Managing Humanitarian Operations: The Impact of Amount, Schedules, and Uncertainty in Funding

Karthik V. Natarajan

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Kenan–Flagler Business School (Operations).

Chapel Hill
2013

Approved by:

Dr. Jayashankar M. Swaminathan, Chair

Dr. Adam Mersereau, Committee Member

Dr. Dimitris Kostamis, Committee Member

Dr. Ann Marucheck, Committee Member

Dr. Vidyadhar G. Kulkarni, Committee Member

Abstract

KARTHIK V. NATARAJAN: Managing Humanitarian Operations: The Impact of
Amount, Schedules, and Uncertainty in Funding
(Under the direction of Dr. Jayashankar M. Swaminathan)

Global health spending has increased manyfold in the last few decades reaching US$6.5 trillion in 2012. Despite these increases, humanitarian organizations from around the world, working on different diseases including Malaria and Tuberculosis, have warned about potential funding shortfalls in the near future. Facing a growing need for health services and commodities, resource–constrained organizations are constantly looking for ways to maximize health outcomes through efficient and effective use of available resources. In this dissertation, we develop approaches to make efficient operational decisions under variable and unpredictable donor funding, a situation that is commonly faced by many humanitarian organizations. In the first chapter, we study the problem of managing inventory of a nutritional product under variable funding constraints. Despite the complexities associated with funding, we show that the optimal replenishment policy is easy to compute and straightforward to implement. We also provide several insights into how the funding amount, funding schedules and uncertainty in funding impact operating costs in this setting. In chapter 2, we look at the problem of dynamically allocating a limited amount of donor funding to patients in different health states in a humanitarian health setting. We show that the optimal allocation policy is state–dependent and prove several structural properties of the optimal policy that would help simplify its computation. Due to the complexity involved in calculating the optimal policy, we develop two heuristics to handle real–size problems with longer planning horizons. Computational results suggest that both heuristics perform well in many cases but one of the heuristics is more robust across a wide variety of settings. In addition to the allocation policy, we also provide some interesting insights into the impact of funding level and funding uncertainty in the multiple health states setting.

In the third chapter, we focus on the supply– vs. demand–side investment dilemma frequently faced by public health managers who have a limited budget at their disposal. First, we consider a centralized setting where a single entity, referred as the principal, makes both supply– and demand–side investment decisions. We determine the principal's optimal investment mix in this budget constrained environment and provide insights into how the investment mix varies with the different supply– and demand–side parameters. We then consider a decentralized setting where the principal invests in improving the supply chain while demand mobilization activities are contracted to an agent, who is a profit maximizer. For the decentralized setting, we identify two contracts that ensure that the coverage in the decentralized setting is at least as high as the centralized case.

Acknowledgements

This dissertation would not have been possible without the support, advice and encouragement provided by faculty members at the Kenan-Flagler Business School, friends and colleagues both within and outside the school, and my family members.

My heartfelt thanks goes to my advisor, Dr. Jayashankar M. Swaminathan, whose guidance and mentorship has made the last few years a very rewarding experience. Working with Jay has developed my critical thinking and taught me the importance of looking at the big picture when working on different projects. His passion and commitment to research continues to amaze me and he has shown me, by example, the value of perseverance and a strong work ethic that are essential qualities to be a successful researcher. In many ways, he is my role model and I feel absolutely privileged to have had the opportunity to work with him. Jay, thank you.

My special thanks also go to Dr. Adam Mersereau and Dr. Dimitris Kostamis for agreeing to serve on my committee, and for their support over the years, beginning in the first year of the doctoral program. I am very grateful for their patience and guidance while working on my summer paper, and I hope that I will have the opportunity to collaborate with them again at some point in the near future. Adam, a special thank you to your family for inviting us to the annual Thanksgiving lunch. We very much enjoyed it.

I am greatly indebted to Dr. Ann Marucheck for serving on my committee, and for her outstanding support and encouragement throughout the years I have been in the doctoral program. Ann is always keen to offer professional and personal support and help students in any way she can throughout the dissertation process. I would also like to express my gratitude to Dr. Vidyadhar Kulkarni for serving on my dissertation committee and providing valuable inputs to improve my dissertation. I also very much enjoyed taking his "Stochastic Processes"

class in my first year in the Ph.D. program.

I would like to thank my senior colleagues — Vidya Mani, Yen-Ting Lin, Gokce Esenduran, Olga Perdikaki and Sriram Narayanan for their help and advice regarding navigating the different stages of the program. I will always cherish the friendship of Aaron Ratcliffe, Adem Orsdemir and Gang Wang, and I very much enjoyed the McColl Cafe and patio lunches we had together over the years. I would like to wish the upcoming doctoral students — Zhe Wang, Hyun Seok Lee and Ying Zhang the very best in successfully completing the program, and I look forward to hearing about their graduations. I would also like to extend my thanks to Virgnia Kay, Jyotishka Ray, Valmik Khadke, Kevin Miceli and Chang Hyun Kim for their friendship and support, and sharing many light moments.

Above all, I would like to take a moment to say special thanks to my parents Mr. Natarajan Veeraswami and Mrs. Bhagirathi Natarajan, my sister Mrs. Priya Suresh and my wonderful fiancee Ms. Kadambari Chandra Shekar. My parents have always been a source of inspiration for me, supporting me through tough times, and have always been there for me despite the fact that we have not been together for many years now. My sister deserves my sincere thanks for her love and affection and providing me the comforting feeling that she is there to take care of my parents in my absence. Kadambari, I cannot thank you enough for all that you have done and for enduring the long periods of absence in the final stages of writing this dissertation (and even before). You make my day better everyday and I am already looking forward to spending many more great years with you. Mom and dad, Priya, and Kadambari, I want to share the joy of completing this dissertation with you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Against the backdrop of escalating costs, growing demand for health commodities, and tightening budgets, humanitarian organizations have come under increasing pressure to target aid financing effectively and justify the enormous levels of aid spending. Consequently, there is a growing interest within the humanitarian community to look at ways to improve the efficacy and efficiency of their operations. Although global health spending has increased tremendously in the last decade, many organizations like The Global Fund to Fight AIDS, Tuberculosis and Malaria, the Global Alliance for Vaccines and Immunization (GAVI Alliance) and the World Health Organization (WHO) face serious funding shortfalls. This situation is expected to further deteriorate in the current economic climate with many governments, including the U.S. and Spain, cutting down their foreign aid. In addition, global health funding is highly volatile and unpredictable — this is partially driven by the budget and funding cycles of the donors, some aspects of which are often outside their control.

In light of the above mentioned problems with global health funding, managing operations in the humanitarian health sector is a complex task. Currently, there is lack of knowledge, training and oversight required to operate humanitarian supply chains, resulting in significant inefficiencies — the WHO estimates the associated losses to be between 20–40% of the total health funding. The first two chapters of this dissertation are aimed at bridging this knowledge gap by (i) determining ways to make effective and efficient operational decisions, given the inherent uncertainties associated with global health funding, and (ii) shedding light onto the impact of funding uncertainty and variability in humanitarian operations. The motivation for the problems studied in the first two chapters comes from the ready–to–use–therapeutic–food (RUTF) supply chain in the Horn of Africa but the problems we study and the underlying issues are also relevant to many other humanitarian organizations operating in the global health sector.

In the first chapter, we study the problem of managing inventory of a health commodity in the presence of funding constraints over a finite planning horizon. Funding from donors, which finances the procurement, is received in installments throughout the planning period with uncertainty around both the timing and amount received. In the problem that motivated this study, the country office of UNICEF, which managed the RUTF procurement, was constrained by the timing of receipt of previously promised funding from donors. Given the highly variable and unpredictable nature of the inflow of funds, we (i) identify the optimal procurement strategy taking into account the current financial position as well as the funding due to arrive in the future, and (ii) quantify the impact of funding timing, funding level and funding uncertainty on the operating costs.

We model this problem using a stylized multi–period inventory model with financial constraints. Among other results, we show that a capital–dependent modified base stock policy is optimal. Remarkably, we are able to show that the capital–dependent modified base stock policy can be simplified to a state independent policy, which greatly simplifies the implementation of the optimal policy and enhances its appeal. We also prove several analytical results regarding the impact of funding uncertainty and increased variability in the funding timing.

Through an extensive numerical study, we also capture the magnitude of impact of both funding amount and funding timing on operating costs. Among other results, we provide the following insights. (1) When there is no uncertainty with respect to funding timing, avoiding delays in funding should be one of the top priorities. We find that, contrary to popular belief, receiving funding in equal installments is not the optimal funding pattern due to its inability to accommodate large demand surges upfront; (2) An all–out effort focusing solely on raising as much capital as possible is not likely to be very effective — raising sufficient capital is important but so is the funding timing. Our results indicate that even with less overall funding, performance may be better if the funding is received earlier or more steadily. (3) There is a nonlinear increase in costs with increased uncertainty in funding timing but this effect is moderated by the uncertainty in demand.

A key message of this research is that humanitarian supply chain managers need to consider the funding level as well as the operating environment while undertaking initiatives to improve the funding situation.

In the first chapter, one simplifying assumption we make is that a single dose/unit of the product suffices to meet the needs of a patient/customer. This assumption makes the model general enough to be applicable to a wide variety of products — e.g., malaria bed nets and reproductive health supplies like contraceptives — but a multi-dose framework is more appropriate in certain other settings. For example, in the outpatient treatment of severely malnourished children, the affected children are given Plumpy'Nut for several weeks until they are deemed fit. This multi-dose framework is also common in the treatment/prevention of many other diseases including certain types of vaccines.

In Chapter 2, we extend the work in the first chapter by relaxing the single dose assumption, and allow for the possibility that patients enrolled in a humanitarian health program could be in different health states requiring treatment over different lengths of time (corresponding to different amounts/doses of the product) before they are completely cured. The treatment duration and the response to treatment or non–treatment could also vary depending on the health state. In this setting, the problem of dynamic allocation of a scarce resource, which in our case is funding, between patients in different health states assumes significance. Given that the total available funding is limited and funding inflow is unpredictable, in certain situations, it might be beneficial to reserve a portion of the funding available on–hand to serve more severe patients in the future periods.

Using a two health states model, we study the problem of dynamic allocation of a limited amount of donor funding to patients in different health states over a finite horizon, with the objective of minimizing the number of disease–adjusted life months lost. Of the two health states, one is assumed to be less severe and the other one is more severe. Funding is received in installments throughout the planning period with uncertainty around the timing as well as how much funding is received in each installment. New patients of both health states enter the program in every period and we also make certain assumptions regarding how patients in the two health states respond to treatment and non–treatment in every period.

We use a multiperiod stochastic dynamic programming framework with health–state dependent per–period and terminal penalty costs to analyze the allocation problem. We characterize the optimal policy to be a state–dependent policy and we prove several monotonicity properties that could help reduce the computational burden involved in determining the optimal policy. Despite the potential simplifications offered by the monotonicity results, determining the optimal policy is time–intensive for problems with long planning horizons. So we develop two heuristics (referred as PNS and FCFS heuristics) that can solve real–size problems fairly quickly. Our computational results suggest that the PNS heuristic performs well in terms of the solution quality and running time across a wide range of scenarios. The FCFS heuristic also performs well in many cases but it is less robust than the PNS heuristic and in certain settings, there is a noticeable performance gap between the two heuristics.

Our computational study also provides several insights regarding the impact of funding level and uncertainty in funding. We find that the impact of uncertainty in funding timing could be very different depending on the length of the planning horizon and the system funding level. For short–planning horizons, uncertainty in funding leads to a loss of disease–adjusted life months while in case of longer planning horizons, receiving the funding in fewer, lumpy installments involving more uncertainty in funding timing might be preferable only in under–financed systems ($<$ 100% funding level). In well–funded systems ($\geq$ 100% funding level), having a smooth and predictable funding pattern is always preferred. This finding highlights the importance of taking into account the system characteristics when making funding–related decisions so as to maximize the per–dollar impact of funding provided to global health programs. Regarding the system funding level, we find that low funding availability leads to a significant loss of disease–adjusted life months in under–financed systems. Hence, it might be beneficial to receive additional funding even at the cost of increased funding uncertainty. In well–funded systems, the losses from the increased funding uncertainty outweigh the potential benefits from the additional funding, and hence less overall funding should not be traded for larger but more uncertain funding.

The third chapter of this dissertation focuses on the the supply– vs. demand–side investment dilemma faced frequently faced by in–country public health managers. Low aid effectiveness and poor coverage have often been attributed to a combination of failures on both the demand and supply side. Supply–side factors determine the availability of essential health supplies and services when and where they are needed, and supply–side investments are geared towards reducing or eliminating supply chain inefficiencies that could lower product availability. Demand–side factors focus on the consumer, and investments on the demand–side are aimed at mobilizing demand for the service or product by increasing community awareness and eliminating or reducing the social, economic and cultural barriers to access. While countries would ideally like to invest as much as possible on both sides, oftentimes they only have a limited budget to spend on interventions aimed at improving health outcomes. In light of the limited funding availability, choosing the right investment mix is critical since too much or too little supply could significantly affect program coverage.

We address this question in the third chapter using a simple one–period model with stochastic demand. We assume that the fraction of the procured quantity available to meet demand increases linearly in the investments in the supply chain, and the deterministic part of the demand is also assumed to increase linearly with the demand–side investment. We first consider a centralized model where a single entity (e.g., Ministry of Health at a host government) that manages the health program makes both supply– and demand–side investments. The centralized setting helps us to understand the effect of the operating environment on the optimal investment mix, and it also acts as a benchmark for the decentralized case that we consider later in the chapter. For a given budget level, we identify the optimal mix of investments to maximize coverage and also analyze how the investment mix changes with respect to the different supply– and demand side parameters. We find that the optimal investment strategy is threshold type in both supply– and demand–side investment effectiveness. Some interesting insights emerge when we consider the impact of expected demand and variability in the demand. Our analysis shows that investment in demand mobilization activities need not necessarily go down in anticipation of a higher expected demand. With respect to demand variability, we prove that demand–side investment may increase or decrease in response to increased variability, depending on whether or not a critical ratio that we identify in the chapter is below a threshold point. In addition to the investment strategies, we also analyze the impact of mean demand and demand variability on the program coverage. We find that higher mean demand may not necessarily imply higher coverage but a higher variability in the demand always leads to a lower coverage for many of the commonly used demand distributions.

In the second part of the third chapter, we consider a decentralized setting where the principal, the decision maker in the centralized setting, is not physically present on ground, and as such, cannot directly engage in demand–related activities. In this case, the principal invests only in the supply–side and contracts with a third–party, who we refer to as the 'agent', to carry out community mobilization efforts on her behalf. Contracting with third–parties to carry out certain services or activities occurs frequently in the public health sector for several reasons including a low efficiency of public health systems, lack of expertise, and human resource constraints. In the decentralized case, the agent makes the demand–side investments but his objective is to maximize profits, which creates incentive issues that could lower coverage levels. Motivated by the growing interest in performance–based funding in the humanitarian sector, we explore the use of contracts that depend only on the program coverage to create incentives for the agent to invest in demand mobilization. We identify two types of contracts that guarantee that the expected coverage level under the decentralized case is at least as high as the centralized case but one of them has some important advantages over the other from an implementation standpoint.

Chapter 2

Inventory Management in Humanitarian Operations

2.1 Introduction

Financial flows play an important role in humanitarian operations and impact their scope, effectiveness and efficiency. While the total amount of donations received can impact the efficacy of such operations, the timing, predictability, and flexibility of usage around those funds also have a strong influence (Wakolbinger and Toyasaki 2011). In a global health context, unpredictability and delays in donor funding are often cited as the reasons behind impaired supply chain management and reduced coverage (Fininnov 2011). A recent study by the Brookings Institution (Lane and Glassman 2008) estimates that for every dollar received in funding, 7 to 28 cents is lost due to funding delays.

Innovative financing mechanisms reduce funding delays and improve predictability in humanitarian operations. For example, IFFIm that supports the GAVI Alliance programs, issues bonds against long–term pledges from donors to convert the pledges into "readily–usable" funds in an effort to achieve front–loading, i.e., push forward the time of receipt of funding for the program. While there is a general consensus that innovative financing mechanisms hold the key to increasing aid–effectiveness, they come at a cost. Moreover, the magnitude of the benefits from front–loading and reducing the uncertainty around the funding timing are not known. With too much front–loading, the costs might outweigh the benefits while very little front–loading could lead to a potential loss of benefits. Celasun and Walliser (2007) make a related observation: "Although predictability has been highlighted as a key issue for aid effectiveness, little systematic information is available on the magnitude of the predictability problem and thus its potential impact on aid recipients."

Motivated by the ready–to–use therapeutic food (RUTF) supply chain in Africa, we study the problem of managing inventory of a nutritional product in the presence of funding constraints over a finite horizon (Swaminathan 2009, 2010). Funding is received in installments throughout the planning period with uncertainty around the timing and amount. This unpredictable nature of donor funding is typical of the funding situation in many global health programs. Even in the best of situations there might be uncertainty in terms of timing of the receipt of those installments. In the problem that motivated this study, the country office of UNICEF that had to procure RUTF were constrained by the timing of the receipt of previously promised funding from donor agencies. Given the highly variable and unpredictable nature of the inflow of funds, an important question that arises is how to effectively manage inventory, taking into account both the current financial position as well as the funds that are due to arrive in the future periods.

In this chapter, we model the above situation with a multi–period stochastic inventory model with financial constraints. Our goal is to (i) determine the optimal ordering policy given the complexities associated with the funding and (ii) characterize the impact of funding timing, funding level, and funding uncertainty on the operating costs. Among other results, we show that despite the uncertainty in funding, the optimal replenishment policy is a state–independent modified base stock policy, which greatly enhances the appeal and implementation of the optimal policy. We also prove that uncertainty in funding timing (in comparison to a deterministic financial schedule) increases operating costs and so does the stochastically dominated late arrival of funds. Finally, we also show that increased variability in the funding timing (as measured by convex ordering) leads to higher costs.

Our work explicitly captures the impact of both funding amount and funding timing on operating costs. Through an extensive numerical study, we offer insights into several issues including (1) the impact of funding patterns on the operating costs by comparing different types of funding patterns ranging from front–loaded funding (where a majority of the total funding is received in the initial periods) to evenly–spread funding (equal installments) to back–loaded funding (where a major chunk of the total funding is received in the later periods), (2) the interaction between funding level (total funding received as a % of the amount required to meet total expected demand) and funding pattern, (3) the effect of uncertainty in funding timing and, (4) the relative importance of funding level and pattern vis a vis funding uncertainty.

Among other results, we provide the following computational insights. (1) Avoiding funding delays should be a top priority for humanitarian supply chain managers. In case of deterministic funding schedules, while evenly–spread funding facilitates planning, it is not the optimal funding pattern due to its inability to accommodate large demand surges upfront; (2) Front–loading the funding brings significant benefits in under–financed systems ($<$ 100% funding level) while avoiding back–loading is critical in fully–financed systems (100% funding level). (3) Front–loaded funding at 75% funding level is better than back–loaded funding at 100% funding level. Depending on the level of front– and back–loading at 75% and 100% funding levels, the operating costs under back–loaded funding at 100% funding level vary between 1.5–5.5 times the operating costs under front–loaded funding at 75% funding level; (4) There is a non-linear increase in costs with increased uncertainty in funding. Further, this effect decreases with demand uncertainty.

2.1.1 Literature Review

Our work is related to three streams of literature.

The Interface between Operations and Finance: In this stream of literature, a firm's available capital at the start of any given period is endogenously dependent on the revenues generated in the previous period. Papers that study the interaction between operational and financial decisions include Archibald et al. (2002), Babich and Sobel (2004), Xu and Birge (2004), Gaur and Seshadri (2005) and Chao et al. (2008). Our work is more closely related to Chao et al. (2008). They study a periodic–review inventory replenishment problem faced by a self–financing retailer whose objective is to maximize the terminal wealth at the end of the planning horizon. For their model, they show that a capital–dependent base stock policy is optimal. One aspect that distinguishes our work from the existing literature is the presence of a donor funding stream that is exogenous to realized demand. In our work, demand fulfilled in the previous periods does not generate any revenue due to the non–profit nature of the business.

