per-unit commission m_c for all cash sales and receives e-float equivalent in value to those sales. The amount of cash inventory decreases (i.e., $q_{\tau+1} = q_\tau - \min(q_\tau, D^c_\tau)$), while the e-float available in the following period increases (i.e., $f_{\tau+1} = f_\tau + \min(q_\tau, D^c_\tau)$). The min represents the possibility of a stockout: the agent cannot give the customer more cash than she has. The agent thus earns a total commission of $m_c \cdot min(q_\tau, D^c_\tau)$. The equations are analogous for e-float demand (a customer deposits cash in order to receive e-float). When an agent experiences a positive D^e_τ, the e-float inventory decreases (i.e., $f_{\tau+1} = f_\tau - \min(f_\tau, D^e_\tau)$, while the cash inventory increases (i.e., $q_{\tau+1} = q_\tau + \min(f_\tau, D^e_\tau)$). The agent's per-unit commission for this transaction is m_e, while the total commission is $m_e \cdot \min(f_\tau, D^e_\tau)$. Note that the number of arrivals (time periods with either positive cash demand or positive e-float demand) over the course of the day is not known ex-ante. Thus, let t be the index of a sub-sequence of all time-periods (indexed by τ) in which either $D^c_\tau > 0$ or $D^e_\tau > 0$. The total number of arrivals

is thus represented by the random variable M, where $M = \sum_{\tau=0}^{N} (\mathbb{1}_{D_\tau^c > 0} + \mathbb{1}_{D_\tau^e > 0})$. Table A.1 lists these parameters and summarizes their respective descriptions. In both of the modeling approaches described in the section, it is assumed that the agent does not re-balance during the day and that any demand not satisfied is lost to the agent.

4.2 The role of demand sequencing

Uncertain sequencing of cash and e-float arrivals complicates decision-making in this setting. To illustrate this challenge, take the following example where initial inventories of cash and e-float are $q_1 = 100$ and $f_1 = 100$, and there is exactly one arrival in period $\tau = 1$ and one arrival in period $\tau = 2$. If the agent experiences a demand sequence $\{D_1^c = 100,\ D_2^e = 200\}$, all of this demand is satisfied. This is because the e-float quantity increases after satisfying demand for units of cash, resulting in sufficient additional units of e-float available for the next time period ($f_2 = f_1 + \min(q_1, D_1^c) = 100 + 100$). Now consider the same arrivals, but in the reverse sequence $\{D_1^e = 200,\ D_2^c = 100\}$: only 200 of the 300 units of total demand are satisfied. This is because the agent is only able to satisfy 100 of the initial 200 units of e-float demand. It is possible that the agent can satisfy all demand for cash and e-float with relatively little inventory (potentially far less than the sum of all demand) if the sequencing of that demand is favorable. Figure A2 presents a simplified depiction of inventory positions of cash and e-float over the course of a day for an agent who begins with 100 units each of both cash and e-float. In this case, the agent stocks out of both cash and e-float over the course of the same day as a result of unfortunate demand sequencing.

4.3 Combining cash and e-float demand distributions

The agent's problem can be simplified significantly by combining cash and e-float demand into a single demand variable.

Lemma 1 *Without loss of generality, a single random variable can completely represent both*

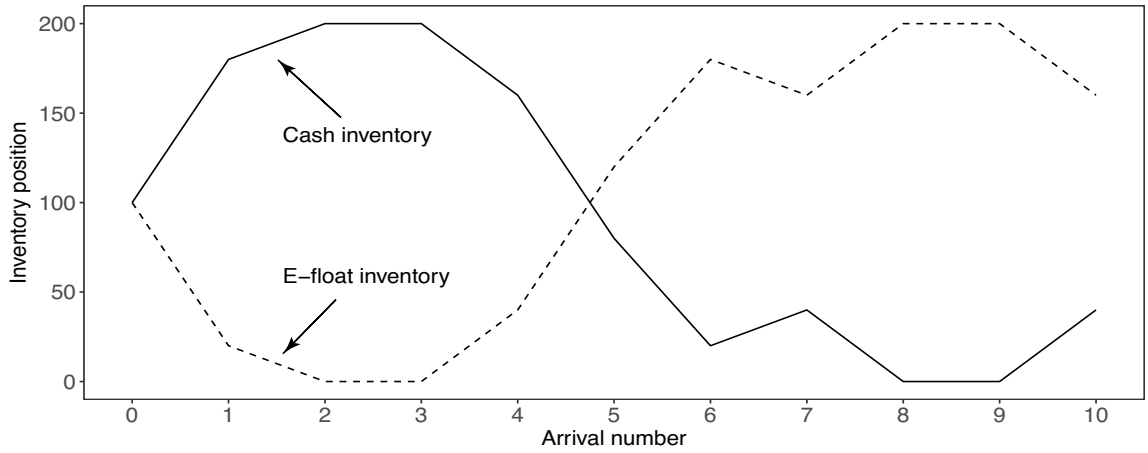


Figure A2: An example of an agent's inventory positions of cash and e-float throughout an illustrative day. The example features both an e-float stockout (between arrivals 2 and 3) and a cash stockout (between arrivals 8 and 9).

cash demand and e-float demand. Specifically, $D_t = D_t^c - D_t^e$. *Recall that* t *is the index of time periods in which there is exactly one of either cash demand or e-float demand.*

All proofs, including that for Lemma 1, are provided in Appendix B. For purposes of analysis and construction of an inventory policy, it is useful to define p_t as denoting the probability that the arrival in period t is a positive cash demand, and $(1 - p_t)$ as the probability that the arrival is a positive e-float demand. Let $p_t \equiv P\left(D_t^c > 0\right)$. Recall that based on the definition of t, in each period t $P(D_t^c > 0 \cup D_t^f > 0) = 1$ and $P(D_t^c > 0 \cap D_t^f > 0) = 0$. We can also define F_{D_t}, the CDF of the combined demand in the lemma, based on the respective *pdf*s for cash and e-float demand: $f_{D_t^c}$ and $f_{D_t^e}$.

$$F_{D_t(x)} = (1 - p_t) \int_{-\infty}^{x} f_{D_t^e}(-x)dx \qquad \text{if } x < 0$$

$$F_{D_t(x)} = (1 - p_t) + p_t \int_{0}^{x} f_{D_t^c}(x)dx \qquad \text{if } x > 0$$

Through this transformation, e-float demand is represented by negative values of D_t and cash demand is represented by positive values of D_t. Now, the evolution of both cash and

e-float inventories, as well as cash and e-float underages, can be characterized with D_t.

4.4 Inventory evolution and cost function

Using the combined demand variable definition, daily agent demand can be represented as a sequence of random variables: $\vec{D} = (D_1, D_2, ..., D_M)$. Recall that M is the number of non-zero demand realizations, and t is the index for only non-zero demands. The agent chooses the budget b and starting cash quantity q_1 such that $0 \leq q_1 \leq b$. The e-float quantity is $f_1 = b - q_1$. Because unfilled demand is lost, neither the cash quantity nor e-float quantity can be negative. Thus, the cash inventory evolution is defined recursively:

$$q_{t+1} = \min\left(b, \max\left(q_t - D_t, 0\right)\right) \ \forall t \in \{1, 2, ..., M\}$$

Again, because the budget is fixed throughout the day, the e-float quantity is always $f_t = b - q_t$. The expected total cost incurred by the agent each day is the sum of expected cash stockout cost, expected e-float stockout cost, and the cost of capital.

$$G(q_1, b) = \sum_{t=1}^{M} \left(\mathbb{E}_{D_t} \left[m_c (D_t - q_t)^+ + m_e (-D_t - b + q_t)^+ \right] \right) + b \cdot \gamma$$

The agent's objective is to minimize $G(\cdot)$ by choosing her optimal budget b^* and the optimal starting cash quantity q_1^*. The optimal starting e-float quantity again follows directly as $b^* - q_1^*$. However, this cost function is deceptively simple, as it obscures significant complexity. First, demand is non-stationary because there can be intra-day seasonality in demand. For example, agents may experience heavy cash-in in the morning and heavy cash-out in the afternoon. Thus, it matters when in the day and in what order those demands occur. Second, the inventory process is not necessarily Markovian. This is the case because there may be a relationship between demand in one period and demand in a later period; for example, customers may do one large cash-in in the morning, or choose to space out the same

total volume of cash-in over the course of the day. Even if the process were assumed to be Markovian, developing solutions for non-stationary models are difficult and often intractable (e.g., Choi et al. 2000, Ghate and Smith 2013). Furthermore, a non-stationary model would require estimates of demand distributions for all time intervals, each of which would need to be small enough such that there is at most one arrival per interval (e.g., seconds) – this is especially important in light of the salience of arrival sequencing outlined in §4.2. This estimation process is both impractical and prone to error, as even stationary models with mis-specified transition matrices can yield poor results (El Ghaoui and Nilim 2005). However, the problem can be solved to optimality by brute force with a tractable Markov model by making simplifying assumptions about the nature of demand. The problem can also be solved approximately without the simplifying assumptions required by the Markov model by making an interesting observation about the sales-inventory connection in this setting. This heuristic approach involves making an approximation of underage cost that vastly simplifies the problem, allowing for the development of a simple analytical heuristic. Both of these approaches are now presented.

4.5 Markov model

Two assumptions can make the model described in section 4.4 tractable. With these two assumptions, the problem can be solved and computed directly by modeling inventory evolution as a Markov chain. To identify the optimal budget and inventory positions, the optimal cash quantity for a single budget is found, and then this process is iterated over the range of possible budgets. The assumptions and their implications are now described.

Assumption 1 *The distributions of demand arrivals are i.i.d. for each given agent-day.*

Assumption 1 allows us to represent demand arrivals for each agent-day as a sequence of realizations of the same random variable, D. This allows for a single transition matrix to represent every inventory transition – after every arrival, the probability of transitioning

from one inventory level of cash to another is the same. This effectively neutralizes the sequencing effect, because previous realizations of D are equally likely to be cash or e-float, and the distributions of magnitudes are identical.

Assumption 2 *The number of arrivals, M, is either constant or a geometrically-distributed random variable that is independent of the value of demand arrivals.*

Assumption 2 allows us to separate the number of arrivals from demand realizations, either by simply assuming constant arrivals or by capturing uncertainty in number of arrivals (M), independently of capturing uncertainty in demand (D). Assuming M is geometrically distributed allows us to account for uncertainty with a single, binary state variable that records whether or not the final arrival has occurred. This indicator variable is denoted as s_t, taking the value of 1 if the final arrival had not occurred prior to the t^{th} arrival and 0 otherwise (i.e., if $s_t = 0$, there would be no more arrivals for the day). The geometric assumption thus allows for capturing the likelihood of any given number of discrete arrivals each agent-day. This geometric distribution is equivalent to conducting a large number of Bernoulli trials, with a "success" being that the final arrival has occurred. Thus we let θ_t be a binary random variable that takes a value of 1 if the final arrival occurs at t and 0 otherwise. Let λ represent the time-invariant probability that any given arrival will be the last, such that $P(\theta_t = 1) = \lambda$ for all t. Using these definitions, we can construct a state transition matrix, P, using the following relationship to calculate each element:

$$\langle q_{t+1}, s_{t+1} \rangle = \langle (1 - s_t) \cdot q_t + s_t \cdot \min\left(b, \max\left(q_t - D, 0\right)\right), s_t \cdot (1 - \theta_t) \rangle \tag{2.1}$$

Letting the maximum number of arrivals be some arbitrarily large integer Z, set the terminal reward to $R_Z(q_Z, s_Z, b) = 0$ and define the stage reward function to be:

$$R_t\left(q_t, s_t, b\right) = s_t \cdot \left(\mathbb{E}_D\left[\left(m_c \cdot (D - q_t)^+ + m_e \cdot (q_t - b - D)^+\right)\right]\right)$$

Backwards induction can be used to determine the optimal starting quantities of cash and e-float (no decisions are made after the initial quantity selections). The cost-to-go function follows:

$$J_t(q_t, s_t, b) = R_t(q_t, s_t, b) + \mathbb{E}_D\left[J_{t+1}\left(q_{t+1}, s_{t+1}, b\right)\right]$$

$$\mathbb{E}_D[J_{t+1}(q_{t+1}, 0, b))] = J_{t+1}(q_t, 0, b) = 0$$

$$\mathbb{E}_D[J_{t+1}(q_{t+1}, 1, b))] = \sum_{q_{t+1}=0}^{b} \Big(P(\langle q_{t+1}, 1\rangle \mid \langle q_t, 1\rangle) \cdot J_{t+1}(q_{t+1}, 1, b)\Big)$$

Thus, given the daily cost of capital parameter, γ, the optimal cost for each given budget, optimal budget, and optimal cash share of the optimal budget, respectively, are:

$$J_1^*(b) = \min_{q_1} J_1(q_1, 1, b)$$

$$b^* = \operatorname*{argmin}_b \ \left(J_1^*\left(b\right) + \gamma \cdot b\right)$$

$$q_1^* = \operatorname*{argmin}_{q_1} \ J_1\left(q_1, 1, b^*\right)$$

The Markov model, however, requires assumptions that are not reflective of reality, and thus any implementation of the Markov model with real data will be sub-optimal to some degree. Additionally, the Markov model is also computationally expensive to implement, a significant practical limitation in the developing world.

4.6 Net demand heuristic

The second approach we develop does not require Assumptions 1 and 2 relied upon by the Markov model. The net demand heuristic is developed from an approximate underage cost (more specifically, a lower bound on the true model's underage cost) that yields useful results without significant computation. The objective remains to find the agent's optimal daily starting inventory of cash and e-float (q_1^* and $b^* - q_1^*$, respectively) as a function of

demand, the daily cost of capital (γ), and cash-out and cash-in commissions (m_c and m_e). Given that the full satisfaction of a cash arrival would yield a corresponding increase in e-float inventory, intuition suggests that cumulative demand (or equivalently, net demand) after t arrivals would be an important quantity in this setting. Define $\Delta_t = \sum_{j=1}^{t} D_j$. Now define two additional quantities based on cumulative demand: maximum cumulative demand ($\hat{\Delta}_t = \max_{1 \le j \le t} \Delta_j$) and minimum cumulative demand ($\check{\Delta}_t = \min_{1 \le j \le t} \Delta_j$). As in the Markov model, let q_1 represent the starting cash quantity and let $b - q_1$ represent starting e-float quantity.

Proposition 1.A *If the initial cash quantity is greater than the maximum cumulative demand ($q_1 \ge \hat{\Delta}_t$) and the initial e-float quantity is greater than the negative of the minimum cumulative demand ($b - q_1 \ge -\check{\Delta}_t$), then all demand up to and including arrival t will be satisfied.*

Proposition 1.B *Given that there are both cash and e-float arrivals (i.e., $\exists$ x and y such that $D_x > 0$ and $D_y < 0$), all demand can be satisfied with strictly less cash and e-float inventory than the sum of all cash and e-float demand.*

As will be demonstrated in this section, Propositions 1.A and 1.B are foundational to the development of the net demand heuristic described here. The implication of Proposition 1.B is particularly important: agents can satisfy all demand while stocking less inventory (sometimes substantially less) than the sum of all demand. The intuition for this statement is as follows: the arrival of cash demand (and satisfaction of that demand) generates e-float inventory which can be used to satisfy future e-float demand and vice versa. As a consequence, satisfying e-float (cash) demand in one period contributes to the ability to satisfy cash (e-float) demand in a future period.

An illustrative example is presented in Table A1, which shows a sequence of cash-in and cash-out arrivals, inventory positions (with initial inventories $q_1 = f_1 = 100$), lost sales, and cumulative demand. In this case, if an agent begins the day with at least the maximum cumulative demand in cash (i.e., starts with at least 120 units of cash) and at least the

negative of the minimum cumulative demand (i.e., 140 units or more of e-float), then the agent would be able to satisfy all cash and e-float demand. Thus the agent in this case could have satisfied all 460 units of demand by holding only 260 units of inventory.

Arrival, t	1	2	3	4	5	6	7	8	9	10
Demand, D_t	80	30	10	-40	-80	-60	20	-60	-40	40
Cash inventory, q^i	100	20	0	0	40	120	180	160	200	200
E-float inventory, f^i	100	180	200	200	160	80	20	40	0	0
Cash underage, $(D_t - q_t)^+$	0	10	10	0	0	0	0	0	0	0
E-float underage, $(-D_t - f_t)^+$	0	0	0	0	0	0	0	20	40	0
Cumulative demand, Δ_t	80	110	**120**	80	0	-60	-40	-100	**-140**	-100

Table A1: An example of agent demand process and inventory evolution throughout an illustrative day.

While the maximum and minimum cumulative demands are clearly not known ex-ante, they can be represented as random variables. These maximum and minimum cumulative demand distributions are used as the basis for a heuristic inventory policy. Though the number of arrivals an agent will experience in any given day is uncertain, the maximum and minimum of cumulative demand can still be represented for an arbitrary M as $\hat{\Delta}_M$ and $\check{\Delta}_M$. A major benefit of this approach is that the maximum and minimum cumulative demand values can be compared "apples-to-apples" across different values of M. Thus, we can drop the "M" subscript, because $\hat{\Delta}$ and $\check{\Delta}$ capture everything needed to calculate approximate underage costs, regardless of M. Next, we state two additional facts that will be useful in developing the heuristic.

Proposition 2.A *The positive part of the difference between maximum cumulative demand and the initial cash quantity* $(\hat{\Delta} - q_1)^+$ *is a lower bound on the cash underage of the true model. The positive part of the difference between the negative of the minimum cumulative demand and initial e-float quantity* $(-\check{\Delta} - b + q_1)^+$ *is a lower bound of e-float underage in the true model.*

Proposition 2.B *This lower bound is sharp (holds at equality) in all cases except those where the agent experiences both e-float and cash stockouts (when $\exists$ both x and y such that $D_x > q_x$ and $-D_y > b - q_y$).*

For the purposes of developing the heuristic, this lower bound on underage is treated as the underage itself. The overage will be captured by incorporating a daily capital cost, γ, which penalizes inventory holding. A unit of cash or e-float costs the agent γ in capital cost (interest or the lost opportunity to deploy capital toward other ends). This capital cost is not salvageable. The cost function, then, can be written as:

$$\tilde{G}(q_1, b) = \mathbb{E}\left[m_c \cdot \left(\hat{\Delta} - q_1\right)^+\right] + \mathbb{E}\left[m_e \cdot \left(-\check{\Delta} - (b - q_1)\right)^+\right] + \gamma \cdot b \tag{2.2}$$

This cost function is convex in b and q_1 (the proof is provided in Appendix B). Therefore, first order conditions can be generated to determine the cost-minimizing q_1 and b as a function of the cost of capital γ, cash commission m_c, e-float commission m_e, the distribution of maximum cumulative demand, $F_{\hat{\Delta}}(\cdot)$, and the distribution of minimum cumulative demand, $F_{\check{\Delta}}(\cdot)$.

Proposition 3 *An agent's optimal starting values of cash and e-float for the net demand heuristic are: $q_1^* = \left(F_{\hat{\Delta}}^{-1}\left(1 - \frac{\gamma}{m_c}\right)\right)^+$ and $b^* - q_1^* = f_1^* = \left(-F_{\check{\Delta}}^{-1}\left(\frac{\gamma}{m_e}\right)\right)^+$ respectively.*

As intuition would suggest, the heuristic recommends holding less cash and e-float as the cost of capital increases, and more cash and e-float when the commissions for cash and e-float sales, respectively, are greater.