Inventory Management under Capacity Constraints: Financial constraints on procurement are somewhat similar to supply–capacity constraints and a lot of work has been done on inventory management under capacity constraints (see Federgruen and Zipkin 1986 a,b, Kapuscinski and Tayur 1998, Aviv and Federgruen 1997, Ciarallo et al. 1994, Wang and Gerchak 1996). The main difference between these models and ours is that while physical capacity constraints are rigid, financial constraints can be made flexible. Unlike production capacity, unused capital does not go waste and can be utilized in the future periods.

Humanitarian Operations: Our paper also contributes to the growing body of work on humanitarian operations. Within humanitarian operations, a majority of the papers focus on disaster relief, e.g., Duran et al. (2011) and Beamon and Kotleba (2006) focus on inventory management during emergencies. In recent years, long-term public health issues have also received attention from the operations management community, e.g., Taylor and Yadav (2011), Rashkova et al. (2011) and Deo and Sohoni (2011). However, we believe that ours is the first work to look at inventory management from a humanitarian health perspective. In addition, to the best of our knowledge, Rashkova et al. (2011) and our work are the only ones to focus explicitly on the role of funding in humanitarian operations.

The rest of the chapter is organized as follows: In section 2, we introduce the model and provide analytical results regarding the optimal policy and the impact of funding. In section 3, we present the results of our numerical study and discuss the impact of funding on operating costs. In section 4, we demonstrate that the inventory replenishment policy established in section 2.2 is optimal under more general settings. We offer some managerial insights and conclude the chapter in the last section.

2.2 Model

We consider a finite–horizon, periodic–review inventory model for a single product. The planning horizon is divided into N periods with the time indexing done in the reverse order, i.e., the first period is period N, followed by N-1, N-2 and so on. Demands in successive periods, ζ_t, $t = N, N-1, ..., 1$ are independent but not necessarily identically distributed with probability distribution function f_t and cumulative density function F_t. The system is financed through external funding received in $m \leq N$ installments. The installment sizes (amount received in each installment) are known beforehand but the time of receipt of each installment could be uncertain. We refer to a funding scenario with uncertain funding timing as a *stochastic funding schedule*. A special case of the *stochastic funding schedule* is a scenario where both the installment amounts and timing are known beforehand. We refer to this special case scenario as a *deterministic funding schedule*.

We denote the funding vector by $Z = (z_1, z_2, z_3, ..., z_m)$ where z_m and z_1 are the first and last installments received respectively. Therefore, the total funding received over the planning horizon is $\sum_{j=1}^{m} z_j$. For our analysis, we do not impose any restrictions on the installment sizes — the amount received in each installment could be very different from one another. Let c be the unit purchase cost, h denote the unit holding cost per period and b be the penalty cost per unit per period for any unsatisfied demand. We make the following assumptions in our model.

1. Unmet demand is completely backlogged. While this is an approximation in the RUTF context, the backlogging assumption is valid for a variety of health commodities, e.g., malaria bed nets and reproductive health supplies like contraceptives.

2. All the installments are received before the end of the planning horizon. As we mentioned before, donors make a commitment based on the funding proposals and while the amount in the individual installments may vary based on the donors' budget cycles, in most cases, the committed amount is received in full before the end of the planning period for which the donation was sought.

3. Borrowing capital is not an option. In the application that motivated this study, country offices place procurement orders only when they have raised enough money from the donors to fund the procurement. Currently, they neither borrow money to finance the procurement nor do they have access to a credit line.

In addition to the above–mentioned assumptions, we also assume that one dose/unit of the product is sufficient to meet the needs of a customer/patient. Again, this is a simplifying assumption in the context of the problem that motivated this study but it makes the model general enough to be applicable to a wide variety of humanitarian health programs.

The sequence of events in any given period t is as follows: 1. Funding (if any) is received at the start of period t. 2. Procurement decisions are made, subject to capital available on–hand. We assume that replenishments arrive instantaneously. (This assumption is only for simplicity and in section 4, we demonstrate that relaxing this assumption does not change the structure of the optimal replenishment policy.) 3. Finally, demand is realized and holding and backorder costs are calculated based on the ending inventory.

Let x_t, y_t denote the on–hand inventory before and after ordering in period t and r_t be the capital available at the start of period t, after receiving installments (if any) in period t. The state variable O_t keeps track of the number of outstanding installments as of the beginning of period t, after receiving installments (if any) in that period. The recursive equation for r_t is thus given by $r_t = r_{t+1} - c(y_{t+1} - x_{t+1}) + \sum_{j=O_t+1}^{O_{t+1}} z_j$. We let $P_t(i)$ denote a random variable corresponding to the number of outstanding installments at the beginning of period $t-1$, given that the number of outstanding installments at the beginning of period t is i. $P_{N+1}(m)$ is the number of outstanding installments at the beginning of period N. Also, define $P_t(i,j) \equiv Pr(P_t(i) = j)$.

The objective is to come up with an optimal ordering policy that minimizes the total cost incurred over the planning horizon subject to the funding constraints. Since we assume self–financing, the order quantity in period t must satisfy the constraint $c(y_t - x_t) \leq r_t$. Let $J_t\left(x_t, r_t, O_t | Z_t\right)$ be the minimum expected cost with t periods to go, given state variables x_t, r_t, O_t and the future funding vector Z_t. Given $O_t = k$ for some constant k, $Z_t = (z_1, ..., z_k, 0, 0..., 0)$ where the last $m-k$ components of Z_t are zeroes. For brevity, we will write the conditioning on Z_t explicitly only when we are comparing two different funding vectors.

Then, for a fixed funding vector Z, the optimality equations are given by

$$J_t(x_t, r_t, O_t) = \min_{y_t \in \left[x_t, x_t + \frac{r_t}{c}\right]} \left\{ \begin{array}{l} c(y_t - x_t) + b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+ + h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+ \\ +\mathsf{E}_{O_{t-1}}\mathsf{E}_{\zeta_t} J_{t-1}(y_t - \zeta_t, r_t - c(y_t - x_t) + \sum_{j=O_{t-1}+1}^{O_t} z_j, O_{t-1}) \end{array} \right\} \tag{2.1}$$

Since all the installments are received before the end of the horizon, O_1=0 always. The terminal cost is $J_0(x_0, r_0, 0) = 0 \; \forall (x_0, r_0)$.

Some intuitive properties of the cost–to–go function can be readily proven. For example, given x_t, O_t and funding vector Z_t, $J_t(x_t, r_t, O_t)$ is monotone decreasing in r_t. Additionally, if $O_t = k$ for some constant k, $z_i^1 = z_i^2 \; \forall \; i = 1, 2, ..k-1$ and $z_k^2 - z_k^1 = r_t^1 - r_t^2 = K \geq 0$, then $J_t\left(x_t, r_t^2, O_t | Z_t^2\right) \geq J_t\left(x_t, r_t^1, O_t | Z_t^1\right)$, i.e., it is more valuable to have an extra dollar today than receiving it in the next installment.

Our first key result is the joint convexity of the value function in state variables x_t and r_t for fixed O_t and Z_t. We prove this in Lemma 1. The proofs for all the results in this chapter can be found in Appendix A.

Lemma 1. *$J_t(x_t, r_t, O_t)$ is jointly convex in x_t and r_t for fixed O_t and funding stream Z_t.*

In our analysis, we find it convenient to use a modified value function expressed in terms of variables x_t and $R_t = r_t + cx_t$. Define

$$\begin{aligned} \hat{C}_t(y_t, R_t, O_t) = cy_t + b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+ + h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+ \\ + \mathsf{E}_{O_{t-1}}\mathsf{E}_{\zeta_t} J_{t-1}(y_t - \zeta_t, R_t - cy_t + \sum_{j=O_{t-1}+1}^{O_t} z_j, O_{t-1}) \end{aligned} \tag{2.2}$$

Then, in terms of $\hat{C}_t$, we have

$$(P1) \quad \tilde{J}_t(x_t, R_t, O_t) = J_t(x_t, r_t, O_t) = -cx_t + \min_{y_t \in \left[x_t, \frac{R_t}{c}\right]} \left\{ \hat{C}_t(y_t, R_t, O_t) \right\} \tag{2.3}$$

Using the modified value function $\tilde{J}_t$, it is straightforward to show that a (R_t, O_t)–dependent modified base stock policy is optimal in period t. However, we go one step further and demonstrate that the optimal replenishment policy is actually simpler — the optimal policy is a state–independent modified base stock policy where the order up–to levels depend only on t and not on R_t, O_t or the future funding stream Z_t.

To be precise, consider a multi–period inventory management problem with the same cost and demand parameters as problem P1 (see equation (2.3)), but no financial constraints. Let $NV_t(x_t)$ be the minimum expected cost with t periods to go, corresponding to this setting with no financial constraints. Then,

$$\text{(P2)}\quad NV_t(x_t) = \min_{y_t \geq x_t} \left\{ \; c(y_t - x_t) + b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+ + h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+ + \mathsf{E}_{\zeta_t} NV_{t-1}(y_t - \zeta_t) \; \right\}$$

Let $NV_0(x_0) = 0 \ \forall \ x_0$. It is well known that a base stock policy is optimal for problem P2 and there exists an optimal base stock level y_t^* in each period such that if the inventory level in period t is below y_t^*, it is optimal to order up–to y_t^*, and not order otherwise. In Theorem 1, we prove that the unconstrained base stock levels $y_t^*, y_{t-1}^*, ..., y_1^*$, optimal for problem P2, are optimal for problem P1 with funding constraints as well.

Theorem 1. *Let $y_t^*, y_{t-1}^*, ..., y_1^*$ be the optimal base stock levels corresponding to problem P2. Let (x_t, R_t, O_t) be the state of the system in problem P1 at the beginning of period t. Then, the optimal ordering policy for problem P1 has the following simple structure.*

$$\begin{array}{lr} \textit{Order up–to } R_t/c & \textit{if } R_t/c \leq y_t^* \\ \textit{Order up–to } y_t^* & \textit{if } R_t/c > y_t^*, x_t < y_t^* \textit{ and} \\ \textit{Do not order} & \textit{if } x_t \geq y_t^*. \end{array} \tag{2.4}$$

Theorem 1, which demonstrates the optimality of a state-independent base–stock policy for a stochastic funding schedule, raises the question: why are the base–stock levels independent of state variables O_t and R_t? To answer this question, recall that O_t, which keeps track of the number of outstanding installments, determines both the total outstanding amount and the level of uncertainty in the future funding. However, the uncertainty in the future funding does not impact the total demand that can be met between period t and the end of the planning period, since unsatisfied demand is completely backlogged. How much demand is met between period t and the end of the horizon is solely determined by the sum: c×(on–hand inventory) + capital available on–hand + future funding. A closer look reveals that this sum is also independent of O_t. Hence, while the actual costs incurred might vary depending on the number of outstanding installments O_t, the incremental difference in costs obtained by changing the order quantity remains the same, irrespective of how many installments are outstanding.

A related question is why is the base–stock level independent of R_t? Recall that $R_t(=cx_t+r_t)$ determines the maximum inventory level attainable in period t. From a single–period perspective, R_t acts like a production capacity constraint. In the presence of capacity constraints, it is well known that the (capacity–dependent) order up–to levels in each period are higher than the corresponding unconstrained base stock levels (see e.g., Federgruen and Zipkin 1986 a,b). So why is the optimal policy different for our problem with funding constraints ? The key reason is that, although both funding and capacity constraints place an upper bound on the order quantity in each period, there is a fundamental difference between the two. While unused capacity goes waste, unused capital can be used in the later periods, i.e., it acts like transferable capacity. Intuitively, this is the reason why the unconstrained base stock levels continue to be optimal even in the presence of funding constraints. This greatly enhances the appeal and implementation of the optimal policy since the state–independent base stock levels can be easily computed using techniques like IPA (see Glasserman and Tayur 1995). Additionally, the fact that the target inventory level is not tied to the future inflow of funds also makes operations planning easier since purchasing decisions depend only on the current state of the system and not on any future unobservable quantities.

Thus far, by identifying the optimal replenishment policy for any given funding scenario, we have answered the first of our two main research questions: how to efficiently manage inventory given the complexities associated with the funding? We now proceed to answer the second question: how does the funding amount, funding schedule, and the uncertainty around the funding timing impact operational performance? As a first step, we focus on understanding the role of the uncertainty around the time of receipt of the installments.

2.2.1 Impact of Funding Timing

Throughout this section, we fix the funding vector $Z = (z_1, z_2, ..., z_m)$ and vary only the time of receipt of each installment. Recall that $P_t(i)$ denotes a random variable corresponding to the number of outstanding installments at the beginning of period $t-1$, given that the number of outstanding installments at the beginning of period t is i. Consider random variables $\left\{P_t^1(i), i \in \{0,1,2,...,m\}\right\}$ and $\left\{P_t^2(i), i \in \{0,1,2,...,m\}\right\}$ corresponding to funding scenarios 1 and 2 such that

$$P_t^2(i) \geq_{st} P_t^1(i) \; \forall \; i \in \{0,1,2,...,m\} \text{ and } \forall \; t \in \{2,3,...,N\} \text{ and} \tag{2.5}$$

$$P_t^n(i') \geq_{st} P_t^n(i) \; \forall \; i \in \{0,1,2,...,m-1\}, i' > i, n \in \{1,2\} \text{ and } \forall \; t = 2,...,N \tag{2.6}$$

where $\geq_{st}$ means first–order stochastic dominance. The following is the definition of first–order stochastic dominance.

Definition 1. *(Shaked and Shanthikumar 2007) Let X and Y be two random variables such that $P(X > x) \leq P(Y > x)$ for all $x \in (-\infty, \infty)$. Then X is said to be smaller than Y in the usual stochastic order.*

Condition (2.5) implies that the number of outstanding installments at the beginning of any period t is (stochastically) larger under funding scenario 2. Condition (2.6) says that, under both funding scenarios, the number of outstanding installments at the beginning of $t-1$ stochastically increases in the number of outstanding installments at the beginning of period t. We denote the value functions associated with random variables P_t^1 and P_t^2 by J_t^1 and J_t^2 respectively. In the following theorem, we demonstrate that the (stochastically) early arrival of funds offers increased procurement flexibility, resulting in lower operating costs.

Theorem 2. *If conditions (2.5) and (2.6) hold, then $J_t^2(x_t, r_t, j) \geq J_t^1(x_t, r_t, j)$ for every $j \in \{0,1,2,...,m\}$.*

2.2.2 Impact of Variability in Funding Timing

Having established that 'earlier is better' (in a stochastic sense) when it comes to the funding timing, we now focus our attention on the variance aspect of the uncertainty in funding. For a fixed funding vector, we compare two funding scenarios; in the first, there is high variability around the number of outstanding installments at the beginning of any given period while there is relatively less variability in the second. More specifically, we represent the two funding scenarios by random variables $\left\{P_t^1(i), i \ \in \{0,1,2,...,m\}\right\}$ and $\left\{P_t^2(i), i \ \in \{0,1,2,...,m\}\right\}$ such that

$$P_t^2(i) \geq_{cvx} P_t^1(i) \ \forall \ i \in \{0,1,2,...m\} \text{ and } t = 2,3,...,N \tag{2.7}$$

$$\{P_t^n(i), i \ \in \{0,1,2,...m\}\} \in \text{SICX}, n = 1,2 \tag{2.8}$$

Condition (2.7) states that $P_t^2(i)$ is larger than $P_t^1(i)$ in the convex order. The following is the definition of a convex order.

Definition 2. *(Shaked and Shanthikumar 2007) Let X and Y be two random variables such that $\mathsf{E}\left[\phi(X)\right] \leq \mathsf{E}\left[\phi(Y)\right]$ for all convex functions $\phi : \mathbb{R} \to \mathbb{R}$. Then X is said to be smaller than Y in the convex order.*

Condition (2.7) implies that $P_t^2(i)$ is more variable than $P_t^1(i)$ but, $\mathsf{E}P_t^2(i) = \mathsf{E}P_t^1(i)$. The convex ordering helps us to isolate the impact of variability around the funding timing since the mean number of outstanding installments remains the same under scenarios 1 and 2. Condition (2.8) states that $\{P_t^n(i), i \ \in \{0,1,2,...m\}\}, n = 1,2$ belong to the class of stochastically increasing convex family of distributions.

Definition 3. *(Shaked and Shanthikumar 2007) Let $\{X(\theta), \theta \in \Theta\}$ be a set of random variables. Then*

1. *$\{X(\theta), \ \theta \in \Theta\} \ \in SI$ (stochastically increasing) if $\mathsf{E}\phi(X(\theta))$ is increasing in θ for all increasing functions ϕ.*

2. $\{X(\theta),\ \theta \in \Theta\}\ \in$ *SICX (stochastically increasing and convex) if* $\{X(\theta),\ \theta \in \Theta\}\ \in$ *SI and* $\mathsf{E}\phi(X(\theta))$ *is increasing convex in* θ *for all increasing convex functions* ϕ.

The following property can be used to check the SICX property for discrete random variables.

Property 1. *(Shaked and Shanthikumar 2007) Suppose that for each* $\theta \in \Theta$*, the support of* $X(\theta)$ *is in* $\mathbb{N}$*. Then,* $\{X(\theta),\ \theta \in \Theta\}\ \in$ *SICX if, and only if,* $\{X(\theta),\ \theta \in \Theta\}\ \in$ *SI and* $\sum_{l=k}^{\infty} Pr\{X(\theta) \geq l\}$ *is increasing convex in* θ *for all* $k \in \mathbb{N}$.

A number of known distributions satisfy the SICX property. For example, if $X(\mu, \sigma)$ is a normal random variable with mean μ and standard deviation σ, then, for each $\sigma > 0$, $\{X(\mu,\sigma), \mu \in \mathbb{R}\} \in$ SICX. Similarly, if $X(\lambda)$ is a Poisson random variable with mean λ, then $\{X(\lambda), \lambda \in [0,\infty)\} \in$ SICX. Another example could be a random variable $X(n)$, which is uniformly distributed on $\{0, 1, 2, ..., n-1\}$. Then, $\{X(n), n \in \mathbb{N}_+\} \in$ SICX.

In the following lemma, we demonstrate that, for any $\{P_t\}$ satisfying condition (2.8), the minimum expected cost with t periods to go is increasing and convex in the number of outstanding installments at the beginning of t, provided the funding vector Z is front–loaded. We refer to a funding vector Z as front–loaded if $Z_i \leq Z_{i+1}, i = 1, 2, ..., N-1$ and back–loaded if $Z_i \geq Z_{i+1}$. Notice that a funding vector with equal installments also satisfies the definition of a front–loaded funding vector.

Lemma 2. *Let* Z *be a front–loaded funding vector and condition (2.8) hold. Then,* $J_t(x_t, r_t + \sum_{k=j+1}^{i} z_k, j)$ *is increasing convex in* j *for* $j \leq i$ *where* $i \in \{0, 1, 2, ...m\}$.

We should point out that front–loaded funding is a sufficient but not necessary condition for the above result to hold. Our final result characterizes the impact of variability in the funding timing—given that the expected number of installments received remains the same, a higher variability around the number of outstanding installments at the beginning of any given period drives up the operating costs and results in poor performance.

Theorem 3. *Let* Z *be a front–loaded funding vector and conditions (2.7) and (2.8) hold. Then,* $J_t^2(x_t, r_t, i) \geq J_t^1(x_t, r_t, i)\ \forall\ i \in \{0, 1, 2, ..., m\}$.

2.2.3 Deterministic Funding Schedule

As we mentioned before, the *deterministic funding schedule* is a special case of the *stochastic funding schedule* where both the installment amounts and the time of receipt of each installment are known. Under a deterministic funding schedule, without loss of generality, we can assume that funding is received in every period (i.e., exactly in N installments) with some installments possibly being 0. Let $V_t\left(x_t, r_t | Z_t\right)$ be the minimum expected cost with t periods to go under a deterministic funding schedule, given state variables x_t, r_t, and the future funding vector Z_t. Since we receive an installment every period, there is no need to keep track of the number of outstanding installments in case of a deterministic funding schedule. As with stochastic funding schedules, we will explicitly condition on Z_t only when we compare different funding schedules.

In sections 2.2.1 and 2.2.2, we fix the funding vector, and explored one part of our second main research question: how does the uncertainty and variability in the funding timing impact operating costs? Using deterministic funding schedules, our goal is to fix the funding timing and explore the other part of our second research question: what is the impact of funding amount and funding schedules as relates to operating costs? Specifically, we are interested in answering the following questions: *1. Does an increase in the total funding received over the horizon necessarily lead to better performance? 2. Given that the total funds received over the horizon remains the same, does advancing additional funds to a certain period have the same level of impact as delaying the same amount until the next period?*

Consider two deterministic funding schedules Z^1 and Z^2 such that

$$\sum_{j=N}^{N-i} z_j^1 \geq \sum_{j=N}^{N-i} z_j^2, \; i = 0, 1, , ..., N-1 \tag{2.9}$$

In Proposition 1, we show that an increase in total funding is guaranteed to result in lower operating costs if condition (2.9) holds.

Proposition 1. *If condition (2.9) holds, then for any $x_N \in \mathbb{R}$, $V_N^2(x_N, r_N^2) \geq V_N^1(x_N, r_N^1)$.*

Consider two funding vectors, Z^1 and Z^2 such that $\sum_{i=1}^{N} Z_i^1 = \sum_{i=1}^{N} Z_i^2$. If Z^1 is front–loaded and $Z^1_i = Z^2_{N-i+1}, i = 1, 2, ..., N$, then clearly, Z^2 is back–loaded. In this case, a direct application of the above proposition shows that for a fixed amount of total funding, the front–loaded vector Z^1 is guaranteed to perform at least as well as the back–loaded vector Z^2. This result is on expected lines — what would be more interesting is to compare the following two funding scenarios: receiving less total funding, say $10 M, in a front–loaded fashion vs. receiving more total funding, say $ 20 M, in a back–loaded fashion. In this case, it is not clear whether the additional funding received leads to lower operating costs — we explore this in our computational study.

Next, we compare three deterministic funding vectors, $Z = (z_1, z_2, ..., z_{t-1}, z_t, ...z_N)$, $Z^A = (z_1, z_2, ..., z_{t-1} - \delta, z_t + \delta, ..., z_N)$ and $Z^D = (z_1, z_2, ..., z_{t-1} + \delta, z_t - \delta, ..., z_N)$. Let V_t, V^A_t and V^D_t be the value functions associated with the three funding vectors respectively. We have the following result.

Proposition 2. $V^D_N(x_N, r_N) - V_N(x_N, r_N) \geq V_N(x_N, r_N) - V^A_N(x_N, r_N)\ \forall\ x_N, r_N.$

Proposition 2 implies that the cost savings resulting from advancing additional funds to a certain period does not match the extra cost incurred if the same amount were to be delayed by one period. In essence, funding delays hurt more than the benefits from receiving funding early.

2.3 Computational Study

For our computational study, we first consider deterministic funding schedules and analyze how funding patterns impact operating costs. Subsequently, we consider stochastic funding schedules to understand how uncertainty in funding affects system performance.

2.3.1 Deterministic Funding Schedules

Experimental Setup:

The numerical study was conducted for planning horizon, N, of different lengths, N =2, 4, 6, 12 and 24. The unit purchase cost c was normalized to 1 in all numerical experiments. For each N, we varied the following parameters.

Holding Cost: We chose four values for holding cost, h=0.01, 0.05, 0.1 and 0.25.

Penalty Cost: For each value of h, the penalty cost was varied so that the critical ratio (CR), $(b-c)/(b+h)$, took on values 0.2, 0.4, 0.6, 0.8, 0.9 and 0.95 respectively.

Demand: In our work, we consider uniform and truncated normal demand distributions. To test the impact of demand variability on the operating costs, we considered U ~[70,130] and U ~[25,175] for the uniform demand case. For truncated normal demand, the mean was fixed at 100 units and we used CV values of 0.1 and 0.25. The normal distribution was truncated at 3 standard deviations. Thus, for each N, we have 4*6*4=96 problem instances.

Funding Patterns: For each combination of N, h, b and demand distribution, we consider five different funding patterns and four funding levels. The five funding patterns are extremely front–loaded funding (EFL), moderately front–loaded funding (MFL), evenly–spread funding (ES), moderately back–loaded funding (MBL) and extremely back–loaded funding (EBL). The holding cost and the funding patterns were chosen so as to be consistent with an earlier study that analyzes the state of the RUTF supply chain in the Horn of Africa (UNICEF 2009). For the backorder costs, due to lack of precise estimates, we decided to carry out a sensitivity analysis over a wide range of critical ratios. For all the funding vectors, the total funding received remained the same and is equal to N*funding level*mean demand. In our experiments, we consider 25%, 50%, 75% and 100% system funding levels and we label them as severely under–financed, moderately under–financed, mildly under–financed and fully–financed systems respectively. For example, a severely under–financed system receives N*0.25*mean demand over the entire planning period. For illustration purposes, we use the specific case of N=4, U ~[70,130] demand distribution and 100% funding level to explain the difference between the different funding vectors.

EFL: The entire funding (=N*mean demand) is received upfront, i.e., the funding vector is (0, 0, 0, 400). Recall that we count time in the reverse order.

MFL: In the first $N/2$ periods, the installment size is 1.5*mean demand followed by 0.5*mean demand in the last $N/2$ periods. For the specific case considered, the funding vector is (50, 50, 150, 150).

ES: Every installment is equal to mean demand, i.e., the funding vector is (100, 100, 100, 100).

MBL: In the first $N/2$ periods, the installment size is 0.5*mean demand followed by 1.5*mean demand in the last $N/2$ periods. The funding vector would be (150, 150, 50, 50).

EBL: The entire funding is back–loaded to the last period, i.e., the funding vector is (400, 0, 0, 0).

Notice that as we move from EBL to EFL funding, more and more funds are received in the initial periods. Compared to ES funding, the front–loaded vectors can be considered as funding advances and back–loading can be considered as a funding delay. By using ES funding as a benchmark, we investigate how back–loading and front–loading the funding impacts operating costs.

Impact of Funding Pattern

We use the cost incurred under ES funding as a benchmark and compute the relative percentage cost difference for a particular funding pattern, say EBL funding, as follows: 100*($cost_{EBL} - cost_{ES})/cost_{ES}$. Table 2.1 provides the relative percentage cost difference (relative to ES funding) for different funding patterns at 100% funding level, averaged over the 96 problem instances.

From Table 2.1, we can immediately make an important observation: operating costs increase almost exponentially with funding delays (back–loading). By maintaining a consistent and even flow of cash to fund its operations (ES funding), an organization can cut down on price

	EBL	MBL	MFL	EFL
N=2	152.54	69.42	-13.23	-13.69
N=4	425.68	126.30	-23.10	-24.89
N=6	664.40	178.97	-28.63	-30.76
N=12	1269.97	318.55	-37.61	-39.86
N=24	2230.19	547.26	-46.22	-48.30
Overall average	948.56	248.10	-29.76	-31.5

Table 2.1: Average % cost difference for different funding patterns relative to ES funding

premiums due to funding delays and utilize valuable aid dollars to increase coverage. However, it is not the optimal funding pattern since, under ES funding, there is very little flexibility to deal with large demand surges upfront. This is where an additional funding influx in the initial periods proves valuable. Relative to ES funding, the moderate shift in funds to the initial periods (MFL funding) results in average (averaged over all N) savings of 29.8%. However, pushing more and more funds to the initial periods yields little to no return and, even in case of an extreme push (EFL funding), the average savings is only 31.5%.

Also, notice that the gulf between front–loaded and back–loaded vectors increases with horizon length (Table 2.1) and critical ratio (Table 2.2) while it decreases with demand variability (Table 2.3). More interestingly, Table 2.3 also tells us that the benefits from front–loaded funding increase with the demand variance while the negative impact of a funding back–load is mitigated by the increased demand variability.

	critical ratio					
	0.2	0.4	0.6	0.8	0.9	0.95
EBL	265.88	342.74	476.63	726.81	1094.32	1385.91
MBL	71.89	92.84	129.35	196.14	297.75	377.25
MFL	-9.50	-12.58	-17.99	-30.69	-43.13	-55.01
EFL	-9.76	-12.94	-18.52	-31.92	-44.41	-56.65

Table 2.2: Average relative % cost difference as a function of CR for N=6, U ∼[70,130] demand distribution and 100% funding level

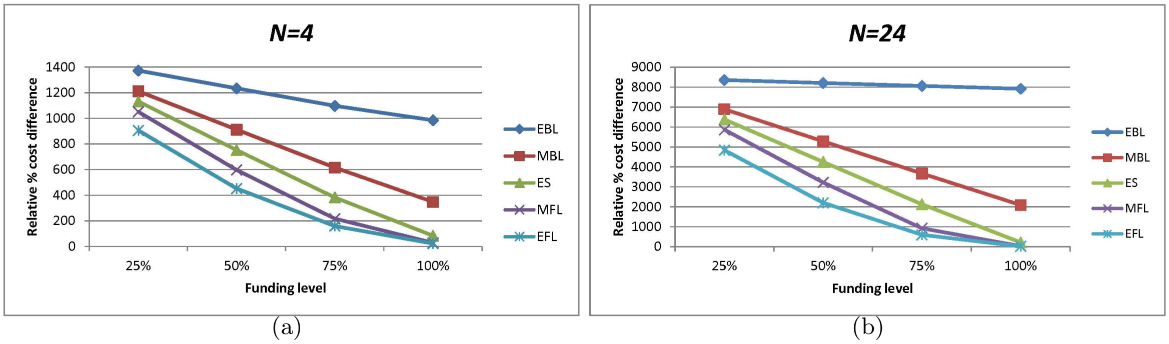
	EBL	MBL	MFL	EFL
U ∼[70,130]	1020.73	268.76	-29.07	-29.85
U ∼[25,175]	554.96	133.46	-36.14	-40.52
N ∼[70,130]	1392.03	378.07	-21.90	-22.05
N ∼[25,175]	826.50	212.11	-31.92	-33.58

Table 2.3: Effect of demand variability on the average relative % cost difference for different funding patterns at 100% funding level

Funding Level vs. Funding Pattern

Having analyzed the impact of funding pattern on operating costs, we now proceed to understand the interaction between funding level and the funding pattern. We are mainly interested in understanding the relative importance of level of funding vis–a–vis funding pattern. Figure 2.1 displays the relative percentage cost difference at different funding levels, ranging from severely under–financed to fully–financed. Here, we use the cost incurred under no funding constraints (NFC) as the benchmark to compute the relative percentage cost difference.

From Figure 2.1, we see that at very low funding levels (25% and 50% funding levels), the funding pattern is inconsequential — 100% funding level almost always outperforms. For a mildly under–financed system, the results are drastically different. From Figure 2.1, we see that back–loaded funding at 100% funding level performs significantly worse compared to front–loaded funding at 75% funding level. However, ES funding at 100% funding level outperforms even EFL funding at 75% funding level. This demonstrates that, at reasonably high funding levels, funding pattern is critical to system performance and a further increase in overall funding should not be traded for a delay in funding.



(a) (b)

Figure 2.1: Average % cost difference for different funding patterns at different funding levels

Figure 2.1 also offers some additional insights into the interaction between funding levels and funding patterns. Notice that the benefits of front–loading (the gap between the lines corresponding to MFL/EFL and ES funding) are significantly higher in under–financed systems when compared to a fully–financed system. However, the maximum benefits of front–loading are not observed in severely under–financed systems as one would expect — for both EFL and MFL funding, the benefits of front–loading follow a U–shaped pattern with the funding level, with maximum benefits seen either in moderately or mildly under–financed systems. The least benefits of front–loading are observed at 100% funding level. An intuitive reasoning for the U–shaped pattern is as follows: for severely under–financed systems, the total amount pushed to the initial periods is relatively less and in fully–financed systems, there is sufficient cash already available in the system for the additional funding to make a large impact.

In case of back–loaded funding, the additional cost incurred due to the delayed receipt of a majority of the funds is monotonically increasing in the funding level. The monotonicity result can be explained as follows: when funding is back–loaded, a major portion of the operating costs can be attributed to backorders. As the funding level increases, the difference between the funds available in each period under the evenly–spread and back–loaded funding scenarios also increase. This implies that the additional backorders ascribed to a funding delay also increase with the funding level.

Impact of Funding Constraints

Most global health programs are already financially constrained, a situation that is only expected to worsen in the near future (Stokes 2011). The numbers in Table 2.4 offer some insights into the role of funding constraints. By taking difference of the relative percentage cost differences in Table 2.4, we get 100*$(cost_{EBL} - cost_{ES})/cost_{NFC}$=2805.93, while 100*$(cost_{ES} - cost_{NFC})/cost_{NFC}$=121.18. This demonstrates that, as we move from EBL funding to NFC, a majority of the resultant benefits actually stem from reducing funding delays (making the funding even) and only a relatively small portion of the cost savings are attributed to the unlimited funding availability. Recall that NFC refers to a schedule where funding is never a constraint. The same insight also holds for MBL funding.

EBL	MBL	ES	MFL	EFL
2927.11	834.39	121.18	27.70	21.34

Table 2.4: Average % cost difference for different funding patterns relative to NFC at 100% funding level

Also, notice in Table 2.4 that the average increase in costs due to the presence of funding constraints is less than 22 % for EFL funding and less than 28 % for MFL funding. Having access to unlimited funding is practically unrealistic and our study shows that front–loaded funding compares favorably with unlimited funding. Thus, even with limited funding, it is possible to achieve reasonable system performance as long as “enough” funding can be secured in the initial periods.

2.3.2 Stochastic Funding Schedules

Experimental Setup

Except for funding patterns, all other elements of the experimental set up remain unchanged from the deterministic funding case. As we mentioned in section 2.2, funding is received in $m \leq N$ installments. For a fixed N, we consider several values of m for the stochastic funding case, details of which are given in Table 2.5.

N	m
2	1,2
4	1,2,3,4
6	1,2,3,4,6
12	1,2,3,4,6,12
24	1,2,3,4,6,12,24

Table 2.5: Number of installments considered for each N

For the stochastic funding case, we assume that the funding level is 100%, i.e., the total funds received over the entire planning horizon is fixed and is equal to N*mean demand in each period. To capture the uncertainty in the funding timing, we vary the number of installments. Depending on the number of installments (m) received, the amount received in each installment varies (=N/m*mean demand in each period).

We assume that the number of installments received in a period is uniformly distributed between 0 and the number of outstanding installments. To understand what happens when we increase the number of installments, consider the example of N=4, m=1, and U ~[70,130] demand. In this scenario, when m=1, any of the four extreme funding scenarios namely, (400,0,0,0), (0,400,0,0), (0,0,400,0) and (0,0,0,400), are equally likely. When we increase m to 2, the probability of extreme funding scenarios like (400,0,0,0) reduce drastically from 1/4 to

1/16 while the probability of a more evenly spread funding vector like (200,0,200,0) increases to 1/8. Hence, the idea is that, as we increase the number of installments, the probability of the funding being more smooth and evenly spread out increases, thereby reducing the volatility in funding received until any given period.

Impact of Funding Uncertainty

We begin by discussing a result that is intuitive and holds true for all problem instances: *reducing the funding volatility, and making the funding more smooth and evenly spread out lowers operating costs* (see Figure 2.2). However, such benefits of reducing the funding uncertainty show diminishing rates of return, i.e., as the number of installments in which funding is currently received increases, the marginal value of receiving the funding in an additional installment decreases.

To understand why reducing funding uncertainty leads to lower expected operating costs, consider the case where m=1 and N=6. The single installment could be received in any of the six periods and all possibilities are equally likely. Of course, there is nothing better than receiving the installment in the very first period (probability 1/6) but we also need to take into account the other possibilities (including an extreme back–loading with probability 1/6). Considered together, receiving funding in one installment is no longer the ideal scenario — in fact, it is the worst. In general, when the number of installments decreases, i.e., funding becomes more uncertain, it increases the possibility of the funding vector being moderately or severely back–loaded. Under uncertain funding, the possibility of the funding vector being front–loaded also exists but the non–linear increase in costs as we move from front–loaded to back–loaded funding means that the overall impact of the funding uncertainty is negative, i.e., in expectation, the operating costs increase.

From Table 2.6, we see that the benefits of reducing funding uncertainty increases with the critical ratio. More importantly, Table 2.7 illustrates that the benefits of reducing funding uncertainty decrease with demand variability. This result brings forth a very important insight: the underlying demand situation has to be taken into account before embarking on initiatives to reduce funding uncertainty. In highly volatile demand environments, the cost of reducing

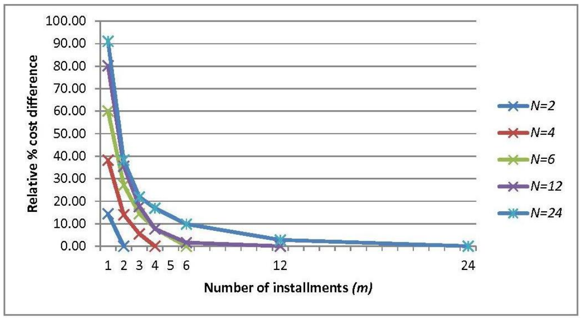


Figure 2.2: Average % cost difference due to funding timing uncertainty

the funding uncertainty might outweigh the benefits.

	critical ratio					
	0.2	0.4	0.6	0.8	0.9	0.95
m=1	39.19	45.91	55.42	69.90	80.43	86.98
m=2	17.83	20.89	25.21	31.81	36.59	39.58
m=3	9.49	11.12	13.42	16.93	19.48	21.06
m=4	5.14	6.02	7.26	9.16	10.54	11.40

Table 2.6: Average relative % cost difference due to funding uncertainty (relative to a funding schedule with m=N) as a function of CR for N=6 and U ∼[70,130] distribution

		m=1	m=2	m=3	m=4	m=6	m=12	m=24
N=4	U ∼[70,130]	40.48	14.96	5.91	0	NA	NA	NA
	U ∼[25,175]	31.11	11.07	3.69	0	NA	NA	NA
N=24	U ∼[70,130]	92.82	39.29	22.57	17.42	10.21	2.96	0
	U ∼[25,175]	85.98	35.64	20.10	15.39	8.75	2.34	0

Table 2.7: Effect of demand variability on the average relative % cost difference due to funding uncertainty (relative to a funding schedule with m=N)

Funding Level vs. Funding Uncertainty

In this section, we aim to address our last but nevertheless, an important question: which of the two hurts system performance more — funding level or funding uncertainty ? To answer this question, we compare deterministic funding patterns at different funding levels to stochastic funding at 100% funding level. The results for N=24 are provided in Tables 2.8 and 2.9. The results are very similar for other values of N.

Comparing rows 1 and 2 in Table 2.8 with Table 2.9, we see that at low (25% and 50%) funding levels, the funding pattern is inconsequential. Receiving less overall funding severely hurts performance, making even the most uncertain funding (m=1) at 100% funding level attractive in comparison in almost all cases (an exception being EFL funding at 50% funding level).

Funding level	EBL	MBL	ES	MFL	EFL
25%	8357.25	6894.14	6377.74	5861.40	4837.84
50%	8208.85	5282.63	4250.12	3220.79	2200.73
75%	8060.46	3671.19	2125.99	923.60	596.72
100%	7917.89	2086.40	213.73	26.65	16.80

Table 2.8: Average relative % cost difference for different deterministic funding patterns at different funding levels

For 75% funding level, the results are very different (compare row 3 in Table 2.8 with Table 2.9). While the back–loaded vectors at 75% funding level perform significantly worse than even the most uncertain funding at 100% funding level, front–loaded funding at 75% funding level outperforms uncertain funding at 100% funding level, even when the uncertainty is considerably reduced (m=24). This demonstrates that at relatively high funding levels, the choice between deterministic funding and an even larger overall but uncertain funding is not obvious — the answer depends on the deterministic funding pattern and the level of uncertainty in the larger overall funding.