5 Performance Evaluation

In this section, we will evaluate the performance of the models under various scenarios. First, we test simulated scenarios where Assumptions 1 and 2 hold (i.e., scenarios when the Markov model solves the problem to optimality). Next, we test simulated scenarios where

Parameter	Model	Description
γ	Both	daily unit cost of capital
m_c	Both	per-unit commission on cash sales
m_e	Both	per-unit commission on e-float sales
τ	Both	time index
t	Both	arrival index
q_t	Both	cash inventory before arrival t
f_t	Both	e-float inventory before arrival t
q_1	Both	beginning of day cash inventory
f_1	Both	beginning of day e-float inventory
D_t	Both	value of demand at arrival number t
s_t	Markov	1 if final arrival has not yet occurred by arrival t
θ_t	Markov	1 if arrival t is final arrival
λ	Markov	probability of any given arrival being the final arrival
Δ_t	Heuristic	cumulative demand after t arrivals
$\check{\Delta}$	Heuristic	minimum cumulative demand
$\hat{\Delta}$	Heuristic	maximum cumulative demand

Table A2: Table of model parameters

the Assumptions are partially relaxed. Finally, we test the models with a full relaxation of Assumptions 1 and 2 (i.e., with real data). To conduct this final evaluation, we utilize a large dataset of East African cash-in and cash-out transactions.

5.1 Performance evaluation under simplifying assumptions

The first set of performance evaluations begin with the construction of a set of scenarios in which Assumptions 1 and 2 hold. Because the Markov model solves this simplified setting's problem to optimality, we are primarily interested in observing the performance of the net demand heuristic under various scenarios with differing sets of empirically-informed parameters. These scenarios vary four parameters which affect the nature of demand; specifically, we vary the number of arrivals (M), cash volume to total transaction volume ratio ($p = \frac{M_c}{M}$), mean demand (μ), and coefficient of variation $CV = \frac{\sigma}{\mu}$. The set of 81 scenarios is constructed as the Cartesian product of the sets containing low, medium, and high values of M, μ, and CV, as well as completely balanced, moderately "imbalanced", and very "imbalanced" levels of p, where imbalance is measured as $|0.5 - p|$. Specifically, $M \in \{6, 12, 24\}$,

$p \in \{.50, .67, .83\}$, $\mu \in \{13000, 24000, 47000\}$, and $CV \in \{1.05, 1.34, 1.75\}$. Values for μ and CV were chosen as the 20^{th}, 50^{th}, and 80^{th} percentiles of our historical dataset (which is described later in this paper). Levels of M were selected to be as close to the 20^{th}, 50^{th}, and 80^{th} percentiles as possible while also ensuring that cash and e-float transaction counts were integers to avoid adding unnecessary noise to this analysis.[1] Finally, values for p were chosen such that the levels of "imbalance" were also chosen as close as possible to the empirical 20^{th}, 50^{th}, and 80^{th} percentiles.[2] Because the distributions of cash and e-float demand magnitudes are discrete and non-negative, we construct the combined i.i.d. demand magnitude from negative binomial distributions. Specifically, we start with the same negative binomial distribution for both cash and e-float magnitude, scale each by p and $1-p$ respectively, and then reflect the scaled e-float magnitude distribution about the y-axis and append it to the scaled cash magnitude distribution. In other words, we begin with a negative binomial distribution $D \sim NegBin(\mu, \frac{\mu}{CV^2 \cdot \mu - 1})$ and then scale and reflect this base distribution to generate the combined distribution, such that $f_D = -(1-p) \cdot f_D(-x) + p \cdot f_D(x)$. We make a choice such that μ and CV are the same for both the base distributions of cash and e-float magnitude. The choice to use the same parameters to generate base negative binomial distributions across e-float and cash (prior to scaling and reflecting) for each scenario keeps the analysis insightful without unnecessarily complicating the simplest set of scenarios. The next two subsections progressively incorporate more complexity by relaxing assumptions on the demand distributions.

For each of the 81 scenarios, we generate 10,000 simulated days for training the heuristic, and 10,000 simulated days for evaluation. The maximum and minimum cumulative demand was generated for each of the 10,000 training days, and these data points formed the empirical distributions of $\hat{\Delta}$ and $\check{\Delta}$ respectively. The respective net demand heuristic fractiles were applied to this empirical distribution to generatre net demand heuristic recommendations for

[1] The exact corresponding empirical values were 6, 11, and 19.

[2] The exact corresponding values were 0.06, 0.18, and 0.33 respectively. Given that these values measure imbalance, we could choose representative p as above .5, below 0.5, or some combination of the two. For simplicity, we choose values of p that are above 0.5, but this choice does not impact the results.

the starting values of cash and e-float. The fractiles were calculated from the empirically-informed parameters $\gamma = .0005$ (annualizing to 20%), $m_e = .0066$, and $m_c = .0105$. An analysis of the effect of altering γ will be described later in this section. The probability transition matrix and stage cost vector for the Markov model were calculated directly from the combined distribution and used to derive the Markov model recommendation. Next, the recommendations from both the net demand heuristic and the Markov model for each of the scenarios were evaluated against 10,000 demand vectors each representing a simulated day. Both stockout costs and capital costs could be computed for both models, and those results are now presented. For all 81 scenarios, we conducted a one-tailed paired t-test to compare the Markov model's performance and the heuristic's performance. The null hypotheses for these tests were that the Markov model's net revenue is not greater than or equal to the heuristic's performance, while the alternative hypothesis was that the net demand heuristic performed worse than the Markov model. Of the 81 scenarios run, the null hypothesis could be rejected in only seven of them, at the $p = 0.05$ level (only eight scenarios at the $p = 0.10$ level, and only two scenarios at the $p = .01$ level). There was not a meaningful pattern in those scenarios where the net demand heuristic underperformed in a statistically significant way, which indicates that these results likely occurred by chance. Furthermore, even among the seven scenarios where the net demand heuristic underperformance was statistically significant, the difference was not economically significant: the average percentage by which the heuristic underperformed was 0.003%. Across all 810,000 simulations, the heuristic captured 99.9998% of the optimal net revenue.

Recall that from Proposition 2.B that the underage cost approximations that are used to formulate the net demand heuristic recommendations are the exact underage cost in all circumstances except for when there are stockouts of both cash and e-float. We shall term these scenarios as "double-stockouts." The frequency and effect of these "double-stockouts," therefore, is salient. First, note that when parameters m_c, m_e, and γ take the empirical values described above, the optimal fractiles imply high service levels: $\frac{1-\gamma}{m_c} = 0.952$ for cash

and $1 - \frac{\gamma}{m_e} = 0.924$ for e-float. If one were to assume that cash and e-float stockouts after starting with net demand heuristic recommendations were independent, then the fraction of simulated days that experienced both stockouts was $(1 - 0.952) * (1 - 0.924) = .0036$. In other words, we would expect only 36 of the 10,000 simulations to experience double-stockouts that threaten to degrade net demand heuristic performance. Increased levels of γ, however, would imply higher cash stockout rates, higher e-float stockout rates, and thus higher double-stockout rates. Interestingly, the level of double-stockouts observed per 10,000 simulations is even lower, instead equalling approximately only six. This is due to the fact that cash stockouts and e-float stockouts are negatively correlated. The magnitude of this negative correlation changes as the cost of capital changes. This dynamic is illustrated in figure A3 which shows the correlation of e-float stockouts and cash stockouts across 10,000 simulated days generated with the median set of parameters ($M = 12$, $p = 0.67$, $\mu = 24,000$ and $CV = 1.34$) as the cost of capital is varied from 5% annualized to 800% annualized.[3] When the cost of capital is very low, stockouts are extremely rare, and thus the magnitude of negative correlation of cash stockouts and e-float stockouts is also low. In very high cost of capital scenarios, stockouts of cash and e-float are both very common, which also yields a small negative correlation. However, when cost of capital values are in between very low and very high, the magnitude of the negative correlation is elevated. Informal interviews with agents yielded a range of plausible cost of capital values. This range, roughly 5% to 95% annualized, is denoted by shading in figures A3, A4, and A5. This negative correlation manifests itself in double-stockout rates that are lower than what would be expected if cash and e-float stockouts were independent. This is illustrated in figure A4, where the dotted line depicts the double-stockout rate implied by independent cash and e-float stockout rates as γ increases, and the solid line depicts the actual double-stockout rate which is lower due to the negative correlation between cash and e-float stockout rates. As implied by the insight from figure A3, we see convergence of the lines at both very low and very high cost of capital

[3]The results presented here are not meaningfully different from the same analysis performed with other sets of plausible parameters.

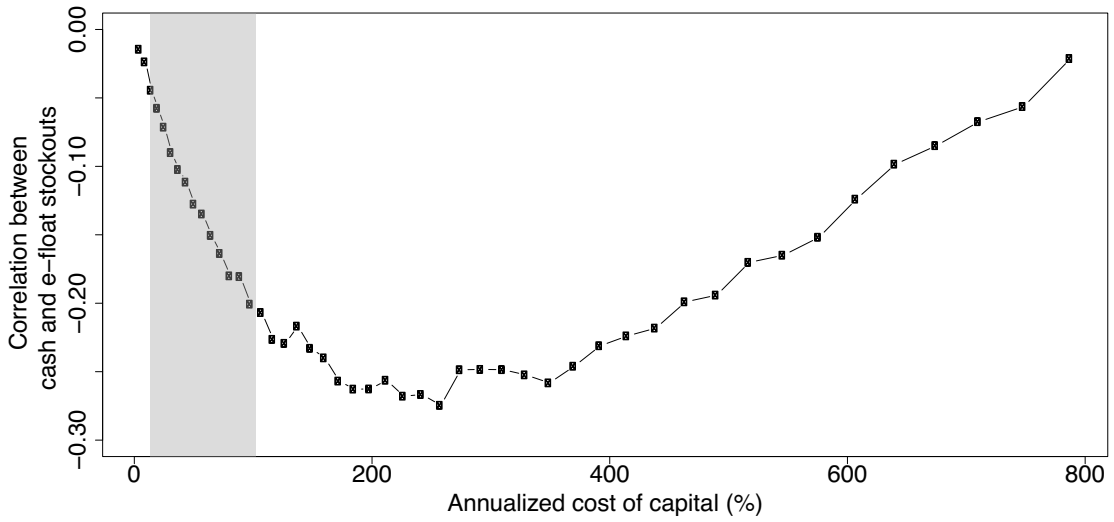


Figure A3: Correlation of cash stockouts and e-float stockouts as cost of capital increases. Note the non-negligible negative correlation in the emprically plausible range (shaded region).

rates.

Figures A5 and A6 illustrate how the heuristic performance degrades as the cost of capital increases and percentage of days with double-stockouts increase, respectively. Non-negligible degradation of heuristic performance appears only after annualized cost of capital exceeds 200%, well past the region of empirical plausibility gauged from interviews with agents.

The region in figure A5 between 100% and 200% annualized cost of capital is interesting; while the percentage of days with a double-stockout grows to be non-negligible, the heuristic is still capturing almost all of the optimal net revenue. This suggests that using mildly sub-optimal recommendations for starting inventories of cash and e-float do not yield meaningful degradation in heuristic performance. In other words, it seems that the cost function is roughly flat near the optimal. Through brute force simulation for each of the 81 scenarios, the cost associated with the entire range of plausible values for budget (where budget is the sum of cash and e-float values) can be found. These values are plotted against the percent by which each budget value deviates from the optimal budget. Each of these curves is then

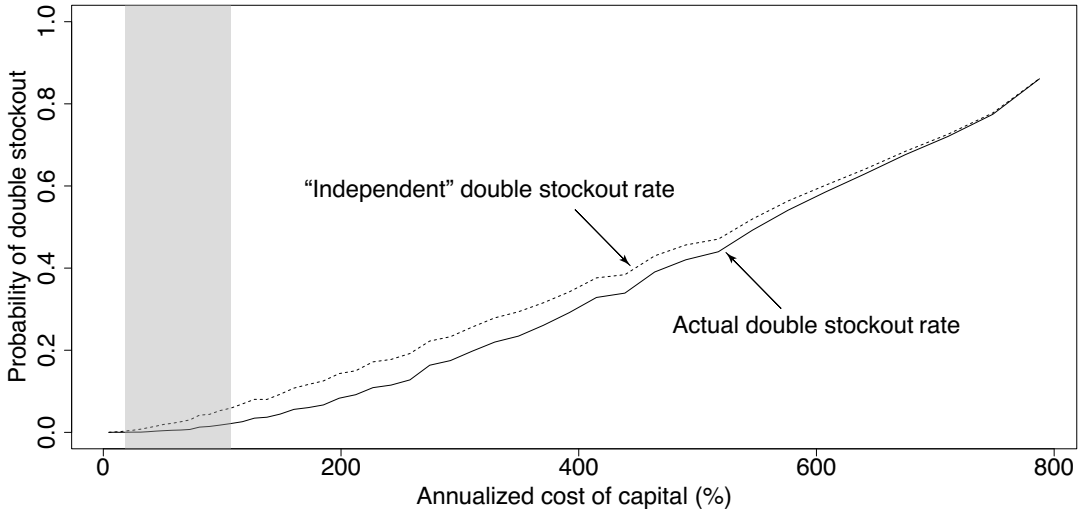


Figure A4: Comparison of actual double-stockout rate with the double-stockout rate if probability of cash stockouts and e-float stockouts were independent. Double-stockout rates are very low in the empirically plausible region (shaded).

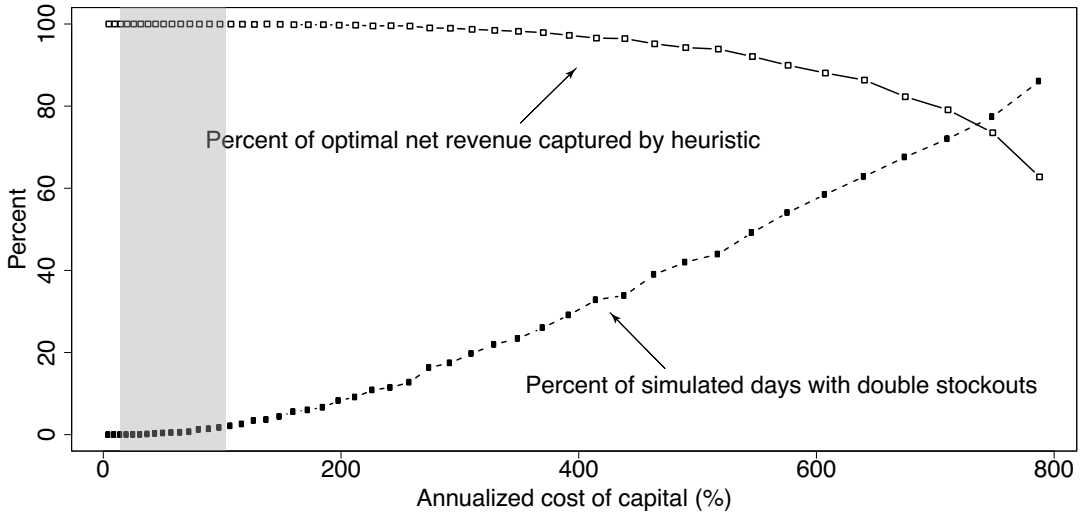


Figure A5: Heuristic performance degradation and percent of simulated days with double-stockouts as cost of capital increases. Note very little performance degradation in empirically plausible range (shaded).

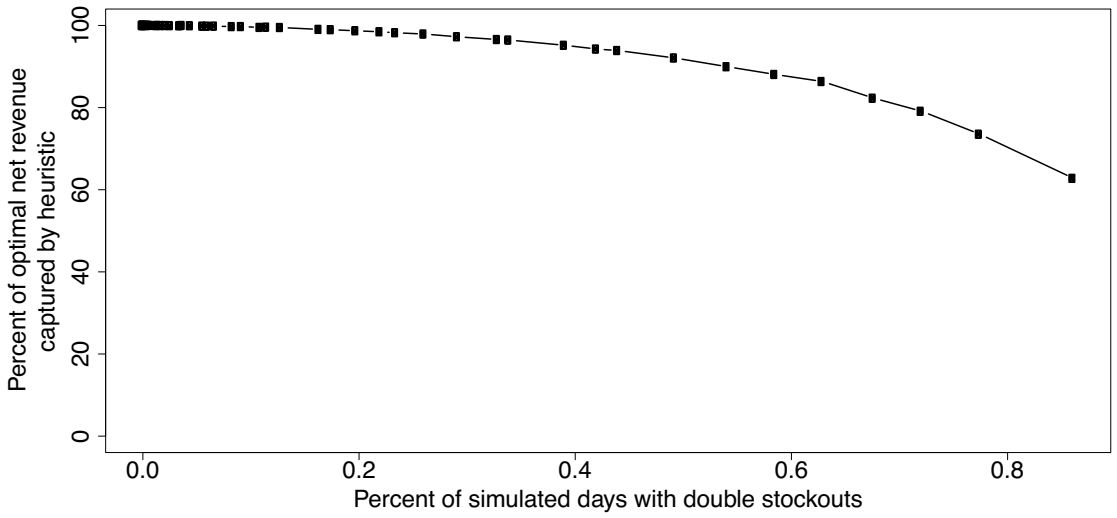


Figure A6: Heuristic performance degradation as the percent of simulated days with double-stockouts increases. Note the slow performance decay.

fit as a quadratic polynomial, and the coefficient on the quadratic term of this fit is recorded. This coefficient is a measure of cost curve flatness – the larger this coefficient, the greater the effect deviations from the optimal budget have on cost. Figure A7 depicts the curves that correspond to the sets of parameters with the smallest, median, and largest quadratic coefficients respectively.

An analogous analysis was conducted with respect to the share of the budget allocated to cash. Using the optimal budget, the cash share of that budget was varied and the resulting net revenue estimates were captured. Again, the deviation from the optimal cash share was plotted against the deviation from optimal cost. Figure A8 depicts the scenarios with the smallest, median, and largest quadratic coefficients, respectively. In general, the cost curves (both with respect to budget and with respect to cash share of budget) are relatively flat near the optimal. Even in the worst-case scenario, deviating by 4% in either direction of the optimal budget yields less than 0.5% higher cost. Analogously, deviating 4% in either direction of the optimal cash share of budget yields at most a 0.5% increase in cost. In

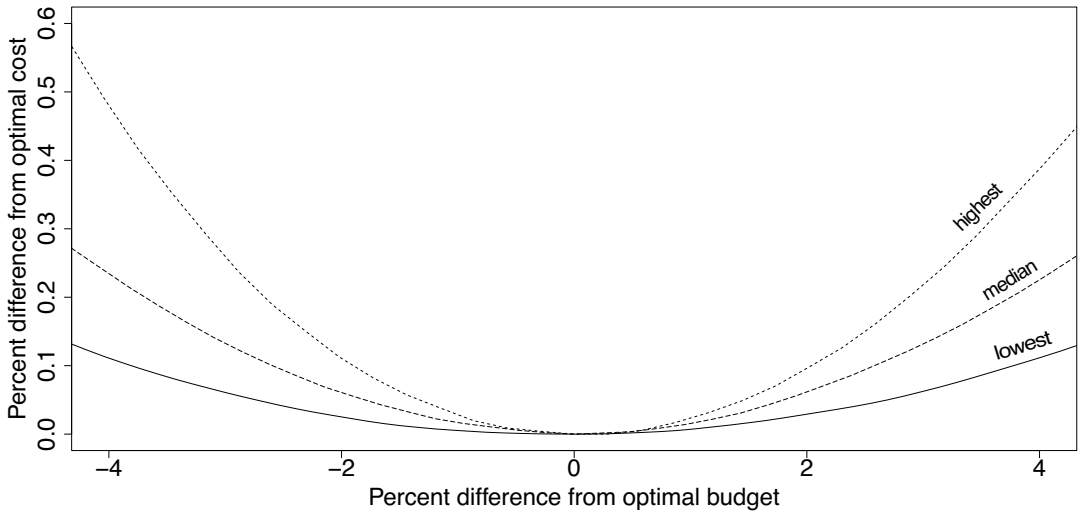


Figure A7: Percent deviation from optimal cost as a function of percent deviation from optimal budget. The lines correspond to the scenarios with progressively flatter cost curves, representing the highest, median, and lowest quadratic coefficient. Even the worst-case scenario has a relatively flat cost curve.

both the budget and cash-share analyses, the median scenario is substantially flatter at the optimal.