2.4 Generalization to Positive Lead Times and Uncertain Installment Amounts

In this section, we demonstrate that one of the key results of our paper — the optimality of a state–independent modified base stock policy — holds under more general settings than the one we considered in Section 2.2. Consider a generalized version of the funding problem P1, which we label problem P3, where the replenishment lead time λ could be ≥ 0. Furthermore, the funding received in period t could be any random variable on [0, outstanding funding at

m=1	m=2	m=3	m=4	m=6	m=12	m=24
2800.83	1968.97	1710.29	1630.68	1518.65	1407.39	1363

Table 2.9: Average relative % cost difference (relative to NFC) due to funding timing uncertainty

the beginning of period t], with no restriction on the specific shape or form of the distribution. Notice that the funding dynamics described in Section 2.2 imply that the funding received in period $t-1$ would equal $\sum_{j=O_{t-1}+1}^{O_t} z_j$ with probability $P_t(O_t, O_{t-1})$. It is not hard to see that this is a special case of the more general funding situation that we assume for problem P3.

Let $G_t^\lambda(x_t, w_t^1, w_t^2, ..., w_t^{\lambda-1}, R_t, OF_t)$ be the minimum expected cost–to–go in this more general setting given that x_t is the on–hand inventory (after receiving shipments at the beginning of the period), w_t^i represents the order placed i periods ago, OF_t is the outstanding funding amount at the beginning of period t, and $R_t = cIP_t + r_t$. Here $IP_t = x_t + \sum_{j=1}^{\lambda-1} w_t^j$ represents the inventory position at the beginning of period t and r_t is the capital available on–hand. Then,

$$\text{(P3)} \quad G_t^\lambda(x_t, w_t^1, w_t^2, ..., w_t^{\lambda-1}, R_t, OF_t)$$

$$= \min_{0 \le z \le \frac{r_t}{c}} \left\{ \begin{array}{l} cz + b\mathsf{E}_{\zeta_t}[\zeta_t - x_t]^+ + h\mathsf{E}_{\zeta_t}[x_t - \zeta_t]^+ \\ +\mathsf{E}_{OF_{t-1}|OF_t}\mathsf{E}_{\zeta_t} G_{t-1}^\lambda(x_t - \zeta_t + w_t^1, w_t^2, ..., z, R_t - c\zeta_t + (OF_t - OF_{t-1}), OF_{t-1}) \end{array} \right\}$$

The terminal condition is $G_0^\lambda(x_0, w_0^1, ..., w_0^{\lambda-1}, R_0, 0) = 0 \ \forall \ (x_0, w_0^1, ..., w_0^{\lambda-1}, R_0, 0)$.

Now consider a multi–period inventory management problem, which we label problem P4, with the same cost and demand parameters and replenishment lead time as problem P3, but no financial constraints. Let $NV_t^\lambda(x_t, w_t^1, w_t^2, ..., w_t^{\lambda-1})$ be the minimum expected cost with t periods to go corresponding to problem P4. Then,

$$\text{(P4)} \qquad NV_t^\lambda(x_t, w_t^1, w_t^2, ..., w_t^{\lambda-1}) = \min_{z \ge 0} \left\{ \begin{array}{l} cz + b\mathsf{E}_{\zeta_t}[\zeta_t - x_t]^+ + h\mathsf{E}_{\zeta_t}[x_t - \zeta_t]^+ \\ +\mathsf{E}_{\zeta_t} NV_{t-1}^\lambda(x_t - \zeta_t + w_t^1, w_t^2, ..., w_t^{\lambda-1}, z) \end{array} \right\}$$

For problem P4, it is well known that there exists an optimal base stock level $y_t^{\lambda*}$ in each period such that if the inventory position in period t is below $y_t^{\lambda*}$, it is optimal to order up–to $y_t^{\lambda*}$, and not order otherwise. In Theorem 4, we prove that the unconstrained base stock levels $y_t^{\lambda*}, y_{t-1}^{\lambda*}, ..., y_1^{\lambda*}$, optimal for problem P4, continue to be optimal for problem P3 with funding constraints as well.

Theorem 4. *Let $y_t^{\lambda*}, y_{t-1}^{\lambda*}, ..., y_1^{\lambda*}$ be the optimal base stock levels corresponding to problem P4* *with replenishment lead time λ. Let $(x_t, w_t^1, w_t^2, ..., w_t^{\lambda-1}, R_t, OF_t)$ be the state of the system in problem P3 at the beginning of period t. Then, the optimal ordering policy for problem P3 has the following simple structure.*

$$\begin{array}{ll} \textit{Order up–to } R_t/c & \textit{if } R_t/c \leq y_t^{\lambda *} \\ \textit{Order up–to } y_t^{\lambda *} & \textit{if } R_t/c > y_t^{\lambda *}, IP_t < y_t^{\lambda *} \textit{ and} \\ \textit{Do not order} & \textit{if } IP_t \geq y_t^{\lambda *}. \end{array}$$

2.5 Conclusions and Managerial Insights

Incorporating funding flows into operational decisions is necessary for making optimal and operationally feasible decisions. In this chapter, we study the problem of managing inventory of a health commodity subject to variable funding constraints. Our work brings out several important insights that would be valuable to humanitarian supply chain managers. To begin with, we find that preventing funding delays should be the top–most priority for humanitarian organizations. Our results regarding the benefits of front–loaded funding are timely in light of the on–going efforts within the global health community to achieve front–loaded funding. Our study suggests that such initiatives to front–load the funding need to be reconciled with the system funding level. Moderate front–loading brings significant benefits at all funding levels, but extreme front–loading, while it looks promising, brings little to no additional benefits in a fully–financed system. Hence, managers need to exercise caution and use careful judgement when deciding the level of front–loading. We believe that our model can serve as a cost–benefit analysis tool to facilitate such decisions.

Often times, humanitarian organizations make an all–out effort to raise as much funding as possible to support the various programs but our analysis shows that such an approach is not the most effective one. Our results indicate that even if the funding level is lower, performance may be better if the funding is received earlier or in a steady fashion.

Finally, managers also need to pay close attention to their operating environment when taking steps to improve the funding situation. One such aspect is the volatility of the underlying demand. While the magnitude of the benefits of front–loading increase with demand volatility, the opposite is true regarding the savings resulting from reducing the funding uncertainty. Given such contrasting results, a good understanding of the operating environment would help in steering the funding–related efforts in the right direction. Front–loading initiatives are likely to yield significant benefits in highly unpredictable environments like HIV/AIDS programs while reducing the funding uncertainty can be expected to result in substantial savings in case of health commodities like reproductive supplies which have a stable and predictable demand pattern.

One of the limitations of our study is the assumption that each patient/customer requires only one dose/unit of the product. In the problem that motivated this study, children diagnosed with severe acute malnutrition are given RUTF for several weeks. Typically, the child's condition improves with every dose while non–provision could lead to health deterioration. Capturing this disease progression would require a more sophisticated model that the one we considered in this chapter. Given that it might take several weeks before a child completely recovers from malnutrition, there will be groups of children in different stages of malnutrition enrolled in the program at any given point in time. An important issue that arises in this context is the allocation of a limited quantity of a scarce resource, e.g. inventory or funding, amongst people in different health states. This will be the focus of our analysis in the next chapter.

Chapter 3

Resource Allocation in Humanitarian Health Settings

3.1 Introduction

In Chapter 2, we analyzed the impact of funding in humanitarian operations in the context of managing inventory of a nutritional product in the presence of funding constraints over a finite horizon. We studied the inventory management problem under the assumption that each customer/patient requires only one dose/unit of the product, and unfulfilled demand is completely backlogged. We characterized the optimal inventory replenishment policy and also offered several insights into the impact of amount, schedule, and uncertainty in funding. The single–dose assumption made in Chapter 2 makes the model general enough to be applicable to a variety of humanitarian health programs but it is a simplifying assumption in view of the specific malnutrition context that motivated the work, since children who suffer from malnutrition are typically treated using RUTF for several weeks before they are declared fit and discharged from the program.

In this chapter, we extend the work in Chapter 2 by relaxing the single dose assumption and allow for the possibility that patients in different health states might require treatment over different lengths of time (corresponding to different amounts/doses of the product) before they are completely cured. In the context of the malnutrition program that motivates our work, children who are screened by the program and diagnosed as malnourished could be of either type: moderately malnourished or severely malnourished. The treatment duration depends on whether they are moderately or severely malnourished and the response to “treatment” or "non–treatment" in any given period could also be different between the two groups. We point out that the multi–dose framework, allowing for different lengths of treatment time, is not specific to our context and is appropriate in many other health care settings as well.

Using a two–health states model, we study the problem of dynamic allocation of a limited amount of resource, which in our case is donor–funding, to patients in different health states over a finite horizon with the objective of minimizing the number of 'disease–adjusted life months' lost. One of the two health states is assumed to be a *less severe* health state and the other one is *more severe.* Funding is received in installments throughout the planning period with uncertainty around the timing and amount. New patients of both health states enter the program in every period. In this setting, a key decision facing public health managers is: how to allocate funding between people in the two health states and in anticipation of a shortage in funding in the near future, should they *reserve* a certain amount of funding for the more severe patients who might show up in the future periods? Answering this question assumes significance in light of the variable and unpredictable nature of the funding in the humanitarian health sector but the decision is significantly complicated by the fact that in the absence of treatment, patients in the less–severe health might deteriorate to the more–severe state. Our goals for this chapter are two–fold: (i) to determine ways to efficiently allocate funding between the two health states in every period, taking into account the current funding availability and future financial inflows, and (ii) characterize the impact of system parameters, funding level (total funding received as a % of the funding required to completely cure the total expected state 1 and state 2 patients), and uncertainty in funding on the number of disease–adjusted life months lost.

Among other results, we show that the optimal allocation policy is state–dependent, which significantly complicates the computation of the optimal policy. We prove several monotonicity results that can help reduce the computational burden. However, despite the simplifications, determining the optimal policy is challenging for longer planning horizons and it may not be practical for realistic–size problems. Hence we propose two heuristics, FCFS heuristic and the PNS heuristic, that are easy to understand and implement. Our computational results show that PNS performs well in terms of the solution quality and running time across a wide range of scenarios. The FCFS heuristic also performs well in many cases but it is less robust than the PNS heuristic and in certain settings, there is a noticeable performance gap between the two heuristics.

Our computational study also provides several insights regarding the impact of funding level and uncertainty in funding. For example, our analysis shows that the impact of uncertainty in funding varies depending on the funding level and the length of the planning horizon. For short–planning horizons, uncertainty in funding leads to a loss of disease–adjusted life months while in case of longer planning horizons, receiving the funding in fewer, lumpy installments involving more uncertainty in funding timing might be preferable only in under–financed systems ($<$ 100% funding level). In well–funded systems ($\geq$ 100% funding level), having a smooth and predictable funding pattern is always preferred. Our analysis also shows that the trade–off between receiving less overall funding in a more predictable fashion and receiving additional funding with increased uncertainty is not straight forward — in under–financed systems, in general, it is preferable to go for the additional funding while in systems with buffer funding ($>$ 100% funding level), the losses from the increased uncertainty outweigh the benefits of additional funding.

3.1.1 Literature Review

Our work is mainly related to two streams of literature. 1. Inventory management with multiple demand classes 2. Resource allocation in humanitarian health settings.

Inventory management with multiple demand classes: The Operations Management (OM) literature is rich with papers studying the problem of managing the inventory of a product when facing demand from multiple customer classes. The customer classes could be different from one another in terms of their penalty costs, and whether unfulfilled demand from each class is lost or backordered. Klejin and Dekker (1998) provide an excellent overview of inventory systems with multiple demand classes. Within the multiple demand class literature, many authors have studied the inventory management problem under both periodic review (e.g., Evans 1968, Kaplan 1969, Veinott 1965, Topkis 1968, Frank et al. 2003) and continuous–review (e.g., Nahmias and Demmy 1981, Deshpande et al. 2003, Arslan et al. 2007) settings. A key differentiating factor between our work and this stream of literature is that, in our work, patients transition between the two health states (due to either treatment or non–treatment) while customers do not switch or move between the classes in the multiple demand class literature. The transition of patients between the two health states complicates the allocation decision further since the number of patients in the two health states in the future periods now depend on the allocation levels to both the health states in the current period.

Resource allocation in humanitarian health settings: A few papers within the OM literature have looked at resource allocation problems from a global health perspective. Deo at al. (2012) study a model of community–based health care delivery system with limited capacity with the objective of maximizing health outcomes through better capacity allocation across multiple patients. In their work, the available capacity is fixed and unlike funding, unused capacity cannot be utilized in the later periods. Deo and Corbett (2010) consider the dynamic allocation of a scarce resource, ARV drugs, used in the management of HIV/AIDS. In their model, the trade–off is between continuing treatment for current patients and initiating treatment for new patients. A key differentiating factor between our work and their model is that, in Deo and Corbett (2010), the supply of ARV drugs in every period is an i.i.d random variable while in our case, funding received is correlated across periods. Yang et al. (2013) develop an optimization model to choose which children (from among a group) should receive ready–to–use therapeutic or supplementary food, based on a child's sex, age, height–for–age and weight–for–height scores, to minimize the mean number of disease–adjusted life years (DALYs) lost. In their model, however, the total funding for the entire planning horizon is available upfront while we study the allocation problem in a setting where funding is received in installments throughout the planning period.

The rest of the paper is organized as follows. In Section 3.2, we describe the model in detail. In Section 3.3, we present results regarding the optimal allocation policy. In Section 3.4, we develop two heuristics to handle realistic–size problems. Section 3.5 provides some analytical results regarding the impact of funding. In Section 3.6, we discuss the results of our computational study to evaluate the performance of the heuristics and also analyze the impact of system parameters and changes in funding. The last section concludes the paper.

3.2 Model

We consider the problem of dynamic allocation of a scare resource, which in our case is donor–funding, to patients in different health states over a finite horizon. The planning horizon is divided into T periods, with time indexing done in the reverse order, i.e., period T is the first period, followed by T-1, T-2,..., and so on. At the start of any period t, we assume that there are n_t^1 patients in the less–severe health state, labeled 'state 1' and n_t^2 patients in a more–severe health state, labeled 'state 2'. In the malnutrition context, state 1 could be thought of as representing children suffering from moderate acute malnutrition and state 2 to be representing severe acute malnutrition. While it is possible that there could be more than two health states depending on the disease, we believe that the two–state model captures the key trade–offs inherent in resource allocation problems faced by public health managers, while keeping the model tractable for computational purposes. Before we get into the dynamics concerning n_t^1 and n_t^2, let us first explain the distinguishing factors between health states 1 and 2.

The two states differ from one another along three key dimensions: 1. how the health state changes in response to "non–treatment" in any given period. 2. the per–period costs associated with being in state $i, i = 1, 2$, and 3. the terminal cost. Let us first discuss the evolution of the patients' health states in the absence of treatment. In any given period, when treatment is not provided, we assume that α_{11} fraction of the patients in state 1 will continue to remain in state 1, while the remaining $\alpha_{12} = 1 - \alpha_{11}$ fraction deteriorate to state 2. In case of state 2, in the absence of treatment, we assume that α_{22} fraction of the patients will continue to remain in state 2, while the remaining α_{2E}=$1 - \alpha_{22}$ fraction exit the system. Patients could exit the system for a variety of reasons, e.g., death, defaulting, loss of confidence in the program etc. In our analysis, we do not distinguish between the different reasons (since in practice it is often difficult to ascertain the exact reason) and assume that for every patient who exits the system in period t, we incur a fixed penalty of l_t^E. In the traditional Operations Management literature, l_t^E is often the lost revenue from not being able to satisfy customer demand. In the health care literature, there exist several approaches to quantify l_t^E, the most popular one being "Years of Life Lost" (YLL) due to premature death. YLL is typically calculated as the "life expectancy" at the age of death, a population–wide estimate that is published and regularly updated by the

World Health Organization (WHO). Variants of the life–expectancy measure include “healthy life expectancy” (HLE) and “disease–adjusted life expectancy” (DALE). In our work, we use DALE (see Mathers et al. 2000) in the calculation of l_t^E. Treating one period of our analysis as roughly a month, we compute l_t^E as follows: $l_t^E = (t+1) + DALE \times 12$, i.e., l^E disease–adjusted life months are lost when a person exits the system. Notice that the addition of the $t+1$ term in the calculation of l_t^E implies that there is a higher penalty incurred for early exits from the system.

Next, consider the per–period costs. Since people in both states 1 and 2 suffer from a “less than ideal” health state, we assume that there is a per–period (penalty) cost $b^i,\ i = 1, 2$ associated with being in health state i, i.e., it captures the burden of not being in a perfectly healthy state. The penalty for being in a perfect health state is 0. Naturally, $0 \leq b^1 \leq b^2$. Again, there exist several approaches in the health care literature to quantify the per–period penalty b^i. One such approach is the use of “disability weights” employed in the calculation of disease–adjusted life years (DALYs). The disability weights for several diseases and conditions are published and regularly updated by WHO (see WHO 2004). To better understand how the disability weights can be used, let us suppose that a person suffers from a particular disease which has a disability score of 0.2. The disability scores for perfect health and death are 0 and 1 respectively. Then, for every year lived with the disease (with a disability score of 0.2), 0.2×1=0.2 DALYs are lost. Notice that the disability weights offer a natural way of quantifying the per–period cost b^i, since they explicitly capture the relative impact of being in a particular health state in any given period.

In addition to the per–period costs, we also consider terminal costs t^1 and t^2 associated with health states 1 and 2 respectively. In contrast to the per–period costs, which capture the short–term effects of being in a particular health state, the terminal costs capture the long–term impact. The long–term impact could be very different depending on the disease. For certain diseases, there is very little to no long–term impact while in case of diseases like HIV/AIDS, and stunting due to chronic malnutrition, the impact could be life–long. In our work, we calculate $t^i,\ i = 1, 2$ to be $\textit{mortality rate}_i \times$ DALE $\times$ 12 where $\textit{mortality rate}_i$ refers to the mortality rate in health state i. We provide additional details regarding the calculation of the per–period and terminal costs in Section 3.6.

So far, we have not discussed the effect of treatment on patients belonging to the two health states. We assume that *treatment is perfect* for both the health states, i.e., after receiving treatment, patients in state 2 transition to state 1 and patients in state 1 are completely cured and discharged from the system (not to be confused with *exiting* the system).

3.2.1 Patient Entry Dynamics

New patients belonging to both health states enter the system in every period. We use an *incidence* model to capture the patient entry dynamics — in every period t, we assume that a *random* number, denoted by n_t^N, of people enter the catchment population. For example, in the malnutrition context, this could represent the number of newborns in a particular month. We assume that n_t^N is independent and identically distributed with a probability distribution function g_t and cumulative density function G_t. A fixed fraction, $\beta_1 + \beta_2$, of the people entering the catchment population are assumed to be infected/suffering from the disease, with β_1 fraction belonging to state 1 and β_2 fraction belonging to state 2. The remaining fraction, $1 - \beta_1 - \beta_2$, are healthy and do not require any treatment. β_1 and β_2 can be thought of as the "disease incidence rates", i.e., the number of new cases of the disease divided by the number of people at risk over a given time period.

3.2.2 Funding Inflows

The system is funded by external donor–funding, received in multiple installments over the planning horizon. Typically, donors make a commitment at the beginning of the fiscal year after a careful evaluation of the funding proposals received from the recipient countries. The total promised amount, which we denote by TPF, is then dispersed in installments of various sizes throughout the year, depending on the budget and funding cycles of the donor. We should remark that, while the amount in the individual installments may vary, in most cases, the committed amount is received in full before the end of the planning period for which funding was sought.

We capture the funding inflow using a very general model of ‘outstanding funding’, denoted by OF_t, at the beginning of every period. OF_t indicates the amount that is yet to be received as of period t, after receiving funding, if any, at the beginning of that period. Of course, $OF_{T+1} = TPF$. Given OF_t, let $OF_{t-1} |_{OF_t}$ denote the random variable corresponding to the outstanding funding amount at the beginning of period $t-1$. We do not impose any specific restrictions on the shape or form of $OF_{t-1} |_{OF_t}$, but we will introduce conditions that $OF_{t-1} |_{OF_t}$ needs to satisfy for certain results to hold, as and when required. Note that $OF_t - OF_{t-1}$ is the funding received at the beginning of period t-1. Also, in light of our earlier comment that donors typically disburse the committed amount in full before the end of the planning period, we assume that OF_1=0 always.