5.2 Performance evaluation under partially relaxed simplifying assumptions

Next, we relax Assumption 1 (which assumes i.i.d. demand) slightly, such that the Markov model is no longer optimal. In these cases, there will now be differences between the morning and afternoon demand distributions. To conduct this analysis, we use the same principles as the previous analysis for constructing scenarios, where M, p, μ, and CV are varied. The base negative binomial distributions are constructed in the same way as the previous analysis (where both Assumptions hold), except for that in this analysis, p is applied to the morning distribution and $1 - p$ is applied to the afternoon distribution, where the morning and afternoon have an equal number of arrivals. For example, if $M = 12$ and $p = .67$, there

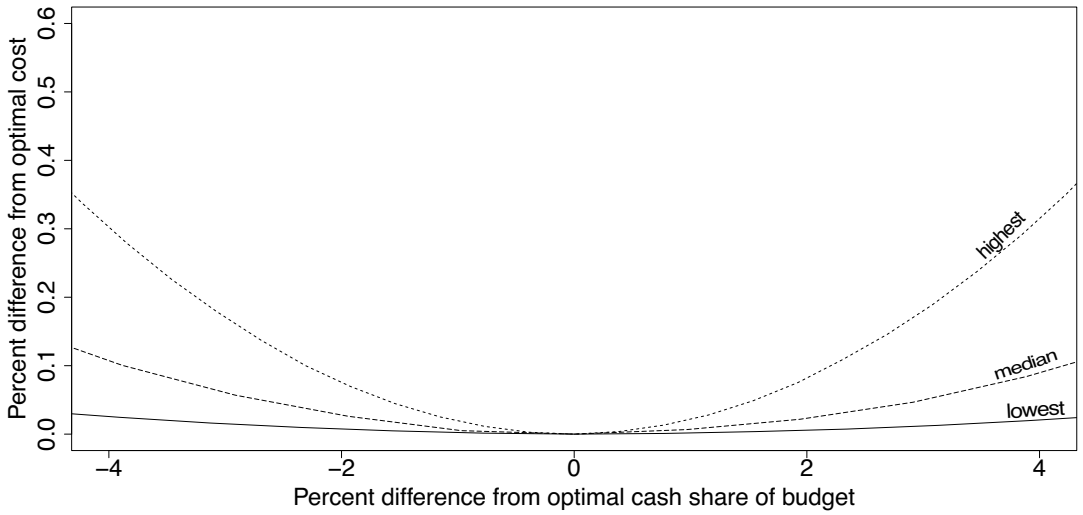


Figure A8: Percent deviation from optimal cost as a function of percent deviation from optimal cash share of budget. Lines correspond to scenarios with progressively flatter cost curves, represnting the highest, median, and lowest quadratic coefficient. Even the worse-case scenario has relatively a flat cost curve.

would be six arrivals each in the morning and afternoon; in the morning, there would be four cash arrivals and two e-float arrivals (in randomized order), whereas in the afternoon, there would be two cash arrivals and four e-float arrivals (also in randomized order). We use the same set of possibilities for μ and CV, but eliminate one value from the sets of p and M respectively. Specifically, we eliminate the value $p = 0.5$ because this represents a completely balanced scenario with no difference between morning and afternoon, which was previously analyzed. Second, we eliminate the value $M = 6$ because applying the remaining values of $p \in \{0.67, 0.83\}$ to $\frac{M}{2}$ result in a non-integer number of arrivals for the morning and afternoon. The inclusion of $M = 6$ would thus introduce unnecessary noise into the analysis. The Cartesian product of these sets yields 36 scenarios that can be analyzed. With each set of parameters, we again generate 10,000 simulated training days and 10,000 evaluation days as was done previously. In this scenario, because the Markov model assumes i.i.d. demand throughout the day, only one demand distribution can be used. While the

net demand heuristic does capture this intra-day seasonality (the maximum and minimum cumulative demand capture the relevant sequencing information), the Markov model is not able to. Intuition thus suggests that the net demand heuristic may over-perform the Markov model.

This intuition is borne out in the results of the simulations. In all 36 scenarios (averaged over 10,000 simulations each), the net demand heuristic outperformed the Markov model recommendations. The mean and median over-performance of the heuristic versus the Markov model recommendations was 0.52% and 0.55% respectively. The null hypothesis that the Markov model's net revenue is greater than or equal to the net demand heuristic's net revenue is rejected in one-tailed, paired t-tests of all 36 scenarios at the $p = 0.01$ level. While the effect of differences in M, μ, and CV are not readily apparent in examining their relationship to the amount by which the heuristic outperforms the Markov model, it is clear that higher levels of imbalance ($|p - .5|$) yield a larger over-performance by the heuristic. The over-performance in the more "balanced" case of $p = .67$ was 0.4%, while the more imbalanced case of $p = 0.83$ was 0.7%. This result is intuitive because while the Markov model is unable to distinguish the difference between morning and afternoon, the effect of this inability is most acute in the highly imbalanced scenarios. It is therefore clear that even with a small relaxation of the ideal conditions, the net demand heuristic can outperform the Markov model.

5.3 Performance evaluation with historical data

Finally, we can relax both Assumptions 1 and 2 using real data. For this analysis, using historical data as input to each of the models, recommendations of starting values of cash and e-float inventories per agent per day can be generated, and the models' performance can be compared to agents' actual performance. This section conducts these analyses, beginning with a description of historical data followed by a description of the results from the evaluation of the Markov model and net demand heuristic against actual decisions for all

agents over the final six months of the sample period. To evaluate agents' actual decisions against recommendations, a large sample of transaction-level cash-in and cash-out historical data from a scaled mobile money operation in East Africa was utilized. Real Impact Analytics, a company specializing in business analytics for mobile networks in the developing world, provided anonymized transaction logs, each of which features a time-stamp, transaction type, transaction value, anonymized IDs of the sender and receiver, as well as the pre and post e-float balances of both the sender and receiver. These transaction logs contained all 35,882,460 cash-in and cash-out transactions from 6,725 agents who conducted at least one cash-in and one cash-out transaction on at least 336 of the 471 days (roughly a five day work-week over the period) between June 1, 2014 and September 14, 2015. The total number of agent-days in the sample was 2,708,385. With this data, each agent's actual daily e-float decision as well as all e-float (cash-in) and cash (cash-out) sales can be directly identified. Because neither cash inventory nor stockouts (since the sales data represents censored demand) are directly observable in the data, these quantities are estimated. Both the methodology for augmenting historical sales data with estimated stockouts to create estimated demand and the methodology for estimating cash inventories are described in Appendix B.

The Markov model requires estimates of both the geometric success parameter λ (to account for uncertainty in the number of arrivals) as well as a transition matrix to describe the probabilities of transitioning between states. The process for estimating these inputs is now described. First, λ is calculated from an estimate of the number of arrivals for each agent-day. This estimate of the number of arrivals is generated with a one-step seasonal point forecast using the number of arrivals seen by that agent on all previous days. The seasonal forecasting method accounts for day-of-week effects on arrivals. Second, the transition matrix (whose entries are specified in equation 2.1 is generated from an empirical demand distribution of the magnitudes of each agent's historical sales. The net demand heuristic requires only estimates of the distributions of maximum and minimum cumulative demand. One-step seasonal forecasts of minimum and maximum cumulative demand generate means

and standard deviations of normal distributions, which are used as the forecasted maximum and minimum cumulative demand distributions, respectively. The net demand heuristic fractiles are applied to these respective distributions to generate recommendations for each agent's starting cash and e-float inventories.

We evaluate model and actual performance over the final 180 days of uncensored transaction data, generating recommendations from the Markov model and net demand heuristic for each agent-day within this horizon. Each model generates recommended starting cash and e-float inventories for each agent-day. For each set of values (Markov model and net demand heuristic recommendations, as well as the actual e-float decision and estimated actual cash decision) we simulate the day using sequenced demand from the augmented transaction log. The commission lost from unmet demand is summed to generate a stockout loss estimate for each model for each agent-day. The cost of capital associated with e-float inventory holding each agent-day is also calculated. Only the stockout losses from cash-in (e-float demand) transactions and cost of capital associated with e-float are presented here, as the agent's actual cash decisions are not as precisely calculable as e-float decisions. These results are aggregated across all agents over the 180-day period. A performance comparison of agents' estimated actual decisions, net demand heuristic, Markov model, and hindsight decisions are presented in figure A9. The hindsight results are calculated by determining the cost of capital associated with stocking the minimum inventory required to satisfy all sequenced demand. Thus the hindsight decisions result in no stockout losses, but some cost of capital. Each bar in figure A9 represents the aggregate amount of commission that could have been earned had all agents satisfied all e-float demand over the evaluation period. The black portion represents e-float commission lost due to e-float stockouts, the gray represents the cost of capital allocated to e-float, and the white represents the net revenue realized by all agents in each model. As can be seen, there are striking differences in performance across the actual, heuristic, Markov model, and hindsight decisions. Paired t-tests reveal statistically significant differences in total e-float inventory costs (stockout + cost of capital) between the

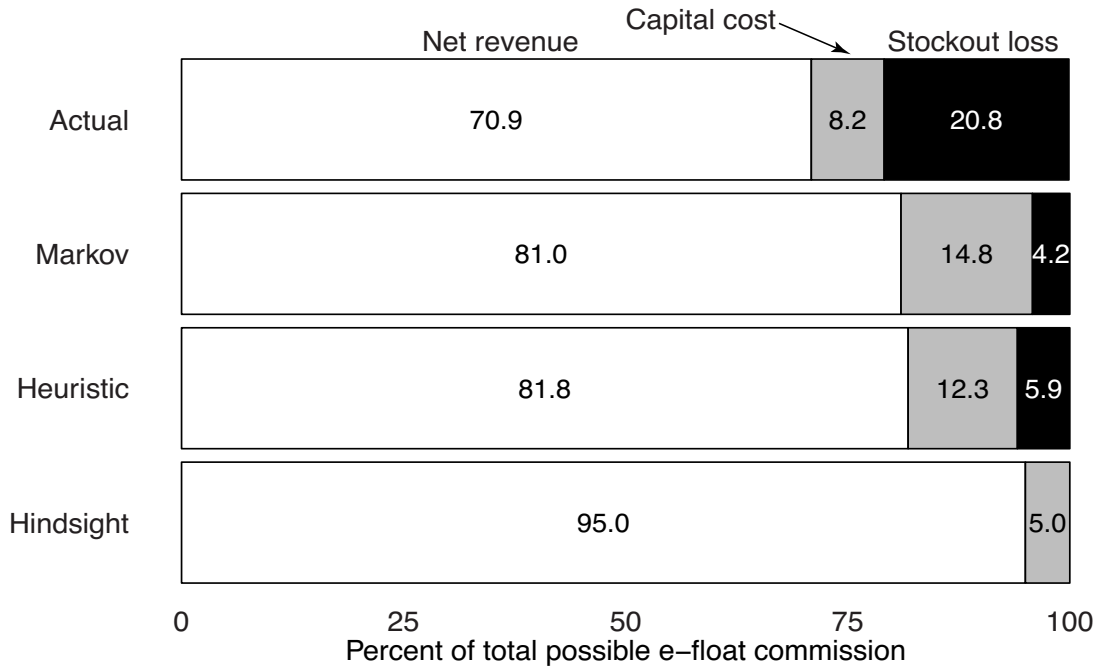


Figure A9: Performance evaluation: aggregate net e-float revenue for actual agent decisions, Markov model, net demand heuristic, and hindsight recommendations over 180 days. The heuristic outperforms both the Markov model and actual agent decisions.

net demand heuristic and actual decisions, as well as between the Markov model and actual decisions. Both of these tests result in values of $p < .001$. The net demand heuristic results in e-float stockout losses as a percent of total possible e-float commission that are nearly 15 percentage points (20.8% to 5.9%) less than those that correspond with the actual decisions made by agents, while only requiring 4.1 percentage points of extra capital cost (8.2% to 12.3%). The resulting 10.9 percentage point increase of the net demand heuristic net revenue over actual net revenue is a key result of this paper. While the net revenue in the hindsight optimal scenario is 13.2 percentage points, greater than the net demand heuristic, the exact sequencing and magnitudes of demand arrivals must be known ex-ante to achieve this level of improvement. Some of this performance gap between the net demand heuristic and hindsight optimal decisions can likely be closed through more accurate forecasts for the minimum and maximum cumulative demand distributions (upon which the net demand fractiles are applied). More sophisticated forecasting to improve net demand heuristic performance is

left for future work. Results presented in figure A9 assume an annualized cost of capital (γ) of 20%; an analysis of how the heuristic performs relative to actual decisions as a function of cost of capital is presented in figure A11. As expected, the net demand heuristic slightly outperforms the Markov model. This is likely due to the fact that the net demand heuristic captures intra-day seasonality, whereas the Markov model cannot. Also, due to the Markov model's requirement of two sets of separate inputs (estimates for both the number of arrivals and the distribution of demand magnitude per arrival); the compounding of estimation errors may be hampering performance. The net demand heuristic, on the other hand, requires only a single set of inputs (the distributions of the maximum and minimum cumulative demand). It is also noteworthy that the heuristic's stockout losses are higher than the Markov model (5.9% to 4.2%), while the capital costs are lower (12.3% to 14.8%). This observation is interesting in light of the fact that the net demand heuristic is built upon the lower bound on underage cost – which leads to lower inventory recommendations than the Markov model (which, under Assumptions 1 and 2, is equivalent to the true model). This lower inventory recommendation results in both higher stockout losses and lower capital costs. The results presented are robust to changes in the 180-day evaluation horizon: comparisons using an evaluation horizon of the final day, week, month, and three months of transaction data do not yield materially different results.

At an agent level (as opposed to aggregate savings presented in figure A9), most agents would have experienced significant benefit from the net demand heuristic. As seen in figure A10, while a small subset (roughly 10%) would be worse off (none by more than 20 percentage points), the vast majority of agents could increase net revenue by using the heuristic's recommendations. The mean agent increase is 9.9 percentage points, and the median agent increase is 8.2 percentage points. This discrepancy arises because some agents would benefit substantially more, with the largest increase in net revenue for an individual agent nearly 50 percentage points.

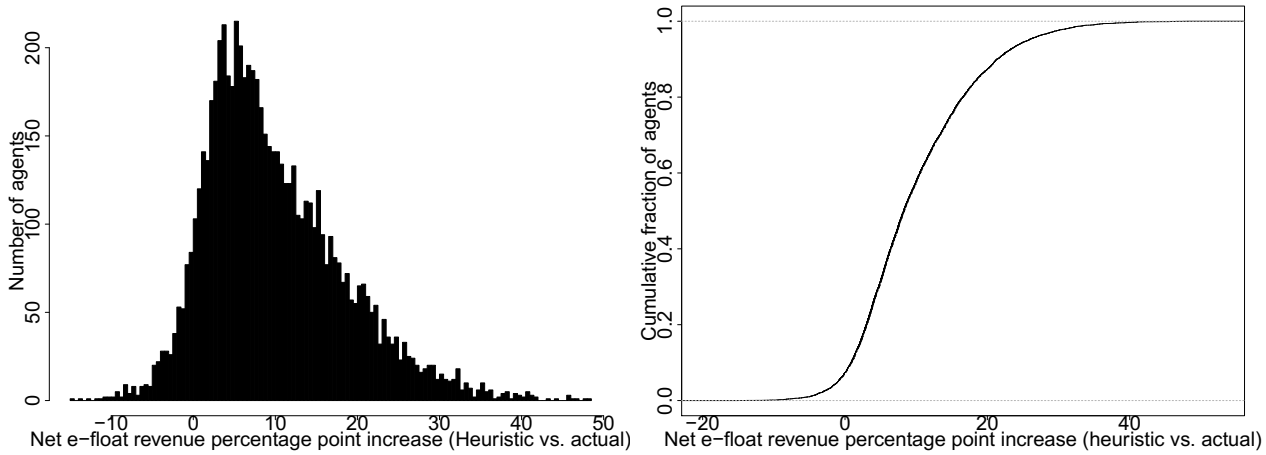


Figure A10: Histogram and empirical cdf plots of heuristic's net e-float revenue percentage point improvement over each agent's decisions. Most agents would benefit from heuristic recommendations.

Based on informal interviews with mobile money agents, the daily cost of capital parameter γ was estimated and assigned a value of 0.05% (20% annualized). Agents were asked about the terms of credit they had received (if they borrowed money to finance their inventory) or the terms of credit they would be willing to extend (if they had financed inventory without borrowing). While there was a range of responses between 5% and 80% annualized, 20% seemed to be the most representative single number. However, given that the cost of capital parameter has a significant effect on the net demand fractiles, and thus inventory recommendations, it is important to calculate the net demand heuristic's performance over a range of plausible cost of capital values. Figure A11 illustrates the number of percentage points by which the actual inventory cost (lost sales and cost of capital) exceeds the inventory cost under the net demand heuristic recommendation for various levels of cost of capital. For the range of annualized cost of capital that is likely to be applicable to the vast majority of agents (from 2% to 100%), observe that the net demand heuristic can increase aggregate net revenue significantly, ranging from 8 to 18 percentage points of total possible revenue. Note that while the stockout loss reduction decreases as the cost of capital increases, above an annualized cost of capital of 68% the net demand heuristic recommends holding less inventory

than actual agent decisions in aggregate.

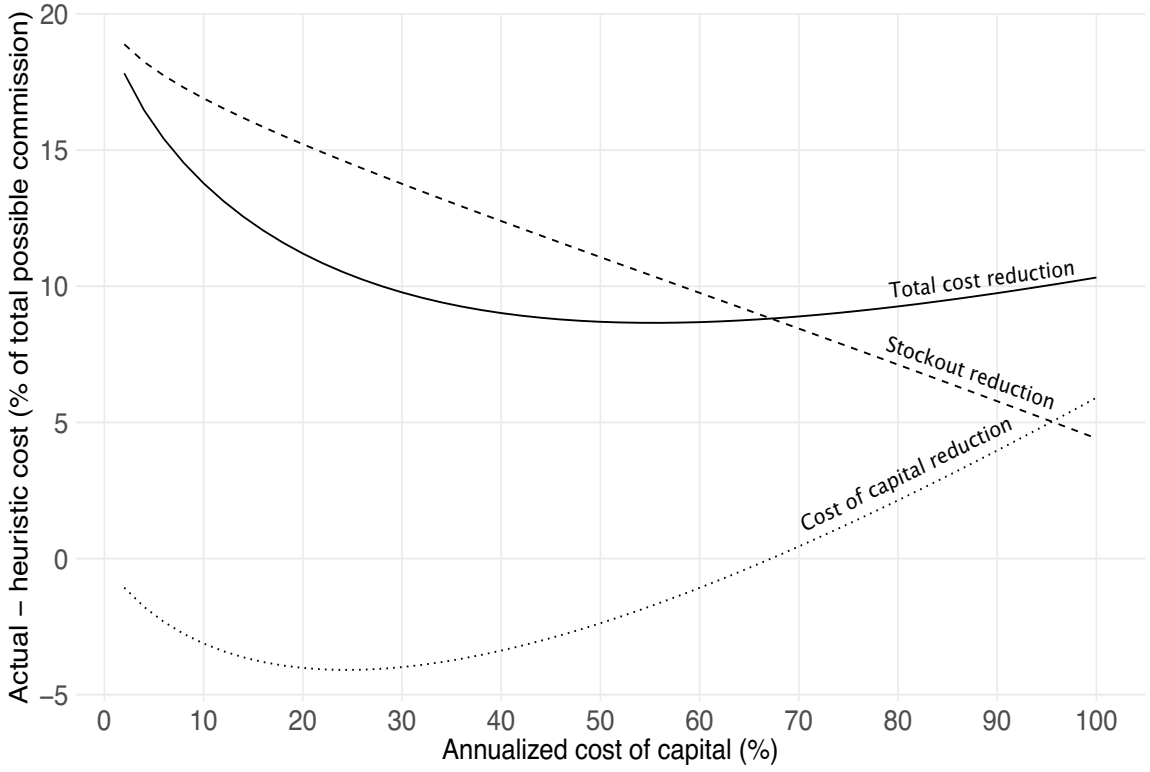


Figure A11: Cost savings sensitivity as a function of cost of capital. Heuristic performance improvement over actual agent decisions are economically significant over a range of cost of capital values.

6 Discussion and Conclusion

In this section, the implications of these findings, limitations of this work, and promising areas for future mobile money research with an operations management lens are discussed. Both the Markov model and the net demand heuristic can produce per-agent per-day recommendations for cash and e-float stocking decisions that can increase agent profitability. While it is likely that more sophisticated estimation processes could enhance the performance of the Markov model, the net demand heuristic performs the best "out-of-the-box" and is thus well-positioned to be operationalized in the field given that it requires limited computational resources and comparatively minimal estimation. The net demand heuristic

also has the added benefit of being intuitive: the concepts of minimum and maximum cumulative demand, in addition to balancing overage and underage costs, can be taught to agents. This has the potential to contribute to a virtuous cycle: less stockouts (without excess capital) lead to more profitable agents as well as happier customers. However, there are limitations to this work. First, the estimated benefits presented here assume that agents follow the heuristic's budget and inventory recommendations. Behavioral biases that might influence how an agent interprets and acts on these recommendations are not addressed. Implementing the net demand heuristic in the field to determine how well agents adhere to the recommendations, as well as determining how much agents benefit in practice, is thus an important area for future work. Second, the lack of data related to cash inventory balance hampers our ability to precisely determine agent cash stocking decisions and cash stockouts. While the method employed here to calculate cash balances (and cash stockouts) is likely too conservative (likely overestimating agent cash balances and underestimating agent cash stockouts), it is likely that agents would benefit from following net demand heuristic cash holding recommendations in addition to e-float holding recommendations. Given that e-float stockout losses are significant, it is likely that cash stockout losses are also significant. In fact, cash stockout losses are likely to be more severe than e-float stockout losses because cash commissions are larger than e-float commissions, by approximately 50% for the average agent. Third, it is assumed that agents do not rebalance during working hours daily. This assumption is supported by a large survey of mobile money agents in Uganda, Tanzania, and Kenya. The survey, conducted by the Helix Institute of Digital Finance, finds that the median number of rebalances (proactively converting e-float to cash or vice versa) per month in East Africa is less than eight (McCaffrey et al. 2014, Githachuri et al. 2014). This finding – combined with the fact that in order to rebalance during working hours agents must either close their shop or be short-staffed – suggests that most agents, in general, do not rebalance multiple times daily. The other rebalancing assumption is that agents are able to costlessly rebalance each day (either before the first transaction and/or after the last trans-

action of each day). This assumption is also supported by the agent survey; the survey finds that most agents can rebalance easily and inexpensively (excluding working hour rebalances which would incur potential lost sales): 72% of agents in Uganda were within 15 minutes of a rebalancing point, and the transit cost to a rebalancing point was nominal for most agents. Conditions for rebalancing in Kenya and Tanzania were found to be similar or even more favorable (McCaffrey et al. 2014, Githachuri et al. 2014). Some mobile money markets are competitive in that there are multiple competing mobile money platforms. In some of these markets, the operators do not have the market power required to demand agent exclusivity; in these markets, most agents provide CICO services for multiple platforms. In this case, agents must make stocking decisions for each operator's platform (no scaled mobile money market has developed systems that allow simple and free exchange across platforms). However, this scenario is more complicated because while e-float is not interchangeable between platforms (i.e., e-float of platform A cannot satisfy a cash-in arrival for platform B), cash is fungible (i.e., the same pool of cash can be used to satisfy cash-out transactions on both platform A and B). Thus, the multi-platform agent's problem is another potential area for future research. Finally, the mobile money agent's inventory problem is a simplification of the inventory problem of a generic N-currency exchange agent. For example, an airport currency exchange agent might need to stock up to 15 different currencies, selling one currency that increases the inventory of another currency. In addition to the obviously complicating fact of more than two currencies, airport currency agents immediately realize the commission/margin on sales of currency A (which immediately increases inventory of currency B more than the equivalent value of currency A). Furthermore, the exchange rate (and possibly even the exchange premium) may also fluctuate with market conditions. This is a promising area for future theoretical work.

6.1 Conclusion

This paper introduces the context of mobile money and the mobile money agent's challenge of balancing inventory costs (expected stockout losses with cost of capital) for both cash and e-float. This setting presents a unique inventory challenge: how should a firm stock when sales of one good generate inventory of another? Because the "true" model is intractable, two approaches are taken to generate recommendations for starting inventories of cash and e-float. These two approaches, one based on a "brute force" Markovian model and the other involving a simple analytical heuristic are tested in a variety of simulated settings, as well as against the actual decisions made by mobile money agents in an East African country. While recommendations from both models can substantially increase revenue net of cost of capital, the net demand heuristic does so while also being substantially simpler to implement than the Markov model.

Acknowledgments

Sincerest thanks to Gautier Kings, Sebastien Deletaille, Pierre Boel, Olivier Thierry, Thibault Rouby, and Maxime Temmerman of Real Impact Analytics for anonymizing data and coordinating data access. Thanks also to the mobile network partner in East Africa who supported this research. Sincerest thanks to Jason Acimovic, Joel Goh, and Kris Johnson Ferreira for extremely helpful guidance and encouragement. Thanks also to Andrew Marder of Harvard Business School's Research Computing Services for his invaluable code optimization assistance. The authors are grateful to the Harvard Business School Doctoral Programs Office for generous research support.

Appendix

A Proofs

Proof of Lemma 1: It is sufficient to show that the cash and e-float underage, as well as the inventory evolution can be written in terms of D_t. Let $D_t = D_t^c - D_t^e = \mathbb{1}_{D_t^c>0} \cdot D_t^c - \mathbb{1}_{D_t^e>0} \cdot D_t^e$.

$$
\begin{aligned}
Underage &= (D_t^c - q_t)^+ + (D_t^e - f_t)^+ \\
&= \mathbb{1}_{D_t^c>0}(D_t^c - q_t)^+ + \mathbb{1}_{D_t^e>0}(D_t^e - f_t)^+ \\
&= (\mathbb{1}_{D_t^c>0} \cdot D_t^c - \mathbb{1}_{D_t^c>0} \cdot q_t)^+ + (\mathbb{1}_{D_t^e>0} \cdot D_t^e - \mathbb{1}_{D_t^e>0} \cdot f_t)^+ \\
&= (\mathbb{1}_{D_t^c>0} \cdot D_t^c - \mathbb{1}_{D_t^e>0} \cdot D_t^e - \mathbb{1}_{D_t^c>0} \cdot q_t)^+ + (-\mathbb{1}_{D_t^c>0} \cdot D_t^c + \mathbb{1}_{D_t^e>0} \cdot D_t^e - \mathbb{1}_{D_t^e>0} \cdot f_t)^+ \\
&= (D_t - \mathbb{1}_{D_t^c>0} \cdot q_t)^+ + (-D_t - \mathbb{1}_{D_t^e>0} \cdot f_t)^+ \\
&= (D_t - q_t)^+ + (-D_t - b + q_t)^+
\end{aligned}
$$

For inventory evolution:

$$
\begin{aligned}
q_{t+1} &= \mathbb{1}_{D_t^e>0} \cdot \min(b, q_t + D_t^e) + \mathbb{1}_{D_t^c>0} \cdot \max(q_t - D_t^c, 0) \\
&= \min(b, \max((q_t - \mathbb{1}_{D_t^c>0} \cdot D_t^c + \mathbb{1}_{D_t^e>0} \cdot D_t^e), 0) \\
&= \min(b, \max(q_t - D_t, 0)) \square
\end{aligned}
$$

Proof of Proposition 1.A: The inventory evolution equation is:

$$
q_{t+1} = q_t - \mathbb{1}_{D_t>0} \cdot \min(q_t, D_t) + \mathbb{1}_{D_t<0} \cdot \min(b - q_t, -D_t)
$$

By the definition of $\hat{\Delta}$ and $\check{\Delta}$, as well as $q_1 \geq \hat{\Delta}$ and $(b - q_1) \geq -\check{\Delta}$:

$$q_1 \geq \max_{1 \leq t \leq M} (\sum_{t=1}^{M} D_t) \Rightarrow q_1 \geq D_1$$
$$b - q_1 \geq -\min_{1 \leq t \leq M} (\sum_{t=1}^{M} D_t) \Rightarrow b - q_1 \geq -D_1$$

Substituting into the inventory evolution equation:

$$\begin{aligned} q_2 &= q_1 - D_1 \\ q_3 &= q_2 - \mathbb{1}_{D_2>0} \cdot \min(q_2, D_2) + \mathbb{1}_{D_2<0} \cdot \min(b - q_2, -D_2) \\ &= q_1 - D_1 - \mathbb{1}_{D_2>0} \cdot \min(q_1 - D_1, D_2) + \mathbb{1}_{D_2<0} \cdot \min(b - q_1 + D_1, -D_2) \end{aligned}$$

Using the fact that $q_1 \geq D_1 + D_2$ and $b - q_1 \geq -(D_1 + D_2)$:

$$q_3 = q_1 - D_1 - D_2$$
$$\ldots$$
$$q_M = q_1 - \sum_{t=1}^{M-1} D_t$$

Then:

$$q_t - D_t \geq 0 \ \forall \ t \leq M$$
$$b - q_t + D_t \geq 0 \ \forall \ t \leq M$$

It follows immediately that:

$$m_c \sum_{t=1}^{M} (D_t - q_t)^+ + m_e \sum_{t=1}^{M} (-D_t - (b - q_t))^+ = 0 \square$$

Proof of Proposition 1.B: Starting from proposition 1.A:

$$(\hat{\Delta})^+(-\check{\Delta})^+ = (\max_{1\le j\le t}\sum_{j=1}^{t} D_j)^+ + (-\min_{1\le j\le t}\sum_{j=1}^{t} D_j)^+$$
$$< \sum_{j=1}^{t}(D_j)^+ + \sum_{j=1}^{t}(-D_j)^+ \square$$

Proof of Proposition 2.A: Let l_t and u_t represent the cumulative underage cost of cash and e-float respectively of the true model after t arrivals. These values can then be compared to the net demand heuristic underage to directly prove the result. For cash underage:

$$\begin{aligned} l_t &= \max_{1\le j\le t}(q_1 - \Delta_j - u_j)^- \\ &= \max_{1\le j\le t}(\Delta_j - q_1 + u_j)^+ \\ &\ge \max_{1\le j\le t}(\Delta_j - q_1)^+ \\ &= (\hat{\Delta} - q_1)^+ \end{aligned}$$

Analogously, for e-float underage, the result follows from:

$$\begin{aligned} u_t &= \max_{1\le j\le t}(b - q_1 + \Delta_j - l_j)^- \\ &= \max_{1\le j\le t}(-\Delta_j + l_j - b + q_1)^+ \\ &\ge \max_{1\le j\le t}(-\Delta_j - b + q_1)^+ \\ &= (-\check{\Delta}_t - b + q_1)^+ \end{aligned}$$

Proof of Proposition 2.B: There are four cases: Case A: $l_t > 0$ and $u_t = 0$, Case B: $l_t = 0$ and $u_t > 0$, Case C: $l_t > 0$ and $u_t > 0$, and Case D: $l_t = 0$ and $u_t = 0$. Note that when there is no true model cash and/or e-float underage (i.e., $l_t = 0$ and/or $u_t = 0$), sharpness follows from Proposition 2.A and the fact that heuristic underage is non-negative. This demonstrates sharpness in Case D. Sharpness is further demonstrated for positive underage

in Cases A and B, while non-sharpness is demonstrated in Case C.

$$\begin{aligned}\textbf{Case A: } l_t &= \max_{1\le j\le t}(q_1-\Delta_j-u_j)^- \\ &= \max_{1\le j\le t}(q_1-\Delta_j-0)^- \\ &= \max_{1\le j\le t}(q_1-\Delta_j)^-\end{aligned}$$

$$\begin{aligned}\textbf{Case B: } u_t &= \max_{1\le j\le t}(b-q_1+\Delta_j-l_j)^- \\ &= \max_{1\le j\le t}(-\Delta_j+0-b+q_1)^+ \\ &= (-\check{\Delta}_t-b+q_1)^+\end{aligned}$$

$$\begin{aligned}\textbf{Case C: } l_t &= \max_{1\le j\le t}(q_1-\Delta_j-u_j)^- \\ &> \max_{1\le j\le t}(q_1-\Delta_j)^-\end{aligned}$$

$$\begin{aligned}u_t &= \max_{1\le j\le t}(b-q_1+\Delta_j-l_j)^- \\ &> (-\check{\Delta}_t-b+q_1)^+\square\end{aligned}$$

Proof of convexity of equation 2.2: The cost function used to derive the net demand heuristic:

$$\tilde{G}(q_1,b)=\mathbb{E}[(m_c\cdot(\hat{\Delta}-q_1)^+]+\mathbb{E}[m_e\cdot(-\check{\Delta}-(b-q_1))^+]+\gamma\cdot b)$$

The function's Hessian matrix is positive-semidefinite:

$$\mathbf{H}\tilde{G}(q_1, b) = \begin{bmatrix} \frac{\partial^2 \tilde{G}q_1, b)}{\partial q_1^2} & \frac{\partial^2 \tilde{G}q_1, b)}{\partial q_1 \partial b} \\ \frac{\partial^2 \tilde{G}q_1, b)}{\partial b \partial q_1} & \frac{\partial^2 \tilde{G}q_1, b)}{\partial b^2} \end{bmatrix}$$

To show that this matrix is positive-semidefinite, it is sufficient to show that the Hessian's principal minors are non-negative: $\frac{\partial^2 \tilde{G}q_1, b)}{\partial b^2} \geq 0$, $\frac{\partial^2 \tilde{G}q_1, b)}{\partial q_1^2} \geq 0$, and $\frac{\partial^2 \tilde{G}q_1, b)}{\partial b^2} \frac{\partial^2 \tilde{G}q_1, b)}{\partial q_1^2} - (\frac{\partial^2 \tilde{G}q_1, b)}{\partial b \partial q_1})^2 \geq 0$.

m_c, m_e, and probabilities are non-negative, so:

$$\begin{aligned} \frac{\partial^2 \tilde{G}q_1, b)}{\partial q_1^2} &= m_c \cdot f_{\hat{\Delta}}(q_1) + m_e \cdot f_{\check{\Delta}}(q_1 - b) \geq 0, \\ \frac{\partial^2 \tilde{G}q_1, b)}{\partial b^2} &= me \cdot f_{\check{\Delta}}(q_1 - b) \geq 0, \\ \frac{\partial^2 \tilde{G}q_1, b)}{\partial b^2} \frac{\partial^2 \tilde{G}q_1, b)}{\partial q_1^2} - (\frac{\partial^2 \tilde{G}q_1, b)}{\partial b \partial q_1})^2 &= m_c \cdot f_{\hat{\Delta}}(q_1) \cdot m_e \cdot f_{\check{\Delta}}(q_1 - b) \geq 0 \square \end{aligned}$$

Proof of Proposition 3: The first order condition of equation 2.2 with respect to q_1 yields the optimality condition:

$$m_c \cdot (1 - F_{\hat{\Delta}}(q_1)) = m_e \cdot (F_{\check{\Delta}}(q_1 - b)) \tag{2.3}$$

The first order condition with respect to b results in:

$$F_{\check{\Delta}}(q_1 - b) = \frac{\gamma}{m_e} \tag{2.4}$$

Substituting equation 2.4 into equation 2.3 yields:

$$F_{\hat{\Delta}}(q_1) = 1 - \frac{\gamma}{m_c}$$

Thus, the optimal q_1 and f_1 are:

$$q_1^* = \left(F_{\hat{\Delta}}^{-1} \left(1 - \frac{\gamma}{m_c} \right) \right)^+$$
$$f_1^* = \left(-F_{\check{\Delta}}^{-1} \left(\frac{\gamma}{m_e} \right) \right)^+ \square$$

B Estimating cash inventory and stockouts

B.1 Estimating cash inventory

While historical e-float balances (and thus, agent e-float stocking decisions) are directly observable over time in our historical transaction data, cash inventory is not directly observable in the data and thus must be estimated. In order to estimate historical cash decisions made by agents, agent-level e-float balances over time are utilized in two-week blocks prior to any given day. Two assumptions are made. First, we assume that each agent's budget does not fluctuate over the two-week block; and second, we assume each agent has stocked out of cash (and thus has allocated all of her budget towards e-float) at some point over the two weeks. Under these two assumptions, the maximum value of e-float balance represents the agent's inventory budget. From the budget estimate, the agent's cash balance at any given time can be estimated by subtracting the current e-float balance from the maximum balance over the two-week block. The assumption of budgets not fluctuating from the maximum value of e-float over a two-week block, in general, is not likely to hold precisely. It is therefore helpful to observe that, if this assumption is violated, this methodology overestimates the budget, and likely also overestimates cash inventory. Because this process likely over-estimates the agent's cash inventory at any given time, this paper presents only comparisons across models of cash-in revenue net cost of capital attributable to e-float. Our results showing that the net demand heuristic performs significantly better than actual agent decisions are robust to the choice of the duration used to calculate agents' budgets; using a week, a month, and the

entire sample period as the duration (rather than two weeks) all yield similar results.

B.2 Reconstructing censored demand

Due to demand censoring, historical sales data are not equivalent to historical demand. There are various methodologies proposed in the literature to derive demand from sales (van Ryzin and Talluri (2005) provide a description of many common methodologies). However, in general, these methods focus on scenarios where firms are unable to satisfy demand until the next period once a stockout has occurred. These approaches are not applicable in the mobile money setting. In the mobile money setting, inventory of e-float (cash) can be generated mid-period from sales of cash (e-float) — making it possible for stockouts to occur and be resolved without external intervention in a given day. For this reason, we opt to employ a simple three-step imputation process to estimate demand. To recreate total historical demand for each agent-day, we 1) estimate the timing and duration of stockouts of e-float (cash), 2) estimate e-float (cash) demand arrival rates, and 3) estimate the magnitude of e-float (cash) lost sales to insert in each e-float (cash) stockout interval.