3.2.3 Objective Function

Before we state our objective function, let us specify the sequence of events in any given period. 1. First, funding (if any) is received at the start of period t and simultaneously, new patients enter the system. 2. Decisions concerning allocations to states 1 and 2 are made, subject to capital available on–hand and the number of patients in the two health states. 3. Based on the allocation decisions, state–transitions take place. We assume that the transitions happen instantaneously. 4. Finally, per–period penalties are incurred based on the number of patients in every health state and the number of people exiting the system.

Let r_t represent the funding available on–hand at the beginning of period t, after receiving funding (if any) at the beginning of the period. Given state variables n_t^1, n_t^2, r_t and OF_t, let $V_t(n_t^1, n_t^2, r_t, OF_t)$ denote the ‘minimum expected disease–adjusted life months lost’ with t periods–to–go. Then, the optimality equations are given by

$$V_t(n_t^1, n_t^2, r_t, OF_t) = \min_{\substack{0 \leq a^1 \leq n_t^1 \\ 0 \leq a^2 \leq n_t^2 \\ a^1 + a^2 \leq r_t}} \left\{ \begin{array}{l} b^1 a^2 + \hat{b}^1(n_t^1 - a^1) + \hat{b}_t^2(n_t^2 - a^2) \\ \qquad + \mathsf{E}_{n_{t-1}^N} \mathsf{E}_{OF_{t-1}|OF_t} V_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1}) \end{array} \right\} \tag{3.1}$$

where $n_{t-1}^1 = a^2 + \alpha_{11}(n_t^1 - a^1) + \beta_1 n_{t-1}^N$, $n_{t-1}^2 = \alpha_{12}(n_t^1 - a^1) + \alpha_{22}(n_t^2 - a^2) + \beta_2 n_{t-1}^N$ and $r_{t-1} = r_t - a^1 - a^2 + (OF_t - OF_{t-1})$. In equation (3.1), $\hat{b}^1 = \alpha_{11}b^1 + \alpha_{12}b^2$, $\hat{b}_t^2 = \alpha_{22}b^2 + \alpha_{2E}l_t^E$, and a^1 and a^2 represent the allocations to states 1 and state 2 respectively. The optimal allocation levels are denoted by a^{1*} and a^{2*}. The boundary condition is given by

$$\begin{aligned} V_0(n_0^1, n_0^2, r_0, 0) = (b^1 + t^1)\min\{r_0, n_0^2\} \\ + (\hat{b}^1 + \alpha_{11}t^1 + \alpha_{12}t^2)(n_0^1 - (r_0 - n_0^2)^+)^+ + (\hat{b}_0^2 + \alpha_{22}t^2)(r_0 - n_0^2)^+ \end{aligned} \tag{3.2}$$

In our analysis, we assume that no new patients enter the system at t=0. Relaxing this assumption would not alter any of our results. For convenience, also define

$$C_t(a^1, n_t^1, n_t^2, r_t, OF_t) = \left\{ \begin{array}{c} b^1 \min\{n_t^2, r_t\} + \hat{b}^1(n_t^1 - a^1) + \hat{b}_t^2\left(n_t^2 - \min\{n_t^2, r_t\}\right) \\ +\mathsf{E}_{n_{t-1}^N}\mathsf{E}_{OF_{t-1}|OF_t} V_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1}) \end{array} \right\} \tag{3.3}$$

In the above equation, a^2 is replaced with $\min\{n_t^2, r_t\}$ in the definitions of n_{t-1}^1, n_{t-1}^2 and r_{t-1} as well. Some intuitive properties of the function V_t can be readily proven. For example, for fixed n_t^1, n_t^2 and OF_t, V_t is monotone decreasing in r_t.

Our first key result is the joint convexity of V_t in state variables n_t^1, n_t^2 and r_t, for fixed OF_t. We prove this in Lemma 3. For brevity, we use the following notation in our analysis: $N_t = (n_t^1, n_t^2), S_t = (n_t^1, n_t^2, r_t)$. The proofs for all the results in this chapter can be found in Appendix B.

Lemma 3. *$V_t(S_t, OF_t)$ is jointly convex in S_t for fixed OF_t.*

In section 3.3, we use the convexity of V_t to establish several structural results concerning the optimal allocation policy.

3.3 Optimal Allocation Policy

Throughout our analysis, we will assume that the following intuitive condition holds: $b^1\delta + V_{t-1}\left(n_{t-1}^1 + \delta, n_{t-1}^2 - \alpha_{22}\delta, r_{t-1} - \delta, OF_{t-1}\right) \le \hat{b}_t^2\delta + V_{t-1}\left(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1}\right)$, i.e., it is never optimal in any period t to not serve state 2 patients in that period and instead, reserve that funding to treat potential future state 1 and/or state 2 patients. This implies that $a^{2*}(S_t, OF_t) = \min\{n_t^2, r_t\}$. Determining the optimal allocation level for state 1 patients is more challenging and will be the focus of our analysis going forward. The following theorem offers a first step in determining the optimal allocation policy in period t.

Theorem 5. *Fix OF_t. Then, given state vector S_t, the optimal allocation level for state 1 patients in period t is the following:*

1. *If $n_t^2 \geq r_t$, then $a^{1*} = 0$.*
2. *If $n_t^2 < r_t$, then $a^{1*} = \max\{a^1 : \frac{\partial C_t}{\partial a^1} \leq 0, 0 \leq a^1 \leq \min\{n_t^1, r_t - n_t^2\}\}$ where C_t is given by equation (3.3).*

Theorem 5, while offering a first–step in determining a^{1*}, demonstrates that the optimal allocation policy is state–dependent. The state–dependency significantly complicates the computation of the optimal policy due to the so–called "curse of dimensionality", especially for long planning horizons. In an effort to simplify the computation of the optimal policy, we explore additional structural properties of the optimal policy that would help narrow down the search space for a^{1*}.

3.3.1 Monotonicity of the Optimal Policy

In Theorem 5, we established that a^{1*} is a function of $S_t = (n_t^1, n_t^2, r_t)$. The following proposition demonstrates that the optimal allocation level a^{1*} is monotone increasing in $S_{t\setminus\{n_t^2\}} = (n_t^1, r_t)$ for fixed n_t^2. We refer to a state vector as increasing if all components of the vector are at least weakly increasing.

Proposition 3. *For fixed n_t^2 and OF_t, $a^{1*}(S_{t\setminus\{n_t^2\}}, n_t^2, OF_t)$ is (at least weakly) increasing in $S_{t\setminus\{n_t^2\}}$.*

The monotonicity of a^{1*} established in Proposition 3 can be exploited to reduce the computational effort required to identify the optimal allocation policy for a range of values of r_t. The following corollary collects the relevant results.

Corollary 1. *Let N_t and OF_t be fixed. Then, the following results regarding the optimal allocation policy hold.*

1. *Let $r_t > n_t^1 + n_t^2$ and suppose that it is optimal to allocate $a^2 = n_t^2$ and $a^1 = n_t^1$. Then, $a^2 = n_t^2$ and $a^1 = n_t^1$ are also optimal for every $\hat{r}_t \geq r_t$.*

2. *Let $n_t^2 \leq r_t \leq n_t^1 + n_t^2$ and suppose that it is optimal to allocate $a^2 = n_t^2$ and $a^1 = r_t - n_t^2$. Then, $a^2 = n_t^2$ and $a^1 = \hat{r}_t - n_t^2$ are optimal for all $\hat{r}_t$ such that $n_t^2 \leq \hat{r}_t \leq r_t \leq n_t^1 + n_t^2$.*

3. *Let $r_t > n_t^1 + n_t^2$ and suppose that it is optimal to allocate $a^2 = n_t^2$ and $a^1 < n_t^1$. Then, $a^2 = n_t^2$ and $\hat{a}^1 \leq a_1 < n_t^1$ are optimal for all $\hat{r}_t$ such that $n_t^1 + n_t^2 < \hat{r}_t \leq r_t$.*

4. *Let $n_t^2 \leq r_t \leq n_t^1 + n_t^2$ and suppose that it is optimal to allocate $a^2 = n_t^2$ and $a^1 < r_t - n_t^2 \leq n_t^1$. Then, $a^2 = n_t^2$ and $\hat{a}^1 \leq \min\{a^1, \hat{r}_t - n_t^2\}$ are optimal for all $\hat{r}_t$ such that $n_t^2 \leq \hat{r}_t \leq r_t \leq n_t^1 + n_t^2$.*

Proposition 3 and Corollary 1 could prove to be useful in narrowing the search space for a^{1*}. However, both results, which demonstrate the monotonicity of a^{1*} with respect to r_t, are somewhat restrictive in terms of their applicability due to the requirement that OF_t be fixed. What could potentially be more useful is to characterize how a^{1*} changes when we increase r_t by δ, while simultaneously decreasing OF_t by the same amount. To see why such a result could be useful, notice from equation (3.3) that, to determine a^{1*}, we are essentially looking for a value of a^1, $0 \leq a^1 \leq \min\{n_t^1, (r_t - n_t^2)^+\}$, that minimizes the expression $\hat{b}^1(n_t^1 - a^1) + \mathsf{E}_{n_{t-1}^N} \mathsf{E}_{OF_{t-1}|OF_t} V_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1})$. In order to evaluate $\mathsf{E}_{n_{t-1}^N} \mathsf{E}_{OF_{t-1}|OF_t} V_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1})$, we need to compute $a^{1*}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1})$ for different realizations of n_{t-1}^N and OF_{t-1}. However, given the relation $r_{t-1} = r_t - a^1 - a^2 + (OF_t - OF_{t-1})$, it is easy to see that an increase (decrease) in OF_{t-1} is always associated with a corresponding decrease (increase) in r_{t-1}. Hence, if we could characterize how a^{1*} changes when we increase r_{t-1} by δ, while simultaneously decreasing OF_{t-1} by the same amount, then $a^{1*}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1})$ for one particular realization of OF_{t-1} could be used to narrow down the search space for $a^{1*}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1})$ for other possible realizations of OF_{t-1}.

In order to establish structural results concerning a^{1*} while simultaneously increasing r_t

and decreasing OF_t by the same amount, we require some additional assumptions regarding the funding inflow. We discuss the necessary conditions and associated results in Section 3.3.2.

3.3.2 Additional Monotone Properties of the Optimal Policy

Define a new state variable $TF_t = r_t + OF_t$. TF_t reflects the total funding that is available to treat patients who are already in the system, and new patients who might enter the system between period t and the end of the planning period. Now, rewriting equation (3.1) in terms of TF_t, we have

$$J_t(n_t^1, n_t^2, r_t, TF_t) = \min_{\substack{0 \le a^1 \le n_t^1 \\ 0 \le a^2 \le n_t^2 \\ a^1 + a^2 \le r_t}} \left\{ \begin{array}{l} b^1 a^2 + \hat{b}^1(n_t^1 - a^1) + \hat{b}_t^2(n_t^2 - a^2) \\ \qquad +\mathsf{E}_{n_{t-1}^N} \mathsf{E}_{OF_{t-1}|OF_t} J_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, TF_{t-1}) \end{array} \right\} \tag{3.4}$$

where TF_{t-1}=$TF_t - a^1 - a^2$ and n_{t-1}^1, n_{t-1}^2 and r_{t-1} are the same as defined earlier following equation (3.1). Let $f_{t-1} = OF_t - OF_{t-1}$ denote the funding received at the beginning of period $t-1$. Notice that f_{t-1} is a function of OF_t, the amount that is outstanding as of period t. Then we can rewrite equation (3.4) in terms of f_{t-1} as

$$\begin{aligned} & J_t(n_t^1, n_t^2, r_t, TF_t) \\ = & \min_{\substack{0 \le a^1 \le n_t^1 \\ 0 \le a^2 \le n_t^2 \\ a^1 + a^2 \le r_t}} \left\{ \begin{array}{l} b^1 a^2 + \hat{b}^1(n_t^1 - a^1) + \hat{b}_t^2(n_t^2 - a^2) \\ +\mathsf{E}_{n_{t-1}^N} \mathsf{E}_{TF_t - r_t - f_{t-1}|TF_t - r_t} J_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, TF_{t-1}) \end{array} \right\} \end{aligned} \tag{3.5}$$

In the above equation, $r_{t-1} = r_t - a^1 - a^2 + (OF_t - OF_{t-1}) = r_t - a^1 - a^2 + f_{t-1}$. For convenience, also define

$$\tilde{C}_t(a^1, n_t^1, n_t^2, r_t, TF_t) = \left\{ \begin{array}{l} b^1 \min\{n_t^2, r_t\} + \hat{b}^1(n_t^1 - a^1) + \hat{b}_t^2 \left(n_t^2 - \min\{n_t^2, r_t\}\right) \\ +\mathsf{E}_{n_{t-1}^N} \mathsf{E}_{TF_t - r_t - f_{t-1} | TF_t - r_t} J_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, TF_{t-1}) \end{array} \right\} \tag{3.6}$$

In equation (3.6), a^2 is replaced by $\min\{n_t^2, r_t\}$ in the definitions of n_{t-1}^1, n_{t-1}^2 and r_{t-1} as well. Before we present the structural results established in this section, let us introduce the assumptions concerning f_{t-1} that we use to prove the results.

$$\hat{r}_t + f_{t-1} \mid_{TF_t - \hat{r}_t} \geq_{st} \; r_t + f_{t-1} \mid_{TF_t - r_t} \text{ for fixed } TF_t \text{ and } \hat{r}_t \geq r_t \tag{3.7}$$

$$\{f_{t-1} \mid_{TF_t - r_t}, TF_t - r_t \; \in \mathbb{R}\} \; \in \; \text{SSCV} \tag{3.8}$$

Condition (3.7) states that the total funding available on–hand to treat patients in periods t and t-1 is stochastically decreasing in the outstanding amount at the beginning of period t. Condition (3.8) states that f_{t-1} belongs to the class of distributions satisfying the 'strong stochastically concave' (SSCV) property.

Definition 4. *(Shaked and Shanthikumar 2007) Let $\{X(\theta),\; \theta \; \in \Theta\}$ be a family of random variables. The family $\{X(\theta),\; \theta \; \in \Theta\}$ satisfies the strong stochastically concave property, denoted by SSCV, if there exist $\{\hat{X}(\theta),\; \theta \; \in \Theta\}$ such that $\hat{X}(\theta) =_{st} X(\theta)$ for each $\theta \; \in \Theta$ and $\hat{X}(\theta)$ is concave in θ almost surely.*

While the SSCV assumption might appear to be restrictive at first sight, the interpretation of the requirement that f_{t-1} be strong stochastic concave in OF_t is quite intuitive and natural in our setting: the funding received at the beginning of t-1 is increasing and concave in the amount that is outstanding as of period t. Recall that OF_t indicates the outstanding amount as of the beginning of period t, after receiving funding, if any, at the beginning of that period.

In Section 3.2.3, we proved that V_t is jointly convex in $S_t = (n_t^1, n_t^2, r_t)$ for fixed OF_t. That result, however, neither implies nor guarantees the joint convexity of J_t in (S_t, TF_t). A property that is *sufficient* to guarantee the convexity of J_t is the strong stochastic concavity of f_{t-1} in

OF_t. We prove this in Lemma 4.

Lemma 4. *If (3.8) holds, then $J_t(S_t, TF_t)$ is jointly convex in S_t and TF_t.*

The joint convexity established in Lemma 4 can be used to establish the following monotonicity result concerning a^{1*}. Again, recall that $S_{t\backslash\{n_t^2\}} = (n_t^1, r_t)$.

Proposition 4. *If (3.7) and (3.8) hold, then, for fixed n_t^2 and TF_t, $a^{1*}(S_{t\backslash\{n_t^2\}}, n_t^2, TF_t)$ is (at least weakly) increasing in $S_{t\backslash\{n_t^2\}}$.*

Before proceeding further, it is illustrative to compare the monotonicity properties implied by Propositions 3 and 4. Proposition 3 demonstrates that a^{1*} is monotone increasing in r_t for fixed n_t^1, n_t^2 and OF_t. Proposition 4 expands the scope of Proposition 3 by requiring only the total remaining funding, TF_t, to be fixed while allowing for simultaneous (equal and opposite) changes in r_t and OF_t. The following corollary, derived from Proposition 4, is analogous to Corollary 1.

Corollary 2. *Let N_t and TF_t be fixed. Then, the following results regarding the optimal allocation policy hold.*

1. *Let $r_t > n_t^1 + n_t^2$ and suppose that it is optimal to allocate $a^2 = n_t^2$ and $a^1 = n_t^1$. Then, $a^2 = n_t^2$ and $a^1 = n_t^1$ are also optimal for every $\hat{r}_t \geq r_t$.*

2. *Let $n_t^2 \leq r_t \leq n_t^1 + n_t^2$ and suppose that it is optimal to allocate $a^2 = n_t^2$ and $a^1 = r_t - n_t^2$. Then, $a^2 = n_t^2$ and $a^1 = \hat{r}_t - n_t^2$ are optimal for all $\hat{r}_t$ such that $n_t^2 \leq \hat{r}_t \leq r_t \leq n_t^1 + n_t^2$.*

3. *Let $r_t > n_t^1 + n_t^2$ and suppose that it is optimal to allocate $a^2 = n_t^2$ and $a^1 < n_t^1$. Then, $a^2 = n_t^2$ and $\hat{a}^1 \leq a_1 < n_t^1$ are optimal for all $\hat{r}_t$ such that $n_t^1 + n_t^2 < \hat{r}_t \leq r_t$.*

4. *Let $n_t^2 \leq r_t \leq n_t^1 + n_t^2$ and suppose that it is optimal to allocate $a^2 = n_t^2$ and $a^1 < r_t - n_t^2 \leq n_t^1$. Then, $a^2 = n_t^2$ and $\hat{a}^1 \leq \min\{a^1, \hat{r}_t - n_t^2\}$ are optimal for all $\hat{r}_t$ such that $n_t^2 \leq \hat{r}_t \leq r_t \leq n_t^1 + n_t^2$.*

Propositions 3 and 4, and the implied Corollaries 1 and 2, combined, could significantly reduce the computational effort required to calculate the optimal allocation level a^{1*} in every period. However, despite the (potential) simplifications offered by these results, we expect the computation of the optimal policy to be challenging and computationally–intensive, especially for problems with long planning horizons. This motivated us to look at the use of heuristics in order to handle realistic–size problems in a reasonable amount of time.

3.4 Heuristics

We consider two heuristics for our model. The first heuristic, which we call FCFS, reflects the allocation policy commonly used by global public health managers. The second heuristic, which we refer to as PNS (standing for probability of no shortfall), computes the probability that all state 2 patients will treated in the next period and uses that information to make the allocation decision for state 1 patients in the current period. In what follows, we discuss the two heuristics in greater detail.

3.4.1 Heuristic FCFS

The FCFS heuristic is very simple to understand and easy to implement. In every period, $a^2 = \min\{r_t, n_t^2\}$ and $a^1 = \min\{(r_t - n_t^2)^+, n_t^1\}$, i.e., after treating state 2 patients, treat as many state 1 patients as possible with the funding available on–hand. Notice that the FCFS heuristic is naive since it does not take into account future funding availability for state 2 patients when making the allocation decision for state 1 patients in the current period. Nevertheless, it remains a popular approach in the humanitarian health sector and we are interested in evaluating its performance relative to the PNS heuristic and the optimal policy.