In the first step, we estimate the timing and duration of stockouts. We define the e-float (cash) stocked out state as any interval within which an agent is holding less than a threshold amount of e-float (cash) inventory. This threshold is agent-specific and was chosen as half of each agent's mean e-float (cash) transaction size over the sample period. The results presented in this paper are robust to changes in threshold; using the mean, a fourth of the mean, and median do not produce meaningfully different results. Because transactions are time-stamped with pre and post balances, the time and duration of each estimated e-float stockout interval can be exactly determined. Similarly, with the assumption of fixed budget allowing us to infer cash balances at all points in time, the time and duration of each estimated cash stockout interval can also be determined. In the second step, we estimate an arrival rate for each agent-day. To generate these arrival rates, we calculate the average arrival rate (of units of cash and e-float demand, respectively) for each agent for each day

of the week, using only days that the agent had inventory levels of cash and e-float above the stockout thresholds for the entire day. This is done because the day of the week is a significant factor in demand arrival rate. In the third step, we estimate the magnitude of demand arrivals to insert into each stockout interval. For each duration of stockout, we estimated this quantity by taking the product of the duration of the stockout and the arrival rate of e-float (cash) demand. After imputing these stockouts, they are inserted into their corresponding time-intervals in order to generate the estimated sequenced demand for each agent-day.

Chapter 3

Inventory Pooling for Mobile Money Agents in the Developing World

1 Abstract

In the past decade, systems that enable people to send and receive money with their cell phones, called mobile money platforms, have grown at an astonishing rate in the developing world. However, mobile money agents, who perform the critical functions of converting cash to electronic value and vice versa for customers, are often stocked out of cash or electronic value. Additionally, a significant barrier to opening and operating a mobile money agency is the high working capital requirements to finance inventories of cash and electronic value. We develop a framework for an inventory pool of electronic value that can significantly decrease the working capital burden on agents, while also increasing inventory service levels. This framework achieves these objectives by harnessing not only the power of traditional variation pooling, but also the "recycling effect" resulting from the fact that agents can remit electronic value back to the pool when they satisfy demand for cash. We test this model with a large dataset of mobile money transactions from Zambia, and show that a basic inventory pool can decrease system-wide inventory requirements by over 74% and increase system-wide revenue

net of cost of capital by over 8%. We also describe extensions to these models that should be developed before implementing a pooling framework in the field to ensure regulatory compliance and incentive compatibility.

2 Introduction

In the past decade, systems that enable people to send and receive money with mobile phones have grown at a stunning rate in the developing world, with over 700 million registered accounts in 2017 (Groupe Speciale Mobile Association 2018). Sub-Saharan Africa, not generally known for positive economic news, has been the epicenter of the mobile money explosion. The large proportion of the adult population (over 80%) that does not have access to a bank account is a key reason mobile money has grown so quickly (Demirguc-kunt 2012); "unbanked" people face a critical pain point in sending and receiving money, which traditionally must be done in physical cash. Mobile money has thrived in this fertile environment because it can help people move money over long distances quickly, reliably, and inexpensively. Mobile money is also a promising first step towards a vision of financial inclusion. Firms offering financial products like savings, credit, and insurance are unable to serve the poor due to the cost drivers of transacting in person and in physical cash (Mas 2010). However, if these cost drivers are removed so that people can access financial products exclusively in electronic form, serving even the poorest customers can become a viable business model (Kendall 2011). In the transition period before this tantalizing vision of digital financial inclusion can be realized, a reliable network of mobile money agents – who serve as the "bridges" from cash to electronic value (hereafter referred to as "e-float") and vice versa – is necessary for the growth of mobile money systems (Eijkman et al. 2009). These agents, of whom there are hundreds of thousands globally, convert cash to e-float ("cash-in" transactions) and e-float back to cash ("cash-out" transactions), each for a commission. In conducting these cash-in and cash-out (CICO) transactions, agents receive inventory of

cash when they satisfy demand for e-float, and receive inventory of e-float when they satisfy demand for cash.

2.1 Inventory challenges

In managing CICO transactions, the agent's most basic challenge is deciding how much cash and e-float to carry in order to most profitably support their mobile money business. This inventory problem is a non-trivial challenge because in this setting, the agent not only serves uncertain demand for cash and e-float, but each sale of cash (e-float) also generates equivalent inventory of e-float (cash). Because mobile money agents are independent business owners, they must invest their own capital into their agencies to finance inventory of cash and e-float.[1] Though moving e-float once it has been issued is clearly easier than moving cash, agents can, and do stock out of both e-float and cash. This problem is exacerbated by the fact that many agents in competitive markets serve multiple operators, forcing their investment in e-float to be fractured into separate, non-interoperable liquidity pools. This increases the likelihood of being stocked out of any one type of e-float (Kiarie and Wright 2017). Surveys of mobile money customers and mobile money agents show that stockouts of both cash and e-float are a significant problem and may be holding back the progression of mobile money ecosystems. Agents in three separate surveys in Kenya, Tanzania, and Uganda cited "lack of resources to stock enough cash and e-float" as a major barrier to growing the business – second only to nearby competition (Githachuri et al. 2014, McCaffrey et al. 2014, Githachuri et al. 2013). Indeed, informal interviews conducted by the authors indicated that, in order to avoid the cost of stocking more inventory, many agents deal with stockouts by just waiting until the reverse transaction occurs to replenish inventories. But latency in this setting is pernicious – waiting hours for e-float or cash to arrive may cause customers to go to another agent, switch operators, or even worse, decrease their use of mobile money. Ultimately, these occurrences

[1]Note that because e-float is actual currency, it cannot be "created" on the spot by either the agent or the operator. Each unit of e-float an operator issues must be backed by traditional deposits at a prudentially regulated financial institution.

of stockouts decrease customers' faith in the convenient convertability of cash to e-float and vice versa (Wright 2015).

2.2 Inventory pooling

One possible mechanism to alleviate inventory stockouts and enhance the agent's business case is inventory pooling. Inventory pooling refers to the practice of "pooling" inventory once held in separate locations (e.g., retail stores) into a single location (e.g., a warehouse) that supplies the separate locations on an as-needed basis. Pooling has potential to reduce the overall need for inventory in supply chains by aggregating the variability in demand across many locations. While pooling inventory of cash is not attractive due to high transportation costs and latency, moving e-float across long distances can be achieved at a near-zero cost and can be done nearly instantaneously. It then becomes clear that it may be possible to realize the benefits of pooling inventory of e-float centrally – improving both agent service levels and customer experiences without incurring the significant transportation costs and latency generally associated with physical inventory pooling. The traditional benefit of inventory pooling that has been studied in the operations management literature revolves around the ability to decrease system-wide inventory by taking advantage of "statistical economies of scale" stemming from the reduction of overall variability when demand is aggregated. The other related, salient benefit arises from the fact that when an agent satisfies demand for cash (e-float), she generates e-float (cash); this may allow the agent to stock far less than the total demand of cash and e-float that she satisfies during the day. This "recycling effect" is capitalized upon to some extent by most agents, enabling many inventory turns throughout the day. This off-sets the relatively low margins on satisfying cash and e-float demand. The "recycling effect" can potentially generate even more value when it is aggregated across agents. For example, any excess (idle) inventory of e-float carried by agent A could be lent to agent B who could service a cash-in transaction without making a working capital investment. This debt could be quickly repaid when agent B receives demand for

cash (a cash-out transaction). The e-float generated in the cash-out transaction could be then utilized again by agent A for a cash-in transaction. This process can be formalized and potentially centralized for maximum system-wide benefit. As will be explored in this paper, if the operator were to off-load all responsibility of carrying e-float from agents, the operator would only need to meet the maximum cumulative system-wide e-float demand. By consolidating e-float inventory centrally, pooling allows agents the ability to direct their limited budgets toward holding greater inventories of cash, potentially leading to fewer cash stockouts in addition to reducing e-float stockouts. These reductions in stockouts further benefit agents due to fewer lost sales, and also benefits customers who will face frustrating and confidence-degrading stockouts less often. Reductions in stockouts are also beneficial to the operator, as revenues may grow as a result of being seen as more reliable by customers.

2.3 Research questions and preview of results

This paper presents a series of models that are progressively more sophisticated in capturing "real world" complexity. The paper begins with the description and model of the "status quo," a scenario in which all agents manage their own inventory and maximize their own revenue net of cost of capital independently. Next, a vertically integrated entity (termed the "omni-agent") scenario is presented, where one single entity satisfies all demand centrally. This model, which yields the maximum possible channel net revenue, can be compared to the status quo. We use anonymized data from a mobile money operator in Zambia to demonstrate this maximum pooling benefit: the omni-agent can decrease system-wide inventory requirements by over 74% and increase system-wide net revenue by over 8%. Next, we model a scenario with a central pool that stocks e-float for and shares revenue with agents. We then identify the "double-marginalization" challenge that arises in this context and show how the framework could be tweaked in a simple way to incent first-best stocking decisions. Next, we examine the viable ranges for parameters that ensure welfare gains among all participants. We lastly describe future work that is needed to create a deployable

pooling framework. The first critical extension involves accounting for a common government regulation restricting mobile operators from borrowing from agents. The second extension involves developing lending thresholds as a function of agent characteristics to mitigate default risk. The final extension focuses on determining the characteristics of agents best suited to the pool, and developing incentives that are most attractive to these best-suited agents.

3 Literature Review

Since Eppen's seminal work in 1979, inventory pooling has been a mainstay of academic research and supply chain practitioner interest. Eppen demonstrated the potential for "statistical economies of scale" by showing that for independent and identical normally distributed demands, the amount of inventory required to serve aggregate demand centrally is significantly less than the aggregate inventory required by a decentralized system. Eppen showed that the safety stock required by a centralized system is proportional to $\sqrt{n}$, where n is the number of retail locations served. He further demonstrates that if demands are correlated, then the cost savings depends on the extent to which they are correlated. In the worst case, when demands are perfectly positively correlated, there are no cost savings. But when demand is negatively correlated, the cost savings increase (Eppen 1979). Eppen's assertion that system inventory costs under a centralized system decrease as the individual demand streams are less positively correlated was extended over time to increasingly more general demand distributions by Federgruen and Zipkin (1984) and Corbett and Rajaram (2006). Mak and Shen (2014) show that this result holds even when only partial distribution information (e.g., means and variances) is known. But the benefits of pooling may not always be as large as Eppen described: Bimpikis and Markakis (2015) show that the benefits of pooling decrease as distributions become more "heavy-tailed" (i.e., where extreme events are more likely). However, these papers do not address scenarios where the satisfaction of demand for

one good generates inventory of the other (i.e., currency exchange settings such as mobile money). This "recycling effect" may significantly enhance the benefits of pooling.

This paper also benefits from rich literature on supply chain coordination. The pool frameworks we propose, in which responsibility of stocking e-float is off-loaded completely from the agent, is related to "vendor-managed inventory," a popular supply chain concept in which suppliers handle inventory decisions for retailers (e.g., Cetinkaya and Lee 2000, Choi et al. 2004). As will be seen in §6.1, an incentive mis-alignment can arise between the pool and participating agents. Specifically, when revenue is shared between the pool and agents, the pool's optimal inventory quantity is lower than the central planner's quantity – a well-studied issue termed "double marginalization." (e.g., Tirole 1988). While there are a plethora of mechanisms for supply chain coordination proposed in the literature (see Tayur et al. 2012 for a comprehensive review), the "revenue-sharing" contract studied by among others, Cachon and Lariviere (2005) and Mortimer (2002), is used as a basis for the mechanisms we present here. While contracting mechanisms have been studied extensively, to the best of our knowledge, this paper is the first to analyze supply-chain coordinating mechanisms as they relate to "two-way" inventory pools, where firms can both draw down from and push up to centralized inventory pools. In the mobile money setting, inventory is not depleted by agents as they satisfy demand (which is generally the case with standard retailers). Instead, the agents not only draw e-float from the pool when conducting a cash-in transaction, they can also return e-float back to the pool when they conduct a cash-out transaction.

4 Modeling preliminaries

This section first describes parameters that will be used in modeling, then describes the status quo, and finally describes data that will be used in analyzing the frameworks.

4.1 Description of parameters

A set of mobile money agents, A, is indexed by an identifier $a \in A$. These agents satisfy demand for cash and e-float throughout the day. Each day is split into N time periods of equal duration, indexed by t. The duration of each time period is chosen to be arbitrarily small. Thus, the probability of more than one transaction (conducted by any agent $a \in A$) is small enough that we assume no more than one transaction occurs each time period. Previous work by the authors (Balasubramanian et al. 2017) show that demand for both cash and e-float can be collapsed into a single variable D_{at}, where positive values of D_{at} represent e-float demand for agent a at time t, and negative values represent cash demand. Agent a chooses starting values of e-float q_a and cash s_a to fulfill both e-float demand and cash demand.[2] Successful satisfaction of e-float demand yields per-unit commission m_e and successful satisfaction of cash demand yields per-unit commission m_c. The per-unit commissions are exogenously determined (i.e., no entity discussed in this paper can alter commissions). The sum of each agent's inventories of cash and e-float is b_a, which incurs per-unit daily cost of capital γ. Table A.1 lists these parameters and summarizes their respective descriptions.

4.2 Description of status quo

The status quo involves individual agents operating independently (in an inventory management sense) from the operator and other agents. The agent's inventories of both cash and e-float fluctuate throughout the day (while always summing to b_a due to the sales-inventory relationship). The inventory of e-float evolves as:

$$q_{a,t+1} = \min\left(b_a, \max\left(q_{at} - D_{at}, 0\right)\right) \ \forall t \in \{1, 2, ..., N-1\}$$

[2]Note two slight changes of notation between this paper and previous work: while Balasubramanian et al. (2017) represent e-float demand as negative values of combined demand D_{at}, this paper represents e-float demand as positive. Similarly, while q represented cash inventory in previous work, q represents e-float inventory in this paper.

The total cost incurred by the agent each day is the sum of the cost of capital and stockouts of cash and e-float for each unit of time until the final unit of time, N.

$$G(q_a, b_a) = \sum_{t=1}^{N} \left(\mathbb{E}_{D_{at}} \left[m_e (D_{at} - q_{at})^+ + m_c (-D_{at} - b_a + q_{at})^+ \right] \right) + b_a \cdot \gamma$$

Because solving this problem directly for q_a^* and b_a^* is intractable, Balasubramanian et al. (2017) show that the net demand $\Delta_{at} = \sum_{i=1}^{t} D_{ai}$ for each agent-day can form the foundation for a simple analytical heuristic to determine how much inventory of e-float and cash agents should stock. Specifically, it is shown that if an agent stocks more electronic credit than the maximum cumulative demand $\hat{\Delta}_{aN}$ and more cash than the negative of the minimum cumulative demand $-\check{\Delta}_{aN}$, no stockouts occur. It is also true that the maximum and minimum cumulative demand for a day is sufficient for the inventory calculation, regardless of number of arrivals on that particular day, allowing us to drop the time index subscript on maximum cumulative demand (i.e., $\hat{\Delta}_a$) and minimum cumulative demand (i.e., $\check{\Delta}_a$). It is further shown that the quantities $(\hat{\Delta}_a - q_a)^+$ and $(-\check{\Delta}_a - s_a)^+$ are lower bounds on the cash and e-float underage that hold at equality in all cases other than when there are "double-stockouts" (i.e., stockouts of both cash and e-float on the same day). From these underage cost approximations, a cost function can be constructed and used for a heuristic that generates recommendations of starting cash and e-float inventories. It is shown at reasonable levels of γ (i.e., γ below 200% annualized cost of capital), a heuristic can capture over 99% of the optimal revenue net of capital cost in a variety of simulated settings. The heuristic recommendations are generated as follows:

$$q_a^* = \left(F_{\hat{\Delta}_a}^{-1} \left(1 - \frac{\gamma}{m_e} \right) \right)^+ \tag{3.1}$$

$$s_a^* = \left(-F_{\check{\Delta}_a}^{-1} \left(\frac{\gamma}{m_c} \right) \right)^+ \tag{3.2}$$

Using the lower bound (and close approximation) of underage, an upper bound (and close approximation) of agent a's revenue net of cost of capital can be constructed:

$$\pi_a = m_e \left(\sum_{t=1}^{N} D_{at}^{+} - (\hat{\Delta}_a - q_a)^{+} \right) + m_c \left(\sum_{t=1}^{N} D_{at}^{-} - (-\check{\Delta}_a - s_a)^{+} \right) - \gamma(q_a + s_a)$$

The pooling framework being proposed in this paper focuses only on the e-float side of the agent's business, so the relevant total system-wide e-float net revenue under the status quo is labeled π_A:

$$\pi_A = \sum_{a \in A} \left(m_e \left(\sum_{t=1}^{N} D_{at}^{+} - (\hat{\Delta}_a - q_a)^{+} \right) - \gamma q_a \right)$$

4.3 Description of historical data

In order to test the model frameworks presented in this paper, we obtained anonymized transaction-level data from a mobile money platform operating in Zambia. The sample contained 2,006,793 transactions from 76 agents over the period from January 1, 2015 through December 31, 2016. The data only includes records of agents who had conducted at least one cash-in (e-float demand) and one cash-out (cash demand) transaction for 520 days (roughly a five-day work-week) over the course of the sample period. Each transaction record features a time-stamp, transaction type, and transaction value. This data provides a sequenced vector of e-float and cash sales for every agent (as well as the entire system). This allows for the calculation of maximum cumulative demand and minimum cumulative demand for all agents and the system as a whole. We do, however, note limitations: neither cash inventory nor stockouts (because the sales data represent censored demand) are directly observable in the data. Not having centralized cash inventory information is one reason we focus specifically on the effect of pooling inventory of e-float (the other reason being that transportation of physical cash incurs both cost and latency). While augmenting the sales data with stockouts to produce a demand vector that includes estimated demand during periods of low inventory can be helpful in refining this analysis, we leave that exercise for future work. Instead, we

will treat the sequenced sales information gleaned from the transaction log as the demand vector itself.

5 Vertically integrated pooling framework

Pooling inventory of e-float has the potential to grow the "pie" of net revenue by reducing system-wide inventory costs and increasing revenue capture. The maximum system-wide net revenue can be calculated if one considers the problem of a single entity, acting as an "omni-agent" (central planner) who is able to serve all demand system-wide with one single pool of cash and e-float inventory. The omni-agent faces a different cost of capital, represented as Γ. The problem of the omni-agent is structurally identical to the single agent's problem considered in the previous section, but it will be shown that it has potential to yield substantially higher net revenue. Because all demand is satisfied by a single agent, we can drop the "a" subscript from demand, and we can represent the omni-agent as "M".

$$\pi_M = m_e \left(\sum_{t=1}^{N} D_t^+ - (\hat{\Delta}_M - q_M)^+ \right) - \Gamma q_M$$

The maximum benefit of pooling is $\pi_M - \pi_A$. It is the case that $\pi_M > \pi_A$ for four reasons. First, $\Gamma < \gamma$, because a large entity is generally able to raise capital more cheaply than small-scale entrepreneurs in the developing world (Quaye and Hartarska 2016). Second, because $\Gamma < \gamma$, the fractile that applies to the maximum cumulative demand distributions is larger in the pooled case ($1 - \frac{\Gamma}{m_e} > 1 - \frac{\gamma}{m_e}$). This results in a higher service level and thus decreased overall stockout losses. Thus, more revenue is captured in the vertically integrated case. The third and fourth reasons that $\pi_M > \pi_A$ relate to the fact that for a fixed service level, $q_M^* < \sum_{a \in A} q_a^*$. This statement is true due to both the recycling effect and the traditional pooling effect. The recycling effect relates to the fact that one agent's cash-in demand might be offset by another's cash-out demand, thus lowering overall required inventory. The traditional pooling effect also reduces inventory. Given that the maximum

cumulative demands for agents are not perfectly positively correlated, variation can be pooled such that the overall coefficient of variation is reduced substantially. Notably, the traditional costs of pooling – transportation and latency – are not applicable in this setting because e-float can be transmitted instantly for essentially zero marginal cost.

5.1 Recycling effect

The mobile money setting is unique in that sales of cash generate inventory for e-float and vice versa. This fact allows agents to stock less cash and e-float inventory than the sum of sales of cash and e-float. The omni-agent reaps an even greater benefit from this effect than any individual agent alone – satisfaction of cash demand at one location (previously by an individual agent a) can generate e-float that is used at another location (previously by an individual agent a').The calculation generating q_M^* is structurally identical to that calculation of q_a^*, shown in equation 4.6. In this case, the relevant fractile is $1 - \frac{\Gamma}{m_e}$, which is applied to the omni-agent's maximum cumulative demand distribution, $\hat{\Delta}_M = \max_{1 \leq t \leq N} \sum_{j=1}^{t} D_j$. However, the recycling effect reduces the omni-agent's maximum cumulative demand due to this important observation:

Proposition 4 *The omni-agent's maximum cumulative demand is always less than or equal to the sum of the maximum cumulative demands of all of the agents. That is,* $\hat{\Delta}_M \leq \sum_{a \in A} (\hat{\Delta}_a)$

All proofs can be found in Appendix B. It is established in previous work that if initial stocking quantities of e-float and cash are greater than the maximum cumulative demand and the negative of the minimum cumulative demand, respectively, all demand is satisfied. The combination of this fact with Proposition 4 results in an important conclusion: assuming no cash stockouts, given any sequence of demand an omni-agent would need to carry less e-float inventory than the sum of e-float required by all agents individually. This result arises because a cash satisfaction by one individual agent could provide inventory of e-float required by another agent to satisfy e-float demand. Ultimately, the pool can do no worse

than the sum of all individual agents' inventories; the intuition for this statement is that the pool can always carry individual "pots" of e-float for agents, with each "pot" in the amount of each individual agent's maximum cumulative demand. We can represent the magnitude of the recycling effect mathematically as the ratio of sales (cash and e-float) to the minimum inventory required to satisfy all sales (with ex-ante knowledge of demand). This is analogous to a common metric of inventory efficiency, inventory turns. The only difference between the traditional definition of inventory turns (sales divided by average inventory) and the one presented here is that inventory is not depleted with a sale. In this case, we define inventory turns for the individual agent and omni-agent as T_a and T_M, respectively.

$$T_a = \frac{\sum_{t=1}^{N}(D_{at}^{+} + D_{at}^{-})}{(\hat{\Delta}_a)^{+} + (-\check{\Delta}_a)^{+}} \qquad T_M = \frac{\sum_{t=1}^{N}(D_{t}^{+} + D_{t}^{-})}{(\hat{\Delta}_M)^{+} + (-\check{\Delta}_M)^{+}}$$

An illustrative example is presented in Table A1, which shows a sequence of cash-in and cash-out arrivals for two agents, indexed 1 and 2. In the case of agent 1, if the agent began the day with e-float equal to at least the maximum cumulative demand (i.e., starts with at least 40 units of e-float) and cash equal to at least the negative of the minimum cumulative demand (i.e., 20 units or more of cash), then the agent would be able to satisfy all 100 total units of e-float and cash demand. Similarly, if agent 2 started the day with at least 20 units of e-float and 40 units of cash, agent 2 would be able to satisfy all 100 total units of cash and e-float demand. In this example, $T_1 = \frac{40+60}{40+20} = \frac{5}{3}$, while $T_2 = \frac{60+40}{20+40} = \frac{5}{3}$. Crucially, however, we see that $T_M = \frac{100+100}{20+30} = 4$. This more than doubles the turnover ratio of both agents 1 and 2, and allows for savings of 70 units of total inventory (40 units of e-float and 30 units of cash), while fulfilling all demand. While this contrived example is admittedly extreme, it does begin to illuminate how the recycling effect can create value.

To estimate the recycling effect's magnitude, we use transaction data from 76 Zambian agents to generate the daily difference between the omni-agent's e-float inventory requirements

time index, t	1	2	3	4	5	6	7	8	9	10
Agent 1 demand, D_{1t}	20	0	20	0	-30	0	-10	0	-20	0
Agent 2 demand, D_{2t}	0	-20	0	-20	0	30	0	10	0	20
Agent 1 cumulative demand, Δ_{1t}	20	20	40	**40**	10	10	0	0	-20	**-20**
Agent 2 cumulative demand, Δ_{2t}	0	-20	-20	-40	**-40**	-10	-10	0	0	**20**
Omni-agent cumulative demand, Δ_{Mt}	20	0	**20**	0	**-30**	0	-10	0	-20	0

Table A1: Illustrative example of demand processes for two agents over one day. This example shows how pooling can decrease inventory requirements.

and the sum of individual agent inventory requirements over a two-year period. To isolate the recycling effect, we consider minimum inventory requirements with ex-ante (hindsight) knowledge of demand. While this assumption of deterministic demand does not represent reality, it does allow for determining the rough magnitude of the recycling effect. Figure A1 compares the average monthly system-wide inventory required in the status quo with ex-ante knowledge of demand and the average monthly system-wide inventory required by the omni-agent with ex-ante knowledge of demand. The mean and median inventory reductions over the two-year period are both roughly 60%. This reduction is also roughly 60% when the average daily inventories and the average weekly inventories are compared.

5.2 Traditional pooling effect

Even in the absence of the recycling effect, the traditional benefits of pooling related to decreasing safety stock can apply here. Each agent stocks more than the expected maximum cumulative demand, because the fractile in equation 4.6 is generally quite high (larger than 0.9). If the maximum cumulative demands of agents are not perfectly positively correlated ($\rho_{a,a'} < 1 \quad \forall a, a' \in A$), then the pool could carry less than the sum of all individual agents' e-float holdings. Crucially, in this context, the traditional costs of pooling do not apply; there is no transportation costs and also essentially no latency. Eppen (1979) notes the statistical underpinning of the potential for safety stock reduction through the pooling of variability across locations. We represent the set of all agents with demand being satisfied by the omni-agent as A, with $\hat{A}$ representing the integer corresponding to $|A|$. We further

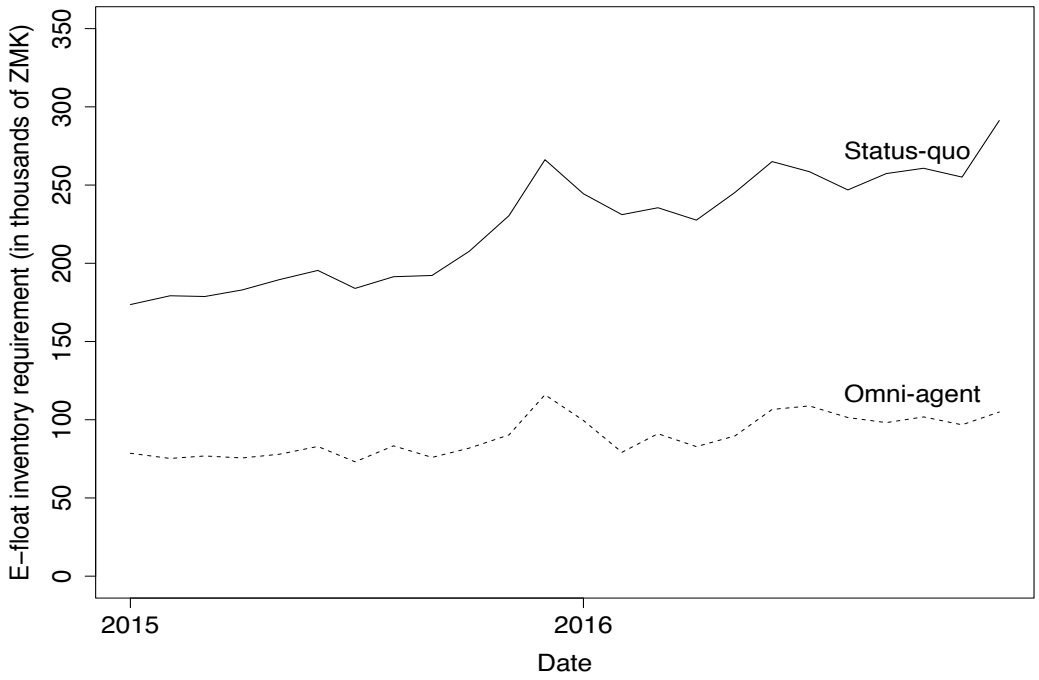


Figure A1: Hindsight inventory requirements of omni-agent and status quo agents averaged over each month. The omni-agent requires 60% less inventory on average.

represent the omni-agent M. The standard deviation of demand for M is:

$$\sigma_M = \sqrt{\sum_{a=1}^{\hat{A}} \sigma_a^2 + 2 \sum_{a=1}^{\hat{A}-1} \sum_{a'=a+1}^{\hat{A}} \sigma_a \sigma_{a'} \rho_{a,a'}}$$

Using the sample data of 76 agents over two years, the mean correlation coefficient between the maximum cumulative demands of these agents is nearly zero, as seen in figure A2. For the purposes of roughly estimating the magnitude of the traditional pooling effect, we will assume uncorrelated demand ($\rho_{a,a'} = 0 \ \forall a, a' \in A$) with similar distributions. In this case, the safety stock reduction possible is roughly $89\% = 1 - \frac{1}{\sqrt{76}}$. Indeed, using the specific empirical distributions of maximum cumulative demands of the sample of agents over the two year period and ignoring the recycling effect yields an approximate safety stock reduction of 88%, and a roughly 56% reduction in total inventory. This dynamic can be seen in figure

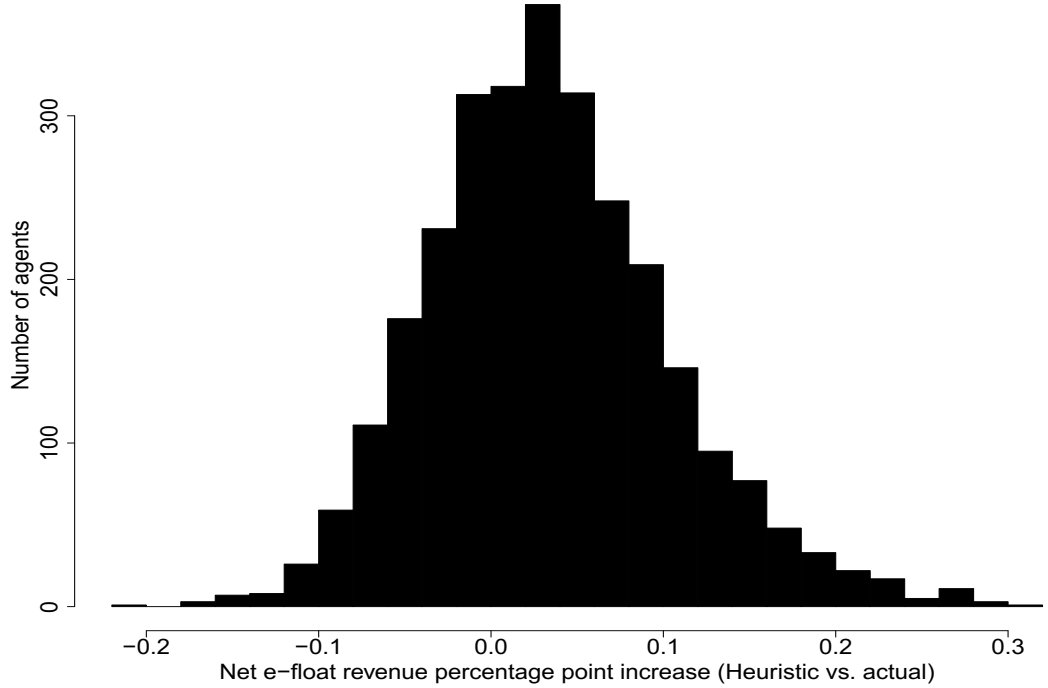


Figure A2: Histogram of correlation coefficients of maximum cumulative demand for each pair of agents. The mean and median correlation are approximately 0.03.

A3, which shows how the traditional pooling effect grows as the number of agents grows.

5.3 Combining the pooling effects

The recycling effect and traditional pooling effect combine to yield even larger pooling savings than each effect individually. This combined effect can be estimated by considering the difference between the omni-agent's recommended inventory and status quo inventories. In this case, we have no ex-ante knowledge of demand, so the net demand heuristic (equation 4.6) is required to determine recommended inventory levels. The heuristic uses the distribution of the relevant maximum cumulative demand. For all individual agents and the omni-agent, the empirical vectors of realized daily maximum cumulative demand over the course of the sample period can constitute the distributions of $\hat{\Delta}_a$ and $\hat{\Delta}_M$, respectively. It is possible to consider other methods of determining the distribution of demand, such as

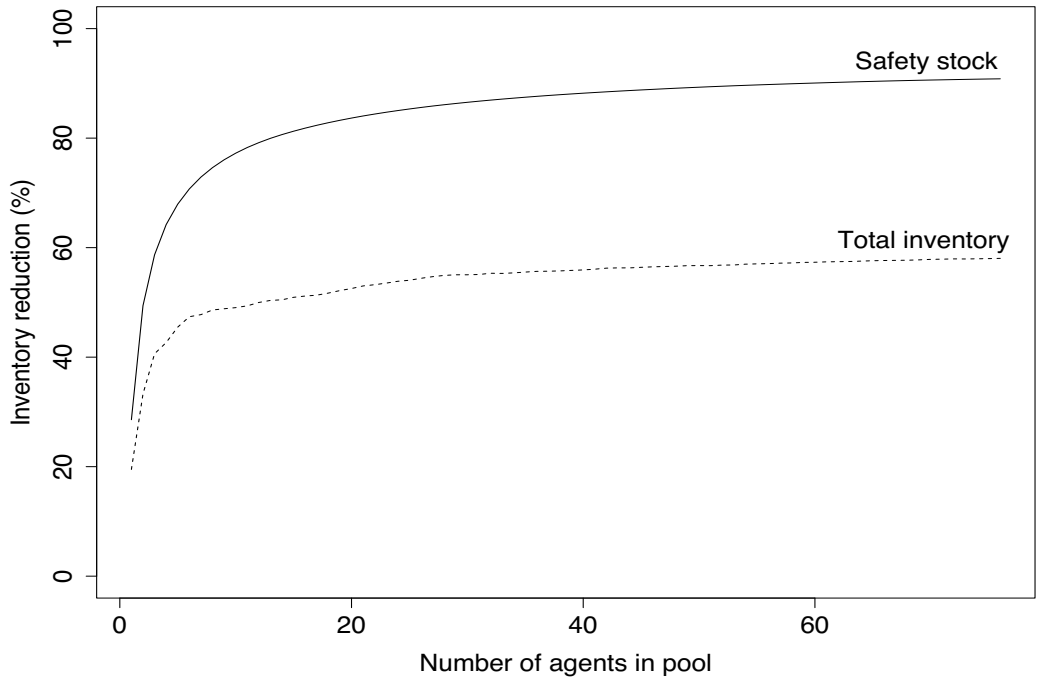


Figure A3: Estimated percent reduction in safety stock and total inventory due to the traditional pooling effect. The traditional pooling effect yields a 56% reduction in total inventory.

forecasting, but we use the simplest method here in order to keep the insight straightforward. The omni-agent's inventory requirement can be found as $\left(F^{-1}_{\hat{\Delta}_M}(1-\frac{\Gamma}{m_e})\right)^+$. Similarly, a total status quo, unpooled inventory requirement can be obtained as $\sum_{a\in A}\left(F^{-1}_{\hat{\Delta}_a}(1-\frac{\gamma}{m_e})\right)^+$. Using the sample of 76 agents over two years, we can calculate and compare the inventory requirements of the status quo and the omni-agent. Figure A4 shows that the pool reduces the system-wide inventory requirement by approximately 74% (shown as omni-agent I). When using an apples-to-apples comparison on service level by substituting γ for Γ in the omni-agent's calculation (shown as omni-agent II), the inventory reduction grows to 76%. Crucially, it is possible that the higher service level can be achieved at a lower cost than the status quo for a number of reasons, including the fact that $\Gamma \leq \gamma$. This can be observed when estimating the net revenue of the omni-agent versus the sum of individual agents.

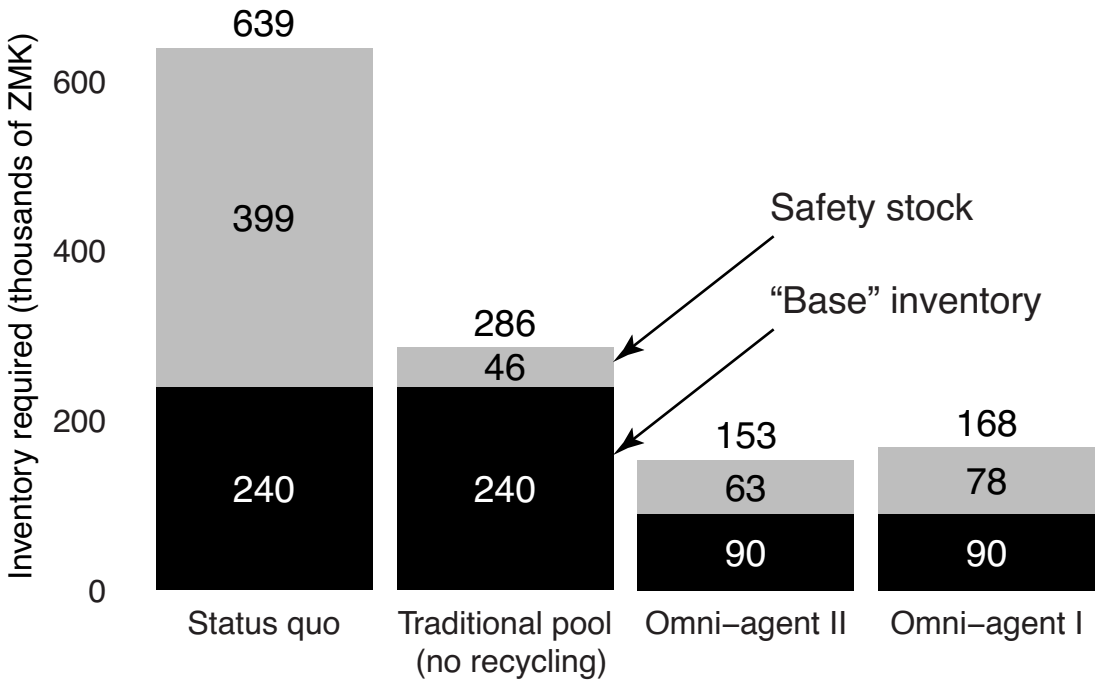


Figure A4: Comparison of inventory requirements between the status quo and the omni-agent (II corresponds to status quo service level, while I refers to higher pool service level). The omni-agent requires 70%+ less inventory than the status quo.

5.4 First-best channel net revenue

The omni-agent can achieve greater net revenue than the sum of the net revenues individual agents can achieve for the four reasons previously outlined: 1) lower cost of capital (i.e., $\Gamma < \gamma$), 2) greater revenue capture, 3) the recycling effect, 4) the traditional pooling effect. The magnitude of the total revenue differential is captured in the following equation:

$$\pi_M - \pi_A = m_e(-(\hat{\Delta}_M - q_M)^+ + \sum_{a \in A}(\hat{\Delta}_a - q_a)^+) - \Gamma q_M + \gamma \sum_{a \in A} q_a \tag{3.3}$$

All four effects combine for additional net revenue worth nearly 7.5 percentage points of total possible aggregate e-float revenue for all agents in the sample for an average day. Over half of the 7.5 percentage point improvement comes from the reduction in inventory due to

the recycling and traditional pooling effects. This analysis gives us a useful upper bound on the benefits of pooling, as the omni-agent stocks the e-float quantity that maximizes system profit.