3.4.2 Heuristic PNS

The PNS heuristic is based on the calculation of the probability that all state 2 patients would be treated in period t-1, given that the allocation for state 1 patients in period t is a^1. Of course, $0 \le a^1 \le (r_t - n_t^2)^+$. When $r_t \ge n_t^2$, for any a^1, this probability can be easily computed as shown below.

$$
\begin{aligned}
&Pr\{\text{all state 2 patients will be treated in } t-1\}\ |_{a^1}\\
&\qquad = Pr\{r_{t-1} \geq n^2_{t-1}\}\ |_{a^1}\\
&\qquad = Pr\{r_t - n^2_t - a^1 + (OF_t - OF_{t-1}) \geq \alpha_{12}(n^1_t - a^1) + \beta_2 n^N_{t-1}\}\\
&\qquad = Pr\{OF_{t-1} \leq r_t - n^2_t - a^1 + OF_t - \alpha_{12}(n^1_t - a^1) - \beta_2 n^N_{t-1}\}
\end{aligned}
$$

Then, the PNS heuristic chooses the allocation level a^1 as follows: $\max\{a^1 : Pr\{r_{t-1} \geq n^2_{t-1}\}\ |_{a^1} \geq K, 0 \leq a^1 \leq (r_t - n^2_t)^+\}$ where K is a threshold value between 0 and 1. If no such a^1 exists, then a^1=0. In our numerical experiments, we optimize over the range [0,1] to determine the optimal value of K.

The PNS heuristic is appealing to us for two reasons. First, notice that the PNS heuristic is the same as FCFS if we set K=0. Thus, by optimizing over the set of possible values for K, the PNS heuristic offers a natural and intuitive way to improve upon the performance of the FCFS heuristic.

To understand the other reason why we are interested in the PNS heuristic, consider expressions (3.9) and (3.10), which represent the minimum expected disease–adjusted life months lost corresponding to allocations a^1 and $a^1 - \delta$ in period t. We are only interested in small values of δ since our aim is to capture the impact of making incremental changes to the allocation level a^1. In writing these equations, we assume that $r_t \geq n^2_t$ since it is only under this situation that the question of how much to allocate to state 1 patients arises.

$$b^1 n^2_t + \hat{b}^1(n^1_t - a^1) + \mathsf{E}_{n^N_{t-1}} \mathsf{E}_{OF_{t-1}|OF_t} V_{t-1}(n^1_{t-1}, n^2_{t-1}, r_{t-1}, OF_{t-1}) \tag{3.9}$$

$$b^1 n^2_t + \hat{b}^1(n^1_t - a^1 + \delta) + \mathsf{E}_{n^N_{t-1}} \mathsf{E}_{OF_{t-1}|OF_t} V_{t-1}(n^1_{t-1} + \alpha_{11}\delta, n^2_{t-1} + \alpha_{12}\delta, r_{t-1} + \delta, OF_{t-1}) \tag{3.10}$$

In the above expressions, $n^1_{t-1} = n^2_t + \alpha_{11}(n^1_t - a^1) + \beta_1 n^N_{t-1}$, $n^2_{t-1} = \alpha_{12}(n^1_t - a^1) + \beta_2 n^N_{t-1}$ and $r_{t-1} = r_t - n^2_t - a^1 + OF_t - OF_{t-1}$. Notice that when we look at the difference between (3.9) and (3.10), we are essentially considering the expected value of $-\hat{b}^1\delta + V_{t-1}(n^1_{t-1}, n^2_{t-1}, r_{t-1}, OF_{t-1}) -$

$V_{t-1}(n_{t-1}^1 + \alpha_{11}\delta, n_{t-1}^2 + \alpha_{12}\delta, r_{t-1} + \delta, OF_{t-1})$, with the possibility of either $r_{t-1} \geq n_{t-1}^2$ or $r_{t-1} < n_{t-1}^2$, i.e., funding available at the beginning of period t-1 may or may not be sufficient to treat all state 2 patients in that period. Now, if $V_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1}) - V_{t-1}(n_{t-1}^1 + \alpha_{11}\delta, n_{t-1}^2 + \alpha_{12}\delta, r_{t-1} + \delta, OF_{t-1}) \leq 0$ when $r_{t-1} \geq n_{t-1}^2$, then we can reasonably expect the difference between (3.9) and (3.10) to be negative if the probability of $r_{t-1} \geq n_{t-1}^2$ is greater than some threshold value. Let us analyze the difference $V_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1}) - V_{t-1}(n_{t-1}^1 + \alpha_{11}\delta, n_{t-1}^2 + \alpha_{12}\delta, r_{t-1} + \delta, OF_{t-1})$ when $r_{t-1} \geq n_{t-1}^2$.

$$\begin{aligned} & V_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1}) = b^1 n_{t-1}^2 \\ & + \min_{\substack{0 \leq a^1 \leq n_{t-1}^1 \\ a^1 \leq r_t - n_{t-1}^2}} \left\{ \begin{array}{l} \hat{b}^1(n_{t-1}^1 - a^1) \\ \qquad\qquad + \mathsf{E}V_{t-2}(n_{t-2}^1, n_{t-2}^2, r_{t-2}, OF_{t-2}) \end{array} \right\} \end{aligned} \tag{3.11}$$

and

$$\begin{aligned} & V_{t-1}(n_{t-1}^1 + \alpha_{11}\delta, n_{t-1}^2 + \alpha_{12}\delta, r_{t-1} + \delta, OF_{t-1}) = b^1(n_{t-1}^2 + \alpha_{12}\delta) \\ & + \min_{\substack{0 \leq a^1 \leq n_{t-1}^1 + \alpha_{11}\delta \\ a^1 \leq r_t - n_{t-1}^2 + \alpha_{11}\delta}} \left\{ \begin{array}{l} \hat{b}^1(n_{t-1}^1 + \alpha_{11}\delta - a^1) \\ + \mathsf{E}V_{t-2}(n_{t-2}^1 + \Delta_1, n_{t-2}^2 + \Delta_2, r_{t-2} + \Delta_3, OF_{t-2}) \end{array} \right\} \end{aligned} \tag{3.12}$$

In equations (3.11) and (3.12), the expectation is taken with respect to n_{t-2}^N and $OF_{t-2}|_{OF_{t-1}}$, and in (3.12), $\Delta_1 = (\alpha_{12} + \alpha_{11}^2)\delta$, $\Delta_2 = \alpha_{12}\alpha_{11}\delta$, $\Delta_3 = \alpha_{11}\delta$. If α_{11}=0, then, clearly, (3.11)-(3.12)$\leq$0. When $\alpha_{11} > 0$, let $\hat{a}^1$ be the optimal solution for expression (3.12). Now, if $\hat{a}^1 \geq \alpha_{11}\delta$, then using $\hat{a}^1 - \alpha_{11}\delta$ as a solution for expression (3.11) again yields $V_{t-1}(n_{t-1}^1, n_{t-1}^2, r_{t-1}, OF_{t-1}) \leq V_{t-1}(n_{t-1}^1 + \alpha_{11}\delta, n_{t-1}^2 + \alpha_{12}\delta, r_{t-1} + \delta, OF_{t-1})$. While it is not possible to guarantee that $\hat{a}^1 \geq \alpha_{11}\delta$ will always hold, it is clear that the possibility of the condition holding increases as the value of α_{11} goes down. This suggests that the PNS heuristic may perform well for low values of α_{11} but the performance may be sensitive to α_{11}. In Section 3.6, we test this hypothesis and more broadly, evaluate the performance of the two heuristics relative to the optimal policy.

So far, we have focused on answering the first of our two main research questions: how to optimally allocate funding between the two health states in every period, taking into account the current funding availability and future financial inflows. In the next section, we turn our attention to the second research question: what is the impact of system parameters and funding changes on the number of disease–adjusted life months lost? Specifically, we focus on the impact of funding changes.

3.5 Impact of Funding Changes

In this section, we are mainly interested in exploring the impact of funding timing along two dimensions: 1. changes to the expected time of receipt of funds 2. changes to the variability in funding timing.

3.5.1 Expected Funding Timing

Consider funding scenarios 1 and 2 such that

$$OF^2_{t-1} \mid_{OF_t=i} \geq_{st} OF^1_{t-1} \mid_{OF_t=i} \;\; \forall \; i \in \mathbb{R}_+ \text{ and } \forall \; t \in \{3, ..., T\} \tag{3.13}$$

where $\geq_{st}$ means first–order stochastic dominance. Condition (3.13) implies that for any given value of outstanding funding at the beginning of period t, the outstanding funding at the beginning of period $t-1$ is (stochastically) larger under funding scenario 2, i.e., *funds arrive earlier under scenario 1* with probability 1 (and hence in expectation as well). Furthermore, assume that condition (3.7), introduced in Section 3.3.2, continues to hold under both funding scenarios. We denote the value functions associated with funding scenarios 1 and 2 by V_t^1 and V_t^2 respectively. The following proposition demonstrates that the increased allocation flexibility afforded by the (stochastically) early arrival of funds results in lower loss of disease–adjusted life months.

Proposition 5. *If condition (3.7) holds for both funding scenarios, and condition (3.13) holds, then* $V_t^2(S_t, OF_t) \geq V_t^1(S_t, OF_t)$.

3.5.2 Variability in the Funding Timing

Next, we investigate the impact of variability in the funding timing. When examining the role of variability, we look at the variance of the random variable $OF_{t-1} \mid_{OF_t}$. First, we compare funding scenarios 1 and 2 where the variance of $OF_{t-1} \mid_{OF_t}$ is higher under scenario 2 relative to scenario 1, but the expected value remains the same across the two scenarios.

Higher Variability with Equal Means: Consider funding scenarios 1 and 2 such that

$$OF^2_{t-1} \mid_{OF_t=i} \geq_{cvx} OF^1_{t-1} \mid_{OF_t=i} \;\; \forall\; i \in \mathbb{R} \text{ and } \forall\; t = 3, ..., T \tag{3.14}$$

where $\geq_{cvx}$ represents the convex ordering. Condition (3.14) implies that the variability of OF^2_{t-1} is higher than the variability of OF^1_{t-1} but $\mathsf{E}(OF^2_{t-1} \mid_{OF_t=i}) = \mathsf{E}(OF^1_{t-1} \mid_{OF_t=i})$. Thus the convex ordering helps in isolating the impact of variability while keeping the expected value the same. Also, assume that conditions (3.7) and (3.8), introduced in Section 3.3.2, continue to hold under both funding scenarios. Then, the following result shows that increased variability in the outstanding funding at the beginning of a period leads to a higher loss of disease–adjusted life months.

Proposition 6. *If conditions (3.7) and (3.8) hold for funding scenarios 1 and 2, and condition (3.14) holds, then* $V^2_t(S_t, OF_t) \geq V^1_t(S_t, OF_t)$.

Higher Variability with Unequal Means: Next, we compare funding scenarios 1 and 2 where the variance of $OF_{t-1} \mid_{OF_t}$ is again higher under scenario 2 relative to scenario 1, but without the restriction that the expected value of $OF_{t-1} \mid_{OF_t}$ remain the same across the two scenarios. In this case, we use the 'dispersive ordering' of random variables to characterize the impact of variability.

Definition 5. *(Shaked and Shanthikumar 2007) Let X and Y be two random variables with distribution functions F and G respectively. Let F^{-1} and G^{-1} be the right continuous inverses of F and G respectively, and assume that*

$$F^{-1}(\beta) - F^{-1}(\alpha) \leq G^{-1}(\beta) - G^{-1}(\alpha) \textit{ whenever } 0 < \alpha \leq \beta \leq 1.$$

Then, X is said to be smaller than Y in the dispersive order (denoted by $X \leq_{disp} Y$) .

The dispersive ordering leads naturally to a comparison of the variability of X and Y. In fact, $Y \geq_{disp} X$ implies that $Var(Y) \geq Var(X)$. The following property, connecting the $\leq_{disp}$ and $\leq_{st}$ orderings, is useful in understanding the impact of variability.

Property 2. *(Shaked and Shanthikumar 2007) Let X and Y be two random variables such that $X \sim (l_X, u_X)$ and $Y \sim (l_Y, u_Y)$, where l_X, u_X are the endpoints of the support of X and l_Y, u_Y are the endpoints of the support of Y. The following results hold.*

1. *If $l_X = l_Y > -\infty$, then $Y \geq_{disp} X \Rightarrow Y \geq_{st} X$.*
2. *If $u_X = u_Y < \infty$, then $Y \geq_{disp} X \Rightarrow Y \leq_{st} X$.*

Consider funding scenarios 1 and 2 such that $OF^1_{t-1} \mid_{OF_t=i} \sim (L_{t-1}(i), U^1_{t-1}(i))$ and $OF^2_{t-1} \mid_{OF_t=i}$ $\sim (L_{t-1}(i), U^2_{t-1}(i))$, i.e., $OF^n_{t-1} \mid_{OF_t=i}$ is distributed between $L_{t-1}(i)$ and $U^n_{t-1}(i)$, n=1,2, with both boundaries included in the support. Suppose that

$$U^2_{t-1}(i) > U^1_{t-1}(i) \ \forall \ i \in \mathbb{R}_+ \text{ and } \forall \ t \in \{3, ..., T\} \text{ and} \tag{3.15}$$

$$OF^2_{t-1} \mid_{OF_t=i} \geq_{disp} OF^1_{t-1} \mid_{OF_t=i} \ \ \forall \ i \in \mathbb{R}_+ \text{ and } \forall \ t = 3, ..., T \tag{3.16}$$

where $\geq_{disp}$ indicates stochastic ordering according to the dispersive order. Condition (3.15) states that the maximum possible outstanding funding at the beginning of period $t-1$ is higher under funding scenario 2. Condition (3.16) indicates that the outstanding funding at the beginning of period $t-1$ is more variable under funding scenario 2. When conditions (3.15) and (3.16) are combined, intuitively, it appears that funding scenario 2 would perform poorly in comparison to scenario 1. We confirm this intuition in Proposition 7.

Proposition 7. *If condition (3.7) holds for both funding scenarios, and conditions (3.15) and (3.16) hold, then $V^2_t(S_t, OF_t) \geq V^1_t(S_t, OF_t)$.*

Now, we consider an alternative funding scenario involving higher variability. Consider funding scenarios 1 and 2 such that $OF^1_{t-1} \mid_{OF_t=i} \sim (L^1_{t-1}(i), U_{t-1}(i))$ and $OF^2_{t-1} \mid_{OF_t=i} \sim$

$(L^2_{t-1}(i), U_{t-1}(i))$. Suppose that

$$L^2_{t-1}(i) < L^1_{t-1}(i) \ \forall \ i \in \mathbb{R}_+ \text{ and } \forall \ t \in \{3, ..., T\} \tag{3.17}$$

Condition (3.17) states that the least possible outstanding funding at the beginning of period $t-1$ is lower under funding scenario 2.

Proposition 8. *If condition (3.7) holds for both funding scenarios, and conditions (3.16) and (3.17) hold, then* $V^1_t(S_t, OF_t) \geq V^2_t(S_t, OF_t)$.

Propositions 7 and 8, viewed together, provide interesting insights into the impact of variability in funding timing. The results demonstrate that when $\mathsf{E}(OF_{t-1} \mid_{OF_t=i})$ does not remain the same, the effect of a change in the variability in funding timing is not easy to predict. In fact, the results dispel the commonly held notion that *variability is always bad.* Whether more variability is good or bad depends on whether the 'variability' is a favorable one as in condition (3.17) or an 'unfavorable' one as in (3.15).

3.6 Computational Study

The goal of our computational study is to (i) evaluate the performance of the FCFS and PNS heuristics relative to the optimal policy and (ii) to understand the impact of system parameters and funding changes on the number of disease–adjusted life months lost.

Experimental setup: For our computational study, we vary a variety of parameters while keeping other parameters fixed. First, we discuss about the parameters that are fixed throughout the computational study.

Per–period penalties: We use the disability weights published by WHO (see WHO 2004) to calculate the per–period penalties b^1 and b^2. One difficulty in estimating b^1 and b^2 is that WHO (2004) publishes only the "average" disability score (=0.053) associated with wasting from malnutrition without specifying the weights associated with moderate and severe acute malnutrition in calculating the average. One way of overcoming this difficulty is to consider combinations of the per–period penalties satisfying the expression $\lambda b^1 + (1-\lambda)b^2$=0.053 for different values of $0 \leq \lambda \leq 1$. However, in our pilot runs, we observed that the number of disease–adjusted life months lost over the planning period are not sensitive to changes in the per–period penalties since the bulk of the contribution to the life months lost comes from the terminal cost. Hence, we fix the per–period penalties to b^1=0.0486 and b^2=0.0632, corresponding to λ=0.67. We chose λ to be roughly 2/3 since the proportion of moderately malnourished children in the population is likely to be higher than the proportion of severely malnourished children.

Terminal costs: We estimate the terminal cost t^i associated with health state i, $i = 1, 2$ to be $mortality\ rate_i \times$ DALE × 12 where $mortality\ rate_i$ refers to the mortality rate in health state i and DALE is the disease–adjusted life expectancy. We use Bachmann (2009), a study of the cost–effectiveness of community–based therapeutic care of severe–acute malnutrition in Zambia, to guide our choice of $mortality\ rate_i$ and the value of DALE. Based on Bachmann (2009), the disease–adjusted life expectancy of a child who recovers from malnutrition is 33.3 years and $mortality\ rate_2$=0.181. There are conflicting views regarding $mortality\ rate_1$, since some studies in public health suggest that the mortality rate among children suffering from moderate acute malnutrition is the same as the overall under–five mortality rate while others suggest that mortality risk is slightly elevated in the presence of moderate acute malnutrition. Based on Chen at al. (1980), an influential work on assessing the impact of nutritional status on morality rates, we set $mortality\ rate_1$=0.0364, which is the overall under–five mortality rate in Zambia (Bachmann 2009).

New patient entry: Due to the computational complexity involved in determining the optimal policy, and to facilitate comparisons between the heuristics and the optimal policy, we make n_t^N, the number of new patients entering the system in any given period t, to be deterministic throughout the numerical study. We fix n_t^N=50 $\forall\ t = 1, 2, ..., T$. Garcia (2012) indicates that on average, 30% of children in Sub-Saharan Africa are malnourished. Based on this estimate, we assume incidence rates of β^1=0.2 and β^2=0.1 (totalling to 30%), since the proportion of children who are severely malnourished is likely to be less than the proportion of moderately malnourished children.

So far, we focused on the parameters that are fixed throughout our numerical study: the per–period penalties, terminal costs and new patient entry. Next, we proceed to discuss about the parameters that are varied.

Transition rates: In our computational study, we vary both transition rates α_{11} and α_{22}. We considered three values of α_{11}, α_{11}=0.2, 0.5 and 0.8 and three values for α_{22}, α_{22}=0.2, 0.5 and 0.8.

Funding: We assume that the total promised funding is received in $m \leq T$ installments. For each T, we consider several values of m details of which are given in Table 3.1.

T	m
2	1,2
4	1,2,3,4
6	1,2,3,4,6
8	1,2,3,4,6,8
12	1,2,3,4,6,8,10,12
24	1,2,3,4,6,8,10,12,16,20,24

Table 3.1: Number of installments considered for each T

To capture the uncertainty in the funding timing, we vary the number of installments. Depending on the number of installments (m) received, the amount received in each installment varies (=total funding/m). We assume that, in each period, either zero or one installment is received. Therefore, for fixed T and m, the total number of possibilities in terms of the time of receipt of the m installments is $\frac{T!}{m!(T-m)!}$ and we assume that all possibilities are *equally likely*. Notice that, for a fixed T, as m increases from 1 to T, the variability in the funding received until any given period decreases and hence, the funding becomes more even and predictable. In the case of $m = T$, there is no uncertainty with respect to the funding timing.

We assume that the total funding received over the entire planning period is equal to funding level$\times(\beta^1 \times n_t^N + 2 \times \beta^2 \times n_t^N) \times T$. Notice that $(\beta^1 \times n_t^N + 2 \times \beta^2 \times n_t^N) \times T$ is the total funding required to completely cure the state 1 and state 2 patients who seek treatment over the entire planning horizon, assuming that funding is available when required, and there is no health state deterioration due to non–treatment. However, if state deteriorations occur due to non–receipt or shortage of funds in certain periods, $(\beta^1 \times n_t^N + 2 \times \beta^2 \times n_t^N) \times T$ may not be sufficient to completely cure all state 1 and state 2 patients who seek treatment, since there is no buffer funding available to meet the increased resource requirements due to state transitions.