In order to generate savings estimates, the following procedure was followed. First, calculate the maximum cumulative demand for each agent (and for the omni-agent) for each day. These vectors are represented as $\vec{q_a}$ for every given agent a and $\vec{q}_M$ for the omni-agent. Next, treat $\vec{q_a}$ and $\vec{q}_M$ as empirical distributions. Next, apply the critical fractile $(1 - \frac{\gamma}{m_e})$ to each distribution to arrive at the recommended starting values of e-float for the omni-agent, represented as q_M^*, and the individual agents, represted as q_a^*. Let $\bar{q}_M$ and $\bar{q}_a$ represent the mean values of the omni-agent and individual agent's historical empirical maximum cumulative demand distributions, respectively. Then, the raw magnitude of the recycling effect can be calculated as $\gamma(\sum_{a \in A} \bar{q}_a - \bar{q}_M)$. The safety stock of each entity can be calculated as $SS_M = q_M^* - \bar{q}_M$ and $SS_a = q_a^* - \bar{q}_a$, respectively. Thus, the traditional pooling effect can be captured as $\gamma(\sum_{a \in A} SS_a - SS_M)$. The savings associated with reduction in cost of capital is measured as $q_M^*(\gamma - \Gamma)$.

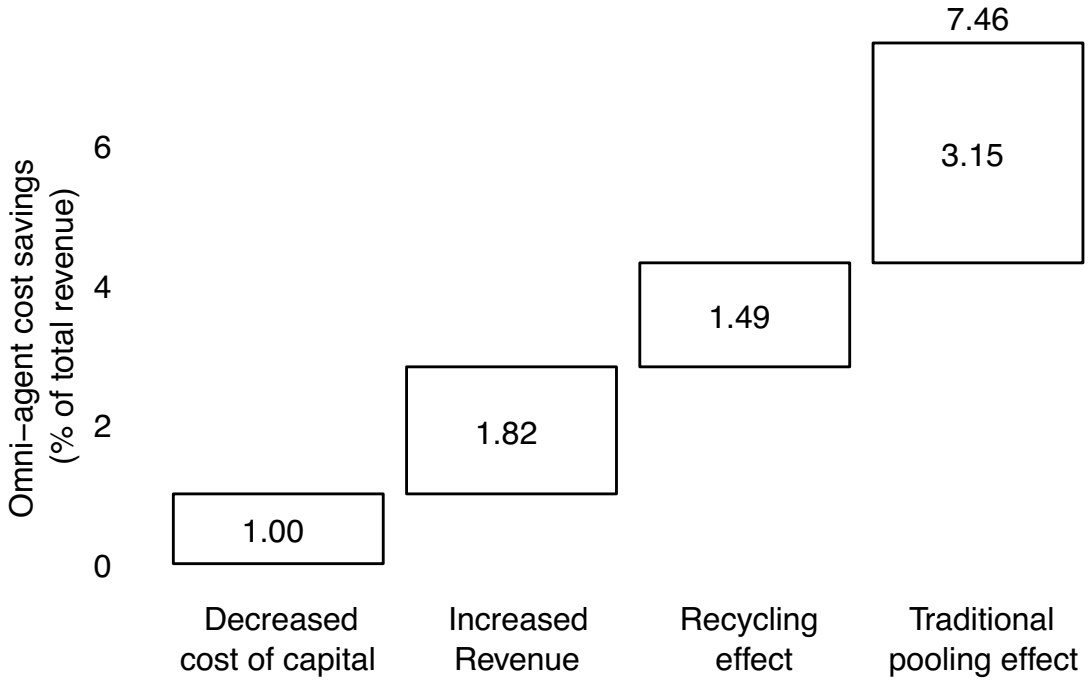


Figure A5: Net revenue improvement by type in percentage points of total possible aggregate e-float commission (omni-agent versus status quo). The omni-agent earns an additional 7.5 percentage points – over half of which stems from inventory reduction.

6 Revenue sharing pooling framework

We now relax the assumption of one omni-agent satisfying all demand centrally. We consider the simplest pooling framework that involves multiple entities – individual agents and a central e-float pool that share revenue. All agents participating in the pool carry no inventory of e-float throughout the day. When an agent is presented with e-float demand, the pool transfers the e-float to the customer, and the customer gives the agent cash equivalent to the e-float requested. The per-unit e-float commission, m_e, is then split between the pool manager and the agent. The agent receives $\alpha \cdot m_e$ and the pool manager receives $(1-\alpha) \cdot m_e$, where $0 \leq \alpha \leq 1$. When an agent is presented with cash demand, the agent gives the customer cash and the e-float generated from this transaction is remitted to the pool. Note

that this basic framework for the pool allows agents to be net lenders to the pool; this occurs when an agent has experienced greater cash demand than e-float demand at a given time in the day. The relaxation of this assumption is a key extension of this work discussed further in §7.

6.1 Pool incentive alignment

The pool's problem, though similar to the omni-agent's problem, differs in a critical way: the e-float demand satisfaction yields only $(1-\alpha)\cdot m_e$, not m_e. As such, the optimal service level (fractile) for the revenue sharing pool is less than the optimal service level for the omni-agent:

$$\left(1-\frac{\Gamma}{(1-\alpha)m_e}\right) < \left(1-\frac{\Gamma}{m_e}\right)$$

Because these fractiles are applied to the same distribution of system-wide maximum cumulative demand, the pool's net revenue maximizing starting quantity of e-float, q_P^*, is less than the omni-agent's optimal e-float, q_M^*. This is an example of "double-marginalization," resulting in a reduction of system-wide net revenue when compared to the omni-agent. Intuitively, this occurs because the pool is now only receiving part of the reward for sales, while being required to bear all of the risk. First-best stocking quantities can be induced by equating the ordering entity's fractile to the fractile of the vertically integrated firm. In newsvendor model parlance, the parameters price p, unit cost c, and salvage value v can be used to construct the fractiles. The values of these parameters for the omni-agent are shown in the first column of table A2. The omni-agent's fractile is constructed as follows:

$$\frac{p_M - c_M}{p_M - v_M} = \frac{1+m_e-(1+\Gamma)}{1+m_e-1} = 1-\frac{\Gamma}{m_e}$$

A common method to align incentives involves the use of supply-chain coordinating contracts (Pasternack 1985), wherein relevant economic parameters are altered to ensure selection of first-best stocking quantities by the ordering entity. One common coordination mechanism

is the "revenue-sharing" contract, where revenues *and* costs are shared across the chain (i.e., the ordering entity earns only part of the revenue but incurs costs that are subsidized). In this case, the ordering entity is the pool, rather than the agent. Specifically, we can alter the per-unit cost faced by the pool from c to c'_P such that:

$$1 - \frac{\Gamma}{m_e} = \frac{p_P - c'_P}{p_P - v_P} = \frac{1 + (1-\alpha)m_e - c'_P}{1 + (1-\alpha)m_e - 1} \Rightarrow c'_P = 1 + (1-\alpha)\Gamma$$

In other words, the per-unit cost of capital incurred by the pool needs to be reduced by $\alpha\Gamma$. This is intuitive and parallel to the reduction of margin from m_e to $(1-\alpha)me$, (a reduction of αm_e). This reduction in unit cost has two effects: 1) it reduces the unit underage cost and 2) it decreases the unit overage cost. The combination of these two adjustments allows for the pool's optimal service level to match the omni-agent's optimal service level to achieve first-best channel stocking. Therefore, we now have the pool's net revenue:

$$\pi_P = (1-\alpha) \cdot m_e \left(\sum_{t=1}^{N} \sum_{a \in A} D_{at}^{+} - (\hat{\Delta}_P - q_P)^{+} \right) - (1-\alpha)\Gamma q_P$$

We see that π_P is decreasing in Γ while it is increasing in e-float demand, $\sum_{t=1}^{N} \sum_{a \in A} D_{at}^{+}$. While the "correction factor" $\alpha\Gamma = c_p - c'_p$ is a variable adjustment to the pool's cost, it is not analogous to a unit cost adjustment for an individual agent. This is because the correction factor is applied to the pool's decision of q_P, not the sum of unpooled agent inventory. As demonstrated earlier, $q_P < \sum_{a \in A} q_a$. This is the source of a large amount of value in pooling. This correction factor can be passed onto and spread over many agents, who would view the cost as a fixed fee. This fixed fee – represented as C_a – can be based on an agent's characteristics, such as transaction volume and ratio of cash-in value to total CICO value.

	Omni-agent	Pool with adjusted c
Unit selling price, p	$1 + m_e$	$1 + (1 - \alpha)m_e$
Unit cost, c	$1 + \Gamma$	$1 + (1 - \alpha)\Gamma$
Unit salvage value, v	1	1
Underage	$m_e - \Gamma$	$(1 - \alpha)m_e - (1 - \alpha)\Gamma$
Overage	Γ	$(1 - \alpha)\Gamma$

Table A2: Newsvendor parameters and their values under the omni-agent and revenue-sharing pooling framework. The decrease in per-unit cost incents first-best stocking quantity.

6.2 Agent incentive compatibility

In order for any given agent to prefer joining the pool, the expected net revenue to the agent for participating in the pool must be greater than or equal to the expected net revenue for not participating in the pool. To account for the different amounts of revenue captured under the status quo versus the pooled scenario, we will use Φ_a and Φ_P to represent the fill-rates when the agent acts independently and as part of the pool, respectively. Thus, the following condition must hold in order to ensure incentive compatibility:

$$\Phi_P \cdot \alpha m_e \left(\sum_{t=1}^{N} D_{at}^{+} \right) - C_a \geq \Phi_a \cdot m_e \left(\sum_{t=1}^{N} D_{at}^{+} \right) - \gamma \cdot q_a^* \quad \Rightarrow \quad \alpha \geq \left(1 - \frac{\gamma q_a^* - C_a}{m_e \sum_{t=1}^{N} D_{at}^{+}} \right) \frac{\Phi_a}{\Phi_P}$$

Thus the agent's threshold α is decreasing in her cost of capital γ and the factor by which pooling can increase revenue $\frac{\Phi_P}{\Phi_a}$. The agent's threshold α is increasing in the fixed payment to the operator C_a, the e-float commission m_e, and her inventory turns $T_a^* = \frac{\sum_{t=1}^{N} D_{at}^{+}}{q_a^*}$.We can construct a lower bound on the threshold α by setting parameters to conservative values (favoring the status quo model). Specifically, we will set $C_a = 0$, $T_a^* = 2$, and $\frac{\Phi_P}{\Phi_a} = 1.05$. We then substitute empirically-informed parameters $\gamma = 0.0005$ (20% annualized cost of capital) and $m_e = 0.01$. We then discover that a lower-bound on the α threshold is actually quite high: $\alpha \geq 0.93$. In other words, in order to incent agents to participate in the pool, the pool must relinquish nearly all of the e-float commission to the agent.

7 Discussion and extensions

This paper describes how pooling can increase system-wide net revenue for mobile money platforms, potentially creating a "win-win-win" scenario for agents, customers, and operators. For agents, there is potential to reap increased revenue from e-float demand satisfaction by carrying no e-float inventory at all. For customers, there would be fewer confidence-degrading stockouts of e-float, because agents would have access to a significantly larger pool of inventory. Finally, operators could benefit from increased service levels, as higher reliability on the part of the platform could increase loyalty (and ultimately usage and revenue) from consumers. This paper shows how a vertically integrated pool can increase system-wide net revenue by over 8% through a combination of lower per-unit capital costs, increased revenue capture, and lower system-wide inventory requirements. The decreased system-wide inventory requirements stem from the "recycling effect" and the traditional pooling effect, which combine to deliver more overall savings than each effect individually. This 8% is in one sense an upper bound on the total system-wide improvement over agents acting optimally. But it is also important to note that managing inventory in this setting is a non-trivial challenge – to the extent that agents are currently under-stocking e-float, the pooling frameworks could increase net revenue to a larger degree.

There are several extensions to this work that are required before it is implementable in the real world. First, the pooling framework must account for regulations that ban "intermediation" of funds – in other words, borrowing from one agent a and using those borrowed funds to lend to another agent a'. This issue can be addressed by restricting the pool from borrowing from agents. Second, the incentives must not encourage agents to permanently keep borrowed e-float – i.e.,, the pooling framework must mitigate the possibility of default risk. The incentives should also encourage the most suitable agents (i.e., those agents that allow for the most efficient use of pool capital) to participate, while discouraging those who would be most damaging to this efficiency.

7.1 Pooling with borrowing restrictions

In most countries where mobile money has scaled, central banks have imposed a "no intermediation" restriction on mobile money platforms (Di Castri 2013). The practical effect of this restriction is that mobile money platforms may not "borrow" funds from one agent and lend those funds to another agent. This restriction may be prudent, as agent B may default, leaving agent A out of luck and without recourse if the pool manager cannot cover the default losses. However, the pooling frameworks presented in §5 and §6 do allow for the possibility of agents to be in a net lending position to the pool. Future work should amend these frameworks so that while each participating agent still carries no e-float, any e-float proceeds from a cash-out transaction beyond what has already been borrowed by that particular agent cannot be utilized by the pool. Agents therefore can only have a net positive (net borrowing) position from the pool. This inventory position with the pool of agent a at system-wide time t is represented as I_{at}:

$$I_{at} = \sum_{j=1}^{t} D_{aj} + \left(\sum_{j=1}^{t} D_{aj} \right)^{-} = (\Delta_{at} + \Delta_{at}^{-})$$

The total pool position (the sum of all agents' borrowing from the pool) at any given time t is thus represented as $\Delta_{Rt} = \sum_{a \in A} I_{at}$. We can continue to describe the important quantity of maximum cumulative demand with our naming convention such that $\hat{\Delta}_R = \max_{1 \leq t \leq N} \Delta_{Rt}$. Given the "no-stockout" condition, it can be seen that in order to satisfy all agent demand, the pool must carry $q_R \geq \hat{\Delta}_{Rt}$. Intuitively, this "no borrowing" restriction decreases the salience of the recycling effect, and thus yields a higher total inventory requirement than a pool unencumbered by this restriction. This gives us an initial indication that the performance of the restricted pool can be degraded by a single agent who is particularly "unbalanced" towards e-float demand (i.e., the agent sees a large amount of e-float demand without meaningful interspersed cash demand for offsetting). Analysis on how much more inventory a pool with this restriction must carry, as well as accompanying incentive issues, should be studied in

future work.

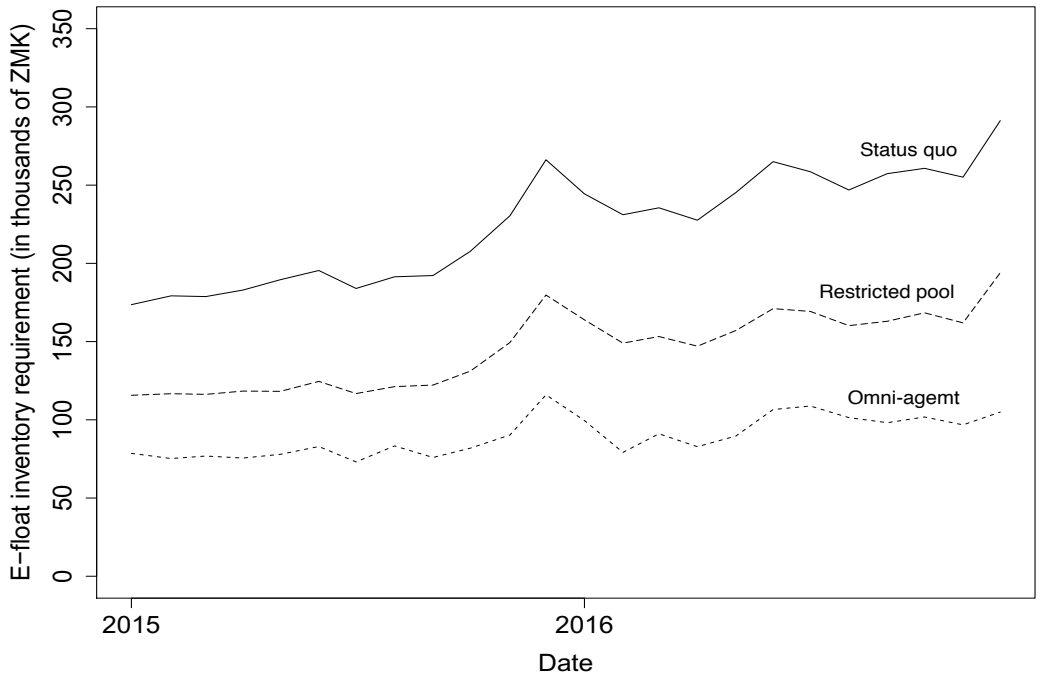


Figure A6: Hindsight inventory requirements of omni-agent, restricted pool, and status quo agents averaged over each month. The omni-agent requires 60% less inventory on average. The restricted pool captures about half of the savings.

7.2 Pooling with default risk

In the frameworks presented in this paper, the pool is effectively extending short-term credit to agents to cover e-float demand. This lending, however, involves default risk. This likelihood of default is connected to the strength of the agent's incentive to default, because there is no collateral in this arrangement. Additionally, there are often no practical or efficient ways to collect on debt from small-scale entrepreneurs through the legal system in the developing world (Armendáriz and Morduch 2010). Thus, the most practical recourse for the pool manager is to bar the agent from future participation in the pool. Therefore, to ensure agent incentive compatibility, the agent's short-term gain from defaulting on a

credit extension from the pool (the magnitude of outstanding borrowing at the end of the day, Δ^{+}_{aN}) must not exceed the agent's expected (and perhaps discounted) future benefit from participating in the pool. This calculation can be used to generate a threshold beyond which an agent is not able to borrow. From an implementation perspective, an agent can be alerted to this threshold upon approaching it, and then could be prompted to proceed to a bank or other re-balancing point in order to exchange cash for e-float to decrease the agent's liabilities with the pool. Additionally, these thresholds and fixed payments to the pool (as addressed in §6.2 can combine to form a set of incentives that could attract the agents who are most "suitable" for the pool. These most "suitable" agents are those who contribute the most to the overall sales of the system while requiring the least amount of incremental additional inventory. Future work should be focused on ensuring maximal system-wide inventory efficiency while simultaneously ensuring incentive compatibility for all players.

Conclusion

Mobile money platforms have grown at a stunning rate in the developing world. However, mobile money agents, who perform the critical functions of converting cash to electronic value and vice versa for customers, are often stocked out of cash and or electronic value. A significant barrier to the opening and operating of a mobile money agency is the high working capital requirements to finance inventories of cash and electronic value. We develop the framework for an inventory pool of electronic value that can significantly decrease the working capital burden on agents, while also increasing inventory service levels. This framework achieves these objectives by harnessing not only the power of traditional variation pooling, but also the "recycling effect" resulting from the fact that agents can remit electronic value back to the pool when they satisfy demand for cash. We test this model with a large dataset of mobile money transactions from Zambia, and show that implementing a pool can decrease system-wide inventory requirements by over 74% and increase system-wide revenue

net of cost of capital by over 8%. We also outline extensions to these models that should be developed before implementing a pooling framework in the field to ensure regulatory compliance and incentive compatibility.