In our study, we consider funding levels of 25%, 50%, 75%, 100%, 125% and 150%. At 25%, 50% and 75% funding levels, the total funding received is clearly not sufficient to completely cure the state 1 and state 2 patients who seek treatment over the planning horizon. At 100% funding level, when there is no uncertainty in funding, the total funding received would be sufficient to meet the state 1 and state 2 demand over the horizon. However, if there is funding uncertainty and funding is not received in any period, the total funding received may not be sufficient due to the increased resource requirements brought about by transitions to a more severe health state. At 125% and 150% funding levels, there is buffer funding available to deal with transitions to a more severe health state and hence, the total funding received might be sufficient to completely cure all patients who seek treatment, even in the presence of funding uncertainties.

Before we proceed to discuss the performance of the heuristics, we point out that all our reported results are in terms of the equivalent disease–adjusted life years (DALYs) lost rather than disease–adjusted life months lost.

3.6.1 Performance of the Heuristics

Table 3.2 displays the average and maximum percentage error of the FCFS and PNS heuristic relative to the optimal allocation policy. Notice that in Table 3.2, we only report the results for $T \leq 6$ since computing the optimal policy is time–intensive and solving for longer planning horizons was not practically feasible. We report the average percentage error for different values of α_{11} separately since the error percentages vary significantly with α_{11}. For each T, the numbers displayed in the table are averages over (3 α_{22} values)×(6 funding levels)×$I(T)$ problem instances where $I(T)$ represents the number of different m values considered for that particular T. For example, we consider m=1,2,3,4 and 6 for T=6 so $I(6)$=5.

From Table 3.2, we see that for very low values of α_{11} (i.e., low probability of remaining in state 1 in the absence of treatment), both the FCFS and PNS heuristics perform very well with an average error of less than 0.5% and maximum error less than 5%. This is not surprising since at α_{11}=0.2, 80% of the people in state 1 will deteriorate to state 2 in the absence of treatment and hence the expected benefit of not treating state 1 patients in the current period and instead

			T=2	T=4	T=6
$\alpha_{11} = 0.2$	FCFS	Avg	0.02	0.01	0.43
		Max	0.09	0.03	4.84
	PNS	Avg	0.02	0.01	0.35
		Max	0.09	0.03	4.84
$\alpha_{11} = 0.5$	FCFS	Avg	0.20	0.17	1.24
		Max	2.09	1.56	11.53
	PNS	Avg	0.03	0.02	0.70
		Max	0.13	0.26	11.53
$\alpha_{11} = 0.8$	FCFS	Avg	1.44	3.41	6.10
		Max	16.65	23.08	20.26
	PNS	Avg	0.05	0.83	2.42
		Max	0.25	9.31	10.85

Table 3.2: Performance of the heuristics: % error relative to the optimal allocation policy

		T=2	T=4	T=6	T=8	T=12	T=24
$\alpha_{11} = 0.2$	Avg	0.00	0.00	0.08	0.05	0.00	0.00
	Max	0.00	0.00	0.48	0.15	0.03	0.00
$\alpha_{11} = 0.5$	Avg	0.17	0.14	0.55	0.23	0.10	0.00
	Max	1.04	0.84	2.02	0.69	0.32	0.01
$\alpha_{11} = 0.8$	Avg	1.39	2.50	3.58	2.35	1.30	0.25
	Max	8.31	10.95	11.94	6.45	3.11	0.42

Table 3.3: Performance of the FCFS heuristic: % error relative to the PNS heuristic

rationing those resources for future state 2 patients is likely to be limited.

At a moderate value of α_{11} (=0.5), both the heuristics still perform well on average. However, the maximum error of 11.5% for T=6 suggests that performance of both the heuristics may be sensitive to system parameters and the particular funding scenario on hand.

As we increase α_{11} to 0.8, the PNS heuristic appears robust with an average error of less than 3%. The performance of the FCFS heuristic is reasonable although there is now a noticeable difference between the performance of the two heuristics. The maximum error percentage associated with the FCFS heuristic is also significantly higher when compared to the PNS heuristic.

To further compare the performance of the FCFS and PNS heuristics, consider Table 3.3 which displays the error percentage of the FCFS heuristic relative to the PNS heuristic for planning horizons of different lengths ranging from 2–24 periods. From the table, it is clear that for α_{11}=0.2 and 0.5, the performance of the FCFS heuristic closely matches that of the

PNS heuristic. However, for α_{11}=0.8, there is a significant difference in performance between the two heuristics both in terms of the average and maximum error percentage, especially for short–medium length planning horizons (4–8 periods). For very long horizons (T=24), the performance gap narrows down and there is no noticeable difference between the two heuristics.

Running times: Table 3.4 shows the average running times for the two heuristics and the optimal allocation policy. The numbers displayed here are averages over $I(T)\times 3\times 3\times 6$ problem instances corresponding to different combinations of T and m, α_{11}, α_{22} and funding levels respectively.

	T=2	T=4	T=6	T=8	T=12	T=24
Opt.Pol	0.04	1.38	3737.66	NA	NA	NA
FCFS	0.01	0.01	0.03	0.13	1.40	7.12
PNS	0.02	0.03	0.11	0.53	5.45	26.10

Table 3.4: Average running time in seconds

From the table, it is evident that the running times in case of the FCFS and PNS heuristics are almost negligent in comparison to the optimal policy. The running time under the optimal policy for $T > 6$ exceeded an hour during our pilot study. Thus, computing the optimal policy may not be practically feasible for real–size problems with longer planning horizons. The table also demonstrates that the PNS heuristic, albeit slower than the FCFS heuristic, is computationally very efficient with an average running time of less than 5 seconds for problems with 12 periods. For 24–period problems, given that the performance gap between the two heuristics is very small (see Table 3.3), it might be beneficial to use the faster FCFS heuristic.

Overall, we make the following observations based on our computational study: 1. Determining the optimal policy is time–intensive and it may not be practically feasible except for short planning horizons (2–4 periods). 2. For $\alpha_{11} \leq 0.5$, both FCFS and PNS heuristics perform well and the performance gap between the two heuristics is narrow. Hence, it might be beneficial to use the faster FCFS heuristic in this case. 3. For higher value of α_{11}, it is better to use the PNS heuristic, especially for short–medium length planning horizons (4-8 periods).

In what follows, we use the PNS heuristic to understand how the different system parameters and funding impact performance. We use the PNS heuristic since it is guaranteed to perform at least as well as the FCFS heuristic and it also appears to be more robust than the FCFS heuristic across a wide range of parameter values. Nevertheless, we verified that all the insights that we obtain using the PNS heuristic are consistent with the results for the FCFS heuristic as well as the optimal policy (for $T \leq 6$). Also recall that in case of the PNS heuristic, we optimize over the value of the threshold K, i.e., the optimal threshold value could be different for different problem instances. A question that naturally arises is: are the results obtained using the PNS heuristic consistent with policies using a fixed threshold value? We compare the results of the PNS heuristic with policies using a fixed threshold for different values of K (=0,0.2.0.4,0.6,0.8,1). Barring a few exceptions, the results presented in section 3.6.3 are generally consistent with the results for fixed threshold value policies. For the results in section 3.6.2, most of the insights are consistent, but there are few differences between the PNS heuristic and the fixed threshold value policies and we highlight the differences as and when they arise.

3.6.2 Impact of Transition Rates

Tables 3.5, 3.6 and 3.7 display the average DALYs lost for different combinations of (α_{11}, α_{22}) at 50%, 100% and 150% funding levels. We chose these three funding levels to present the insights since they are representative of what happens when we increase the funding level. Here, the averages are taken over 36 problem instances corresponding to different combinations of T (ranging from 2 to 24 periods) and m. (see Table 3.1.)

		α_{22}		
		0.2	0.5	0.8
α_{11}	0.2	4657.70	4581.87	3983.51
	0.5	4236.21	4157.07	3598.29
	0.8	3072.69	2973.91	2533.37

Table 3.5: Average DALYs lost for different combinations of (α_{11}, α_{22}) at 50% funding level

		α_{22}		
		0.2	0.5	0.8
α_{11}	0.2	1307.04	1325.70	1264.13
	0.5	1027.16	1014.91	936.06
	0.8	687.74	610.46	454.40

Table 3.6: Average DALYs lost for different combinations of (α_{11}, α_{22}) at 100% funding level

From the tables, we see that for fixed α_{22}, the DALYs lost are monotone decreasing in α_{11}. This is to be expected since a higher value of α_{11} means that a larger fraction of the

		α_{22}		
		0.2	0.5	0.8
α_{11}	0.2	737.25	636.92	439.90
	0.5	673.53	579.51	384.83
	0.8	542.66	465.26	303.80

Table 3.7: Average DALYs lost for different combinations of (α_{11}, α_{22}) at 150% funding level

untreated state 1 patients will continue to remain in state 1 and hence, the expected benefits from rationing funding to treat future state 2 patients is likely to be higher in this case.

Now let us consider what happens when we fix α_{11} and increase α_{22}. For medium and high values of α_{11} (α_{11}=0.5 and 0.8), the DALYs lost are monotone decreasing in α_{22} for all three funding levels, as one would expect. However, notice that for α_{11}=0.2, the monotone property holds at 50% and 150% funding levels but not at 100% funding level. At 100% funding level, the DALYs lost first increase and then decrease as we increase α_{22}. To ensure that the same pattern holds for other low values of α_{11}, we ran additional experiments with α_{11}=0.05, 0.10 and 0.15 and the results were consistent.

The reason behind this nonintuitive behavior with respect to α_{22} is subtle. First, let us look at what happens when we increase α_{22}. As we increase α_{22}, a lower fraction of the untreated state 2 patients exit the system and an increased fraction of them continue to remain in state 2. While this seems beneficial, it also has a potential downside — the increased fraction of untreated state 2 patients remaining in the system implies lesser funding availability for state 1 patients in the next period (since state 2 receives priority over state 1). This results in fewer state 1 patients getting completely cured in the next period, and at low values of α_{11}, a majority of the untreated state 1 patients deteriorate to state 2. The combination of low α_{11} and higher α_{22} leads to a temporary increase in the number of state 2 patients in the system, which, in turn, further reduces the funding availability for state 1 patients in the subsequent periods. Now let us reconcile this discussion of the effect of increasing α_{22} with the funding level.

At a low funding level (50%), the funding received is barely sufficient to meet new state 2 demand. Hence, irrespective of the value of α_{22}, only a low fraction of state 1 patients are treated and a majority of them transition to state 2. Given that only a small number of state 1 patients are treated to begin with, the reduction in the number of state 1 patients treated (and the associated increase in the number of state 2 patients) that is attributable to an increase in α_{22} is likely to be minimal. Thus, increasing α_{22} would have very little impact on the number of state 2 patients in the system but the fraction of untreated state 2 patients exiting the system decreases with α_{22}. This explains why the monotone property holds at 50% funding level.

Now let us consider 100% and 150% funding levels. If the system has buffer funding available (as in the case of 150% funding level) to handle the temporary increase in the number of state 2 patients, the buffer funding can be used to treat the additional state 2 patients without impacting the funding availability for state 1 patients in the subsequent periods. Hence, in this case, the monotone property holds as one might expect. However, in the absence of buffer funding (as with 100% funding level), the system does not have the flexibility to cope up with the increased state 2 demand, leading to a gradual build up of state 2 patients in the system (due to increasingly less funding availability for state 1). The rate of buildup of state 2 patients increases with α_{22}. Given limited total funding, the gradual build up of state 2 patients implies that, after a certain number of periods, an increasing number of state 2 patients are also left untreated. This complex interaction between the number of people completely cured, the magnitude of build up of state 2 patients in the system and the fraction of those patients exiting the system results in the increase followed by a decrease in the number of DALYs lost as we increase α_{22} at 100% funding level.

At 100% funding level, the insights regarding the effect of increasing α_{22} obtained using the PNS heuristic are also somewhat different from the insights obtained using the fixed threshold value heuristics. In case of the fixed threshold policies, we find that, depending on the value of K, the DALYs lost could exhibit a non–monotone pattern even for medium and high values of α_{22} (α_{22}=0.5, 0.8) whereas in case of the PNS heuristic, we observe the non–monotone pattern only for low values of α_{22}(=0.2).

In summary, we see that the effect of a change in the transition rates is on expected lines for the most part. However, as our results demonstrate, the relationship between transition rates and DALYs lost is not always straightforward due to the complex interaction between α_{11}, α_{22} and funding availability in the system. Hence, any conclusions regarding the impact of a change in the transition rates should be drawn only after taking into account the specific system parameters and the funding scenario on hand.

In addition to the impact of increasing α_{11} and α_{22} while keeping the other transition rate fixed, Tables 3.5, 3.6 and 3.7 also throw light onto the relative impact of α_{11} vis–a–vis α_{22}. From the tables, we see that the DALYs lost at (α_{11}, α_{22})=(0.2,0.8) are significantly higher when compared to the DALYs lost at (α_{11}, α_{22})=(0.8,0.2) for 50% and 100% funding levels while the opposite is true at 150% funding level, i.e., the relative impact of α_{11} and α_{22} change as we increase the funding level. The same insight holds for similar comparisons of (α_{11}, α_{22}). At 50% and 100% funding levels, there is no buffer funding to deal with state 1 patients transitioning into state 2 and hence α_{11} has a significant impact. With buffer funding available at 150% funding level, the transition rate from state 1 to state 2 becomes less impactful while α_{22} assumes significance since α_{22} determines the number of people exiting the system in periods when funding is not received.

3.6.3 Impact of Funding

Having investigated the impact of transition rates, we now look at how changes in funding impact performance. Specifically, we investigate the impact of number of funding installments (m) and funding level on the number of DALYs lost. First, we focus on the number of funding installments.

Number of Funding Installments: Recall that for fixed T, the funding received until any given period becomes more smooth and predictable as we increase the number of installments. When m=T, there is no uncertainty with respect to the funding timing.

Tables 3.8, 3.9 and 3.10 display the average DALYs lost for different combinations of T and m at 50%, 100% and 150% funding level respectively. The numbers displayed in the tables are averages over 3×3=9 problem instances corresponding to different values of α_{11} and α_{22}. From the tables, it is clear that for very short horizons ($T \leq 4$), the DALYs lost decrease with the number of installments, i.e., it is preferable to have a relatively smooth funding situation. For medium to long horizon lengths (N=6, 8, 12), at low funding levels (50%), it might be beneficial to receive the funding in fewer, lumpy installments while a smoother funding pattern is preferable at 100% and 150% funding levels. At 50% funding level, when there is no uncertainty in funding, the amount received in every period is sufficient to meet new state 2 demand and only a part of the new state 1 demand. The untreated state 1 patients deteriorate to state 2 and this leads to an increase in the number of state 2 patients who may not receive treatment. When funding is lumpy, there is a possibility that no funding is received until later in the horizon, in which case there would be a significant increase (relative to the case where there is no uncertainty in funding) in the number of state 2 patients who may not receive treatment. However, it is also equally likely that a majority of the funding is received early in the planning horizon, which could be used to prevent state 1 patients from deteriorating into state 2 in the earlier periods. In expectation, at 50% funding level, the benefits of lumpy funding outweigh the potential losses while the opposite is true at 100% and 150% funding levels.

For very long planning horizons (N=24), the optimal number of installments increases with the funding level. At low funding levels (50%), it might be beneficial to receive the funding in fewer, lumpy installments. The reasoning is similar to our earlier discussion for medium to long horizon lengths. At 100% funding level, notice that when there is no uncertainty in funding, the amount received in every period exactly matches the funding required to satisfy state 1 and state 2 demand. Hence, if it is possible to receive the funding in T evenly–spread installments, that should always be preferred. However, in situations where it is not possible to completely eliminate the uncertainty the funding, it is better to receive the funding in a moderate number of installments. The reasoning is as follows: with uncertain funding, the funding received in every period does not exactly match the funding required to meet state 1 and state 2 demand. This mismatch between funding received and funding required decreases with the number of installments. However, on the downside, the amount received in each installment also decreases as we increase the number of installments, resulting in less flexibility to deal with situations where funding is not received in the earlier periods. Given the trade–offs associated with increasing the number of installments, our results show that it is better to avoid the extremes and receive the funding in a moderate amount of installments. At 150% funding level, the additional amount received in each installment (when compared to 100% funding level) mitigates the potential value of flexibility afforded by fewer, lumpy installments and in this case, we see that it is preferable to have a relatively smooth funding situation. The same insight also holds at 125% funding level.

Overall, our analysis provides the following insights regarding the impact of uncertainty in funding: 1. For short planning horizons ($T \leq 4$), reducing the funding uncertainty is always beneficial. 2. For $T \geq 6$, the optimal number of installments increases with the funding level. The flexibility provided by fewer, lumpy installments is beneficial in under–financed systems (<100% funding level) while a smooth funding pattern is preferable in well–funded systems (≥100% funding level).

Interaction between funding level and uncertainty in funding: In this section, we explore the interaction between funding level and uncertainty in funding. Specifically, we are interested in answering questions of the following type: does altering the level of uncertainty in funding lead to better performance even if the overall funding received is lower?

Due to space considerations, we do not provide the tables illustrating how changes in the number of installments affect the DALYs lost at 25%, 75% and 125% funding levels, and discuss only the insights. Our computational results indicate that the number of DALYs lost at 25% (50%) funding level is significantly higher than the DALYs lost at all higher funding levels with one exception — for very long horizons (T=24), receiving 25% (50%) funding in fewer, lumpy installments ($m \leq 2$) is better than receiving 50% (75%) funding in a relatively smooth fashion ($m > 12$). At 75% funding level, again, the number of DALYs lost is significantly higher than the DALYs lost at all higher funding levels except in situations where funding at 100%, 125% and 150% funding are highly unpredictable (m=1) for very long horizons (T=24). Barring these exceptions, in general we see that at 25%, 50% and 75% funding levels, insufficient funding hurts performance and obtaining additional funding could significantly reduce the number of DALYs lost.

More interesting insights emerge when compare the DALYs lost at 100% and 150% funding levels. (A comparison between 100% and 125% funding level yields identical insights.) Notice that for $T \leq 4$, it is better to receive 100% funding in a less uncertain fashion (larger number of installments) rather than receive additional but more uncertain funding. However, the insights change when we increase the length of the planning horizon. For $T \geq 6$, we see that less uncertain funding at 100% funding level is preferable to relatively more unpredictable funding at 150% funding level only when the unpredictability in funding at 150% funding level is very high ($m \leq 2$). Otherwise, the buffer funding available at 150% funding level proves valuable even if it comes at the cost of increased uncertainty in funding.

A comparison of 125% and 150% funding levels shows that it is better to receive 125% funding in a less uncertain fashion rather than receive additional 25% funding but with more uncertainty (except when the funding at 150% funding level is relatively smooth ($m > 8$) for T=24). Intuitively, the reasoning is the following: at 125% funding level, there is already buffer funding available to deal with health state deteriorations occurring due to non–treatment in certain periods. Hence, in this case, the potential losses due to the increased uncertainty in funding outweighs the marginal benefit of receiving an additional 25% in funding.

In summary, we see that, insufficient funding generally hurts performance in under–financed systems (<100% funding level) and an additional influx of funds could bring significant benefits. At 100% funding level, receiving additional but more uncertain funding is beneficial only for medium to long planning horizons ($T \geq 6$) and at 125% funding level, where there is buffer funding available, additional funding should not be traded for more uncertain funding.

3.7 Conclusions and Managerial Insights

In this chapter, we study the problem of dynamic allocation of a scare resource, which in our case is donor–funding, to patients in different health states over a finite horizon. We characterize the optimal policy to be a state–dependent allocation policy and prove several monotonicity properties of the optimal policy that could help reduce the computational burden involved in determining the optimal policy. However, despite the potential simplifications offered by the monotonicity results, determining the optimal policy may not be practical for problems with long planning horizons. This motivated us to consider two heuristics that can handle real–size problems. Our computational results suggest that the PNS heuristic compares favorably with the optimal policy across a wide range of settings and it is easy to understand and use in practice. The FCFS heuristic also performs well for many problem instances but it appears to be less robust than the PNS heuristic, especially when α_{11} is large.