Appendix

A Table of parameters

Parameter	Description
t	unit of time index
N	index of final unit of time
$a \in A$	index in set of agents in pool
γ, Γ	daily unit costs of capital for agent and pool
m_e, m_c	total per-unit commission on e-float and cash sales
q_a, s_a	starting e-float and cash inventory for agent a
D_{at}	value of demand for agent a at time t
M, P, R	indeces for omni-agent and revenue-sharing pool, and restricted pool
α	Agent share of e-float commission in revenue sharing model
Δ_{at}	cumulative demand for agent a after time t
$\hat{\Delta}_a, \check{\Delta}_a$	maximum and minimum cumulative demand for agent a
$\rho_{a,a'}$	correlation coefficient between $\hat{\Delta}_a$ and $\hat{\Delta}_{a'}$
Φ_a, Φ_P	fraction of revenue captured by status quo agent a and pool P
I_{at}	net position with the pool for agent a at time t

Table A.1: Table of parameters

B Proofs

Proof of Proposition 4: $\hat{\Delta}_{A,t} \leq \sum_{i \in A} \hat{\Delta}_{i,t}$

Definitions:

$\Delta_{A,t} = \sum_{j=1}^{t} \sum_{i \in A} D_{i,j}$ and $\Delta_{a,t} = \sum_{j=1}^{t} D_{i,j}$

$\hat{\Delta}_{A,t} = \max_{0 \leq j \leq t} \sum_{j=1}^{t} \sum_{i \in A} D_{i,j}$ and $\hat{\Delta}_{a,t} = \max_{0 \leq j \leq t} \sum_{j=1}^{t} D_{i,j}$

We assume that agents start out with zero cumulative demand, that is $\Delta_{i,0} = 0 \ \forall i \in A$. Furthermore, demand can be realized by only one agent per time period, that is $D_{i,j} * D_{k,j} = 0 \ \forall 0 < j \leq N$ and $\forall i, k \in A$, where $i \neq k$. We will prove the proposition by induction in each of four cases: A) when the next period's demand $D_{a,k+1}$ is not large enough to increase the pool's maximum cumulative demand nor increase any of the maximum

cumulative demand of the individual agents that comprise the pool, B) when $D_{a,k+1}$ is large enough to both increase the pool's maximum cumulative demand and the maximum cumulative demands of one of the individual agents, C) when $D_{a,k+1}$ is large enough to increase the pool's maximum cumulative demand but not large enough to increase any individual agent's maximum cumulative demand, and D) when $D_{a,k+1}$ is not large enough to increase the pool's maximum cumulative demand but is large enough to increase any individual agent's maximum cumulative demand. Proceeding with a proof by induction, the base case holds in all four cases, for $t = 1$:

$$D_{a1} = D_{a1}$$

Let the inductive hypothesis be:

$$\hat{\Delta}_{A,k} \leq \sum_{i \in A} \hat{\Delta}_{i,k}$$

Case A: $D_{a,k+1} \leq \hat{\Delta}_{Ak} - \Delta_{Ak}$ and $D_{a,k+1} \leq \hat{\Delta}_{ak} - \Delta_{ak}$

$$\hat{\Delta}_{A,k} \leq \sum_{i \in A} \hat{\Delta}_{i,k} \qquad \text{by the IH}$$

$$\hat{\Delta}_{A,k+1} \leq \sum_{i \in A} \hat{\Delta}_{i,k} \qquad \text{because } \hat{\Delta}_{A,k+1} = \hat{\Delta}_{A,k}$$

$$\hat{\Delta}_{A,k+1} \leq \sum_{i \in A} \hat{\Delta}_{i,k+1} \qquad \text{because } \hat{\Delta}_{a,k+1} = \hat{\Delta}_{a,k}$$

Case B: $D_{a,k+1} \leq \hat{\Delta}_{Ak} - \Delta_{Ak}$ and $\hat{\Delta}_{ak} - \Delta_{ak} < D_{a,k+1}$

$$\hat{\Delta}_{A,k} \leq \sum_{i \in A} \hat{\Delta}_{i,k} \qquad \text{by the IH}$$

$$\hat{\Delta}_{A,k} \leq \sum_{i \in A'} \hat{\Delta}_{i,k} + \hat{\Delta}_{a,k}$$

$$\hat{\Delta}_{A,k} \leq \sum_{i \in A'} \hat{\Delta}_{i,k} + \Delta_{a,k} + D_{a,k+1} \qquad \text{because } D_{a,k+1} + \Delta_{ak} > \hat{\Delta}_{ak}$$

$$\hat{\Delta}_{A,k+1} \leq \sum_{i \in A} \hat{\Delta}_{i,k+1} \qquad \hat{\Delta}_{A,k+1} = \hat{\Delta}_{A,k}$$

Case C: $\hat{\Delta}_{Ak} - \Delta_{Ak} < D_{a,k+1}$ and $D_{a,k+1} \leq \hat{\Delta}_{ak} - \Delta_{ak}$

$$\Delta_{a,k} \leq \hat{\Delta}_{a,k} - D_{a,k+1} \qquad \text{by definition of Case C}$$

$$\Delta_{A,k} \leq \hat{\Delta}_{a,k} - D_{a,k+1} + \sum_{i \in A'} \Delta_{i,k} \qquad \text{because} \sum_{i \in A'} \Delta_{i,k} + \Delta_{a,k} = \Delta_{A,k}$$

$$\hat{\Delta}_{A,k+1} \leq \hat{\Delta}_{a,k} + \sum_{i \in A'} \Delta_{i,k} \qquad \text{by definition of Case C}$$

$$\hat{\Delta}_{A,k+1} \leq \hat{\Delta}_{a,k} + \sum_{i \in A'} \hat{\Delta}_{i,k}$$

$$\hat{\Delta}_{A,k+1} \leq \sum_{i \in A} \hat{\Delta}_{i,k}$$

$$\hat{\Delta}_{A,k+1} \leq \sum_{i \in A} \hat{\Delta}_{i,k+1} \qquad \text{because } \hat{\Delta}_{a,k+1} = \hat{\Delta}_{a,k}$$

Case D: $\hat{\Delta}_{Ak} - \Delta_{Ak} < D_{a,k+1}$ and $\hat{\Delta}_{ak} - \Delta_{ak} < D_{a,k+1}$

Let A' represent the set of agents in A excluding a.

$$\hat{\Delta}_{A',k} \leq \sum_{i \in A'} \hat{\Delta}_{i,k} \qquad \text{because IH holds } \forall A' \subset A$$

$$\Delta_{A',k} \leq \sum_{i \in A'} \hat{\Delta}_{i,k} \qquad \text{because } \Delta_{A',k} \leq \hat{\Delta}_{A',k}$$

$$\Delta_{A,k} \leq \sum_{i \in A'} \hat{\Delta}_{i,k} + \Delta_{ak} \qquad \text{because } \Delta_{A,k} = \Delta_{A',k} + \Delta_{ak}$$

$$\hat{\Delta}_{A,k+1} \leq \sum_{i \in A'} \hat{\Delta}_{A',k} + \Delta_{ak} + D_{a,k+1} \qquad \text{because } \hat{\Delta}_{A,k+1} = \Delta_{A,k} + D_{a,k+1}$$

$$\hat{\Delta}_{A,k+1} \leq \sum_{a \in A} \hat{\Delta}_{A,k+1} \qquad \text{because } \hat{\Delta}_{a,k+1} = \Delta_{a,k} + D_{a,k+1} \square$$

References

Adelman, Daniel, George L Nemhauser. 1999. Price-directed control of remnant inventory systems. *Operations Research* **47**(6) 889–898.

Agrawal, Mohit. 2009. Mobile Money Transfer. URL `http://www.telecomcircle.com/2009/05/mobile-money-transfer-mmt/`.

Aker, Jenny C, Rachid Boumnijel, Amanda Mcclelland, Niall Tierney. 2011. Zap It to Me : The Short-Term Impacts of a Mobile Cash Transfer Program.

Armendáriz, Beatriz, Jonathan Morduch. 2010. *The economics of microfinance*. MIT press.

Balasubramanian, Karthik, David Drake. 2015. Mobile money: The effect of service quality and competition on demand .

Balasubramanian, Karthik, David Drake, Douglas Fearing. 2017. Inventory management for mobile money agents in the developing world .

Batista, Cátia, Pedro C Vicente. 2013. Introducing Mobile Money in Rural Mozambique: Evidence from a Field Experiment.

BCG. 2011. The Socio-Economic Impact of Mobile Financial Services. Tech. Rep. April.

Bimpikis, Kostas, Mihalis G Markakis. 2015. Inventory pooling under heavy-tailed demand. *Management Science* **62**(6) 1800–1813.

BMGF. 2012. Geospatial Analysis for Financial Inclusion Tracking. URL `http://www.gsma.com/mobilefordevelopment/geospatial-analysis-for-financial-inclusion-tracking`.

BTCA. 2014. Better Than Cash Alliance: for the Development Community. URL `http://betterthancash.org/join/for-the-development-community/`.

Buell, R. W., M. I. Norton. 2011. The Labor Illusion: How Operational Transparency Increases Perceived Value. *Management Science* **57**(9) 1564–1579. doi:10.1287/mnsc.1110.1376.

Buku, Mercy W, Michael W Meredith. 2013. Safaricom and M-PESA in Kenya: Financial Inclusion and Financial Integrity. *Washington Journal of Law, Technology, and the Arts* .

Cachon, Gérard P, Martin A Lariviere. 2005. Supply chain coordination with revenue-sharing contracts: strengths and limitations. *Management science* **51**(1) 30–44.

Cetinkaya, Sila, Chung-Yee Lee. 2000. Stock replenishment and shipment scheduling for vendor-managed inventory systems. *Management Science* **46**(2) 217–232.

Chen, Xin, David Simchi-Levi. 2009. a New Approach for the Stochastic Cash Balance Problem With Fixed Costs. *Probability in the Engineering and Informational Sciences* **23**(04) 545.

Choi, Ki-Seok, JG Dai, Jing-Sheng Song. 2004. On measuring supplier performance under vendor-managed-inventory programs in capacitated supply chains. *Manufacturing & Service Operations Management* **6**(1) 53–72.

Choi, Samuel PM, Dit-Yan Yeung, Nevin L Zhang. 2000. Hidden-mode markov decision processes for nonstationary sequential decision making. *Sequence Learning*. Springer, 264–287.

Collins, Daryl, Stuart Rutherford, Jonathon Morduch. 2009. *Portfolios of the Poor*. Princeton University Press.

Corbett, Charles J., Kumar Rajaram. 2006. A Generalization of the Inventory Pooling Effect to Nonnormal Dependent Demand. *Manufacturing & Service Operations Management* **8**(4) 351–358. URL `http://pubsonline.informs.org/doi/abs/10.1287/msom.1060.0117`.

Demirguc-kunt, Asli. 2012. Measuring Financial Inclusion The Global Findex Database.

Dhaliwal, Dan S., Oliver Zhen Li, Albert Tsang, Yong George Yang. 2011. Voluntary Nonfinancial Disclosure and the Cost of Equity Capital: The Initiation of Corporate Social Responsibility Reporting. *The Accounting Review* **86**(1) 59–100.

Di Castri, Simone. 2013. Mobile money: Enabling regulatory solutions. Tech. rep.

Economides, Nicholas. 2015. Mobile Money in Tanzania.

Economides, Nicholas, Przemyslaw Jeziorski. 2017. Mobile money in tanzania. *Marketing Science* **36**(6) 815–837.

Eijkman, Frederik, Jake Kendall, Ignacio Mas. 2009. Bridges to Cash: The Retail End of M-PESA. *SSRN Electronic Journal* doi:10.2139/ssrn.1655248. URL `http://www.ssrn.com/abstract=1655248`.

Eijkman, Frederik, Jake Kendall, Ignacio Mas. 2010. Bridges to cash: the retail end of m-pesa. *Savings and development* 219–252.

El Ghaoui, Laurent, A Nilim. 2005. Robust solutions to markov decision problems with uncertain transition matrices. *Operations Research* **53**(5).

Eppen, G. D. 1979. Note–Effects of Centralization on Expected Costs in a Multi-Location Newsboy Problem. *Management Science* **25**(5) 498–501.

Evans, David S, Alexis Pirchio. 2014. An empirical examination of why mobile money schemes ignite in some developing countries but flounder in most. *Review of Network Economics* **13**(4) 397–451.

Federgruen, a., P. Zipkin. 1984. Approximations of Dynamic, Multilocation Production and Inventory Problems. *Management Science* **30**(1) 69–84.

Fleischmann, Moritz, Jacqueline M. Bloemhof-Ruwaard, Rommert Dekker, Erwin van der Laan, Jo a.E.E. van Nunen, Luk N. Van Wassenhove. 1997. Quantitative models for reverse logistics: A review. *European Journal of Operational Research* **103**(1) 1–17.

Gamerschlag, Ramin, Klaus Möller, Frank Verbeeten. 2010. Determinants of voluntary CSR disclosure: empirical evidence from Germany. *Review of Managerial Science* **5**(2-3) 233–262.

Garbarino, Ellen, Olivia F. Lee. 2003. Dynamic pricing in internet retail: Effects on consumer trust. *Psychology and Marketing* **20**(6) 495–513.

Ghate, Archis, Robert L Smith. 2013. A linear programming approach to nonstationary infinite-horizon markov decision processes. *Operations Research* **61**(2) 413–425.

Girgis, Nadia. 1968. Management Science, **15**(3) 130–140.

Githachuri, K., M. McCaffrey, L. Anthony, A. Schiff, A.M. van Swinderen, G. A. N. Wright. 2013. Agent network accelerator survey: Tanzania country report 2013. Tech. rep., The Helix Institute of Digital Finance.

Githachuri, K., M. McCaffrey, L. Anthony, A. Schiff, A.M. van Swinderen, G. A. N. Wright.

2014. Agent network accelerator survey :Tanzania country report 2013. Tech. rep., The Helix Institute of Digital Finance.

Gold, Stefan, Rüdiger Hahn, Stefan Seuring. 2013. Sustainable supply chain management in Base of the Pyramid food projects: A path to triple bottom line approaches for multinationals? *International Business Review* **22**(5) 784–799.

Groupe Speciale Mobile Association. 2015. State of the Industry: Mobile Financial Services for the Unbanked. Tech. rep.

Groupe Speciale Mobile Association. 2018. State of the Industry: Mobile Financial Services for the Unbanked. Tech. rep.

Hausman, Warren H, Antonio Sanchez-bell. 1975. The Stochastic Cash Problem with Average Compensating-Balance Requirements. *Management Science* **21**(8) 849–857.

Helms, Brigit. 2006. Access for All. Tech. rep.

Hill, Terry. 1993. *Manufacturing strategy: the strategic management of the manufacturing function*. MacMillian.

Intermedia. 2013. Mobile Money Use, Barriers, and Opportunities. Tech. Rep. February.

International Fund For Agricultural Development. 2011. Rural Poverty Report 2011. Tech. rep.

Jack, William, Tavneet Suri. 2014. Risk Sharing and Transactions Costs: Evidence from Kenya's Mobile Money Revolution . *American Economic Review* **104**(1) 183–223.

Kalkanci, Basak, Erica Plambeck. 2012. Measurement and Improvement of Social and Environmental Performance under Voluntary versus Mandatory Disclosure .

Kendall, Jake. 2011. An Emerging Platform: From Money Transfer System to Mobile Money Ecosystem.

Kiarie, Nancy, Graham N Wright. 2017. Financial Inclusion in Action liquidity solving agents perennial problem. `http://blog.microsave.net/liquidity-solving-agents-perennial-problem/`. Accessed: 2018-03-20.

Mak, Hy, Zjm Shen. 2014. Pooling and dependence of demand and yield in multiple-location inventory systems. *Manufacturing & Service Operations ...* **16**(2) 263–269.

Mas, Ignacio. 2010. Savings for the Poor. *World Economics* **11**(4) 1–12.

Mas, Ignacio, Dan Radcliffe. 2011. Scaling mobile money. *Journal of Payments Strategy & Systems* **5**(3) 298–315.

Maurer, Bill, Taylor C Nelms, Stephen C Rea. 2013. bridges to cash: Channelling agency in mobile money. *Journal of the Royal Anthropological Institute* **19**(1) 52–74.

Mbiti, Isaac. 2011. Mobile Banking: the Impact of M-Pesa in Kenya.

Mbiti, Isaac, David N Weil. 2013. The Home Economics of E-Money: Velocity, Cash Management, and Discount Rates of M-Pesa Users. *American Economic Review* **103**(3) 369–374.

McCaffrey, M., L. Anthony, A. Schiff, K. Githachuri, G. A. N. Wright. 2014. Agent Network Accelerator Survey: Kenya Country Report. Tech. rep., The Helix Institute of Digital Finance.

Morawczynski, Olga. 2009. Poor People Using Mobile Financial Services: Observations on Customer Usage and Impact from M-PESA.

Mortimer, Julie. 2002. The effects of revenue-sharing contracts on welfare in vertically-separated markets: Evidence from the video rental industry .

Neave, Edwin H, Source Management Science, Theory Series Mar. 1970. The Stochastic Cash Balance Problem with Fixed Costs for Increases and Decreases. *Management Science* **16**(7) 472–490.

Olivares, Marcelo, Gérard P. Cachon. 2009. Competing Retailers and Inventory: An Empirical Investigation of General Motors' Dealerships in Isolated U.S. Markets. *Management Science* **55**(9) 1586–1604.

Parasuraman, A., Valerie Zeithaml. 1988. SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *Journal of Retailing* **64**(1).

Pasternack, Barry Alan. 1985. Optimal pricing and return policies for perishable commodities. *Marketing science* **4**(2) 166–176.

Porteus, Evan L, Edwin H Neave. 1972. Stochastic Cash Balance Problem with Charges Levied against the Balance. *Management Science* **18**(11) 600–602.

Prahalad, C K, Allen Hammond. 2002. Serving the Worlds Poor, Profitably. *Harvard Business Review* (September).

Quaye, Frederick Murdoch, Valentina Hartarska. 2016. Investment impact of microfinance credit in ghana. *International Journal of Economics and Finance* **8**(3) 137.

Reeves, Carol A, David A Bednar. 1994. Defining Quality: Alternatives and Implications. *Academy of Management Review* **19**(3) 419–445.

Singh, J., D. Sirdeshmukh. 2000. Agency and Trust Mechanisms in Consumer Satisfaction and Loyalty Judgments. *Journal of the Academy of Marketing Science* **28**(1) 150–167.

Sodhi, ManMohan S., Christopher S. Tang. 2011. Social enterprises as supply-chain enablers for the poor. *Socio-Economic Planning Sciences* **45**(4) 146–153.

Sun, Pi-Chuan, Chia-Min Lin. 2010. Building customer trust and loyalty: an empirical study in a retailing context. *The Service Industries Journal* **30**(9) 1439–1455.

Suri, Tavneet, William Jack. 2016. The long-run poverty and gender impacts of mobile money. *Science* **354**(6317) 1288–1292.

Tatem, Andrew. 2013. Development of Pilot High-Resolution Gridded Poverty Surfaces: Methods working paper.

Tatem, Andrew J, Abdisalan M Noor, Craig von Hagen, Antonio Di Gregorio, Simon I Hay. 2007. High resolution population maps for low income nations: combining land cover and census in East Africa. *PloS one* **2**(12) e1298.

Tayur, Sridhar, Ram Ganeshan, Michael Magazine. 2012. *Quantitative models for supply chain management*, vol. 17. Springer Science & Business Media.

Tirole, Jean. 1988. *The theory of industrial organization*. MIT press.

Toffel, Michael W, Erin M Reid. 2009. Responding To Public And Private Politics : Corporate Disclosure Of Climate Change Strategies. *Strategic Management Journal* **1178**(June) 1157–1178.

van Ryzin, Garrett J, Kalyan T Talluri. 2005. An introduction to revenue management. *Emerging Theory, Methods, and Applications*. INFORMS, 142–194.

Wright, G. A. N. 2015. A Question of Trust: Mitigating Customer Risk in Digital Financial Services. Tech. rep., The Helix Institute of Digital Finance.

Wright, Graham A N, Leonard K Mutesasira. 2001. The Relative Risks of Savings to the Poor. *Small Enterprise Development* **12**(3) 33–45.