Our analysis also provides several interesting insights into the impact of uncertainty in funding timing. For example, our analytical results show that increased variability in the funding timing is not necessarily bad — in fact, we demonstrate that increased variability in funding can be of both types, favorable and unfavorable. Our computational results demonstrate that the impact of uncertainty in funding timing could be very different depending on the length of the planning horizon and the system funding level. Hence, it is important to take into account the system characteristics when making funding–related decisions so as to maximize the per–dollar impact of funding provided to global health programs. We believe that our model can be valuable tool in this regard by demonstrating the impact of alternate funding patterns to donors. In addition to uncertainty in funding, our work also throws light onto the impact of funding level. We find that when there is no buffer funding available, it might be beneficial to receive additional funding even at the cost of increased funding uncertainty but as the system funding level increases and buffer funding becomes available, the losses from the increased funding uncertainty outweigh the potential benefits from the additional funding.

		m										
		1	2	3	4	6	8	10	12	16	20	24
T	2	201.16	199.43	NA	NA	NA	NA	NA	NA	NA	NA	NA
	4	614.80	601.57	588.63	628.27	NA	NA	NA	NA	NA	NA	NA
	6	1121.20	1105.22	1160.82	1169.27	1272.75	NA	NA	NA	NA	NA	NA
	8	1657.27	1680.44	1727.88	1857.37	1934.41	2048.67	NA	NA	NA	NA	NA
	12	2801.92	2831.29	2978.89	3101.24	3480.18	3579.48	3719.29	3819.82	NA	NA	NA
	24	6512.39	6389.04	6572.20	6923.03	7562.93	8229.50	8845.18	9248.30	9491.48	9699.15	9823.97

Table 3.8: Impact of the number of installments on the average DALYs lost at 50% funding level

		m										
		1	2	3	4	6	8	10	12	16	20	24
T	2	45.15	0.04	NA	NA	NA	NA	NA	NA	NA	NA	NA
	4	238.78	90.29	33.39	0.08	NA	NA	NA	NA	NA	NA	NA
	6	522.52	271.92	161.40	105.22	0.12	NA	NA	NA	NA	NA	NA
	8	865.56	503.83	343.65	265.34	172.57	0.16	NA	NA	NA	NA	NA
	12	1663.13	1057.47	816.41	665.50	565.07	559.01	570.31	0.24	NA	NA	NA
	24	4492.92	3180.69	2451.23	2065.36	1748.45	1729.03	1805.46	2020.25	2637.17	2862.13	0.49

Table 3.9: Impact of the number of installments on the average DALYs lost at 100% funding level

		m										
		1	2	3	4	6	8	10	12	16	20	24
T	2	42.01	0.04	NA	NA	NA	NA	NA	NA	NA	NA	NA
	4	233.32	75.63	21.17	0.08	NA	NA	NA	NA	NA	NA	NA
	6	518.39	215.56	97.42	41.80	0.12	NA	NA	NA	NA	NA	NA
	8	862.47	399.62	218.08	119.85	28.69	0.16	NA	NA	NA	NA	NA
	12	1661.06	862.00	551.34	349.32	155.56	65.14	18.71	0.24	NA	NA	NA
	24	4491.90	2759.83	1843.00	1322.06	782.22	525.99	366.37	252.02	129.28	43.69	0.4862

Table 3.10: Impact of the number of installments on the average DALYs lost at 150% funding level

Chapter 4

Supply Vs. Demand Side Investment in Humanitarian Operations

4.1 Introduction

Over the last two decades, international aid commitments and more importantly, development assistance for health (DHA) has increased from \$2.5 billion in 1990 to over \$13 billion in 2005. Over that time, DHA's share of overall development assistance has increased from 4.6% in 1990 to almost 13% in 2005 (WHO 2007). However, despite the increased funding, the progress on health outcomes has been disappointing with many countries falling significantly short of the Millenium Development Goals (MDGs). While numerous factors contribute to the low aid effectiveness, supply– and demand–side factors have been identified as one of the key reasons behind low coverage levels and poor uptake of health services. Supply–side factors include effectiveness and efficiency of the distribution systems, availability of qualified, trained and motivated health personnel at service delivery points, and adequate planning, monitoring and oversight of public health supply chains. Viewed together, supply–side factors determine the availability of essential health supplies and services, and addressing constraints on the supply–side is critical to strengthening health systems. Some examples of supply–side investments that can improve the delivery of health services and products include employee training, investment in physical infrastructure like warehouses and cold–storage systems, health information systems, policy planning support, and general budget support to the Ministry of Health.

While supply–side factors focus on the availability of a product or service, demand–side factors look at the consumer angle of health delivery. Demand–side factors include community awareness of the availability and benefits of using a particular health commodity/service, and social, economic and cultural barriers to access. The combination of the different demand–side factors significantly impact the community uptake of the health services and hence, investing in initiatives that would raise community awareness and reduce the barriers to access is an important step in increasing aid effectiveness. Some examples of demand–side investments include community mobilization activities, informational workshops, and voucher schemes, subsidies and conditional cash transfer programs to remove the economic barriers to access.

Addressing both supply– and demand–side constraints are vital to increasing coverage levels and improving aid effectiveness, but the balance between the two is a delicate one. Stimulating community interest and providing better access would be of little use if there is insufficient capacity or shortage of essential health commodities. For example, in Sierra Leone, the demand for maternal health services increased dramatically after the launch of free public health services for pregnant women and children under the age of five (IPS 2010). However, the public health system faced severe challengies in implementing the free health care program due to a shortage of resources, creating unintended consequences. Similarly, focusing solely on improving the health delivery systems has not yielded the desired results since in many cases, clinics are not acceessible and the opportunity cost of seeking treatment is too high.

The problem of deciding how much to invest in the supply–side as opposed to the demand–side becomes even more significant in light of the fact that countries have a limited budget for interventions aimed at improving health outcomes. Often times, the external aid received is earmarked for procuring health commodities and funding to strengthen health systems and improve uptake typically comes from the host government. In some cases, external development assistance is also available for technical assistance and systems strengthening, but these funds cannot be used for procurement. In this context, a key question faced by in–country public health managers and policy makers is: given the limited funding available to strengthen supply systems and stimulate demand for a health service or product, what is the optimal mix of supply– and demand–side investments to maximize coverage?

We address this question in this chapter using a stylized model with stochastic demand to capture the impact of supply– and demand–side investments on the expected number of people served. We first consider a centralized model where a single entity (e.g., Ministry of Health at a host government) that manages the health program makes both the supply– and demand–side investments. For ease of reference, we refer to this entity as the 'principal' throughout the chapter. Note that in the centralized case, the principal must be physically present on ground to engage in community outreach and mobilization efforts. For a given budget level, we identify the principal's optimal mix of investments to maximize coverage and also present several results regarding how the investment mix changes with respect to the different supply– and demand side parameters as well as the demand distribution. Interestingly, we show that both the supply–side investment and the program coverage may not necessarily increase with expected demand. With respect to the demand variability, we provide a clean characterization that demonstrates that whether or not supply–side investments increase with demand variability depends solely on the value of a critical ratio that we identify in our model. While the supply–side investments may increase or decrease with variability, we show that program coverage, which is the objective of interest, always decreases with demand variability. We believe this result is valuable since it identifies an additional investment opportunity for humanitarian organizations in their quest to improve program coverage. In addition to the centralized setting, we also consider a decentralized model where the principal is not physically present on ground and as such cannot directly engage in demand–related activities. In this case, the principal invests only in the supply–side and contracts with a third–party (a private firm or a local NGO), who we refer to as the 'agent', to carry out community mobilization efforts on her behalf. The agent makes the demand–side investments and his objective is to maximize profits, which creates incentive issues that could lower coverage levels. Motivated by the growing interest in performance–based funding within the humanitarian sector, we explore the use of performance–based contracts to create incentives for the agent to invest in demand mobilization. We identify two types of contracts that guarantee that the expected coverage level under the decentralized case is at least as high as the centralized case.

4.2 Literature review

Our work is related to three streams of literature. The first stream related to our work is concerned with supply chain inefficiencies mainly arising due to inventory misplacement. Due to inventory misplacement, only a fraction of the ordered quantity is available to meet demand. Several authors have studied the benefits of implementing RFID in such settings and some papers have also looked at supply chain coordination issues that arise due to fixed and variable costs of implementing RFID (e.g., Rekik et al. 2007, Rekik et al. 2008, Camdereli and Swaminathan 2010). In our work also, only a fraction of the procured quantity is available to meet demand but we broadly attribute this to supply chain inefficiencies that could arise due to reasons different from inventory misplacement. Moreover, in our setting, the supply–side investment required to increase the fraction of the procured quantity available remains the same irrespective of the actual procurement quantity, which makes it different from the RFID literature, where the cost of tagging depends on the quantity procured. Furthermore, papers focusing on supply chain efficiencies typically consider demand as given while in our work, both supply and demand can be influenced through appropriate investments.

The second stream related to our work is concerned with the operations–marketing interface. In this stream of literature, the focus is on making effective operational (e.g., inventory replenishment, production scheduling) and marketing (e.g., product pricing, advertising budget, sales effort) decisions by explicitly considering the interplay between the two decisions. Examples of papers that focus on the interplay between inventory and sales effort decisions include Khouja and Robbins (2003), Heese and Swaminathan (2010), Wei and Chen (2011) and Xue et al. (2013). These papers differ from our work in two important ways. First, all the four papers jointly optimize the inventory and sales effort decisions with the objective of maximizing profits while our interest is in maximizing program coverage. Second, none of these papers consider budget constraints, which is often a major factor in non–profit operations.

Our work is also related to the literature on contracting and principal–agent models. Principal–agent models have a long history of applications in different fields. See Hart and Holmstron (1987) for an overview of the agency theory. Several papers in OM have used the principal–agent paradigm to study contracting issues in a variety of settings, focusing mostly on the problem of supply chain coordination, e.g., Tsay and Lovejoy (1999), Bassok and Anupindi (2008), Cohen and Agarwal (1998) and Cachon and Lariviere (2005) to name a few. Many papers within the marketing literature have used the agency theory framework to design optimal sales force compensation schemes to maximize profits, taking into account the demand environment and risk–seeking behavior of the sales personnel (see Basu et al. 1985 and Joseph and Thevaranjan 1998). In our work, we consider the problem of a principal contracting with an agent to engage in demand–enhancing activities. Our setting is unique in terms of the nature of the objective of the principal and agent (as opposed to the traditional setting of both being profit maximizers) and also the presence of budget constraints which dictate the feasibility of the contracts offered to the agent.

The rest of the chapter is organized as follows. In section 4.3, we present the centralized model where the principal makes both supply– and demand–side investments subject to a budget constraint. We determine the optimal investment levels for the supply and demand sides and analyze how the investment levels change with respect to a variety of supply– and demand–side parameters. Next we consider the decentralized setting and explore the use of two performance–based contracts to ensure that the coverage in the decentralized setting matches or exceeds the coverage in the centralized setting. The last section concludes the chapter.

4.3 Centralized model

We consider a simple one–period model with stochastic demand. As we discussed earlier, funding for procurement is typically earmarked and hence, for our purposes, we assume that the procurement quantity Q is fixed. However, due to supply chain inefficiencies, only α_0, $0 \leq \alpha_0 \leq 1$, fraction of the procured quantity is available to meet demand in the absence of any supply–side investment. The source of inefficiencies could include inventory loss and misplacement due to lack of supply chain visibility, products perished due to improper storage and handling, or shipment delays resulting in non–availability of products when they are needed. Of course, some of the misplaced items and late shipments could be available at a later date and if we assume that those items could be used to satisfy demand, then α_0 forms an upper–bound on the product availability fraction in the absence of any supply chain investments. The principal can influence the fraction of procured quantity available by investing s_i dollars in supply chain strengthening initiatives and we assume that the available fraction increases linearly in s_i, i.e., $min\left\{(\alpha_0 + \theta s_i), 1\right\} Q$ products are available if s_i dollars are invested in the supply–side. The parameter θ captures how supply chain investments translate into increased product availability.

We assume that the base demand D (without taking into account the effect of any efforts to increase demand) is a continuous random variable with mean μ, variance σ^2, pdf $f(.)$, cdf $F(.)$ and inverse cdf $\bar{F}$=1-F. The principal can influence demand by investing o_i dollars in demand–enhancing activities, and we assume that demand increases linearly in o_i resulting in a total demand of $o_e o_i$+D. Notice that o_e captures the effectiveness of demand–side investments.

The principal has a limited budget B available and the objective is to identify the optimal mix of supply– and demand–side investments so as to maximize expected coverage, i.e., the principal's problem is

$$Max \;\; \mathsf{E}\, min\{D + o_e o_i, min\left\{(\alpha_0 + \theta s_i), 1\right\} Q\} \;\; \text{subject to} \;\; s_i + o_i \leq B \tag{4.1}$$

Notice that for any given o_i, the expected coverage, $\mathsf{E}\, min\{D + o_e o_i, min\left\{(\alpha_0 + \theta s_i), 1\right\} Q\}$ is increasing in s_i and similarly, for any given s_i, the expected coverage is increasing in o_i. Therefore the constraint in expression (4.1) is always binding. Hence the problem can be reformulated as an optimization problem with only one decision variable as shown below.

$$Max \;\; \mathsf{E}\, min\{D + o_e(B - s_i), min\left\{(\alpha_0 + \theta s_i), 1\right\} Q\} \;\; \text{subject to} \;\; s_i \leq B \tag{4.2}$$

Using standard techniques, it can be easily verified that the expected coverage in (4.2) is concave in s_i. This implies that s_i^* is the maximum possible s_i, $0 \leq s_i \leq min\{B, (1 - \alpha_0)/\theta\}$, such that

$$F\left((\alpha_0 + \theta s_i)Q - o_e(B - s_i)\right) \leq \frac{\theta Q}{\theta Q + o_e}. \tag{4.3}$$

Notice that the optimality condition bears resemblance to the well–known "critical fractile" solution for the newsvendor problem. In the newsvendor problem, the critical ratio is a ratio of the underage cost to the sum of underage and overage costs. We have a similar structure in equation (4.3). The right hand side of (4.3) is the essentially a ratio of the effectiveness of supply–side investments to the effectiveness of supply– plus demand–side investments. In case of the newsvendor problem, the left–hand side of the optimality condition is the probability that the stochastic demand is less than the order quantity. Typically, the range of possible values for the order quantity is unrestricted and it does not depend on the underage and overage costs. For the investment problem that we study, the left–hand side of the optimality condition is the probability that the stochastic demand is less than the *quantity made available* to meet demand. Unlike the newsvendor problem, the quantity that can be made available is restricted by the procurement quantity Q and it is also dependent on both the supply–side and demand–side investment effectiveness. This linkage between the supply– and demand–side effectiveness and the quantity made available makes it more difficult (when compared to the newsvendor problem) to explicitly compute the optimal solution to equation (4.3), except for some special demand distributions like the uniform distribution.

In the next section, we use the characterization in equation (4.3) to understand how the optimal investment mix changes depending on the supply– and demand–side parameters and the procurement budget available to the program. Throughout the discussion, we focus on s_i^* since all the insights concerning o_i^* are just the opposite of the insights for s_i^*.

4.3.1 Impact of supply– and demand–side parameters

In this section, we analyze how the optimal investment mix changes with respect to α_0, θ, Q (supply–side parameters) and o_e, which is a demand–side parameter. In the next section, we will focus exclusively on the impact of demand D. Since the different parameters in our model interact in a complicated way, a complete analytical characterization of the sensitivity of s_i^* with respect to all four parameters is not possible. Hence, we first consider the parameter α_0 for which a full characterization is possible and present the results. Then, for the other three parameters, we present analytical results for the special case of uniform demand and then verify if the results can be generalized to other demand distributions like normal and exponential through a numerical study.

Impact of α_0

To understand the impact of α_0, consider the first derivative of the expected coverage with respect to s_i. Notice that $\bar{F}\left((\alpha_0 + \theta s_i)Q - o_e(B - s_i)\right)$ is (at least weakly) decreasing in α_0. Therefore, it follows that s_i^* is non–increasing in α_0. This result is on expected lines since the base product availability (in the absence of any supply–side investment) increases with α_0 and hence, a lower supply–side investment would be sufficient to maintain a pre–determined product availability level.

Now let us consider the other three parameters. In what follows, all the analytical results presented correspond to uniform demand $U \sim [0, D_u]$. In obtaining the results, we assume that there exists an interior solution to problem (4.2). However, in general, that need not be the case and hence, the terms increasing and decreasing in the following results should be interpreted as non–decreasing and non–increasing respectively.

Impact of θ

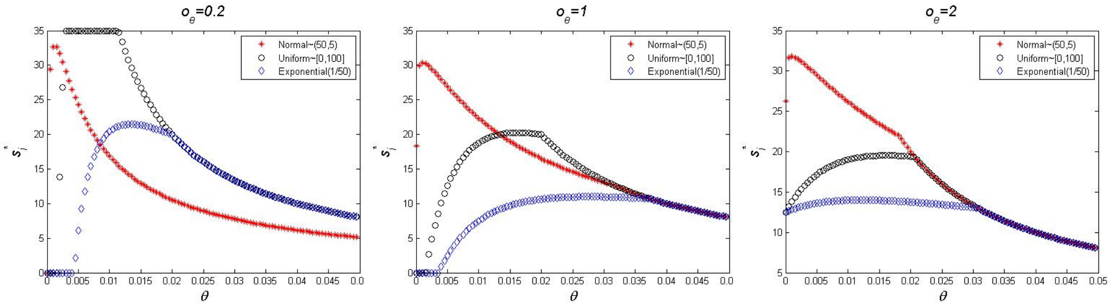
The parameter θ captures how supply chain investments translate into increased product availability. As θ increases, investing in the supply–side becomes increasingly more attractive and it may be beneficial to increase s_i to take advantage of the higher impact and reap the maximum benefits. However, it is also important to keep in mind that the principal is operating in a budget constrained environment and consequently, investment on the demand–side will decrease with s_i^*. Depending on D and o_e, the reduction in o_i could result in a situation where there may not be sufficient demand for the product, in which case increasing the product availability further would only be detrimental. Given the two effects, the act of balancing supply and demand is a delicate one as the following lemma demonstrates. The proofs for all the results in this chapter can be found in Appendix C.

Lemma 5. *s_i^* increases with θ if and only if $\theta < \frac{o_e(D_u - Bo_e + \alpha_0 Q)}{Q(D_u + Bo_e - \alpha_0 Q)}$ and decreases otherwise.*

Lemma 5 demonstrates that the relation between s_i^* and θ is threshold–type. When θ is below a threshold, we see that additional funds should be invested on the supply–side as θ increases. The idea is to take advantage of the higher supply–side investment impact to ramp up product availability. However, as s_i^* increases and we continue to invest additional funds on the supply–side, the supply–demand balance shifts, while continually lowering the marginal benefit of additional supply–side investments. At the threshold point, the scales tip in favor of increasing demand since continuing to further invest in increasing product availability would lead to a situation where there is too much supply but insufficient demand. Hence, in this case, investment on the supply–side should be scaled back and the funds should be diverted to stimulate demand. Note that a lower s_i^* does not necessarily mean reduced product availability — the increase in θ can compensate for the reduction in supply–side investment without impacting product availability.

We numerically checked to see if the threshold result holds true for normal and exponential distributions. To facilitate better comparison, we also plot the results for the uniform distribution. For the experiments, we assume the following parameters: α_0=0.6, Q=75, B=35. The demand distributions considered are normal (50,5), uniform [0,100] and exponential(1/50) distributions, i.e., all distributions have a mean value of 50. To ensure the robustness of the results, we also considered mean demand values of 30 and 70, and the results were consistent. All the distributions were truncated at a lower limit of 0 and upper limit 100.

Figure 4.1 displays how s_i^* varies with θ for different values of o_e. From the graphs, we see that the threshold result holds true for all three distributions, consistent with Lemma 5. The actual value of the threshold, however, varies with the demand distribution and o_e but this is to be expected given the specific nature of the threshold in Lemma 5.



(a) (b) (c)

Figure 4.1: Impact of θ on s_i^* for different values of o_e

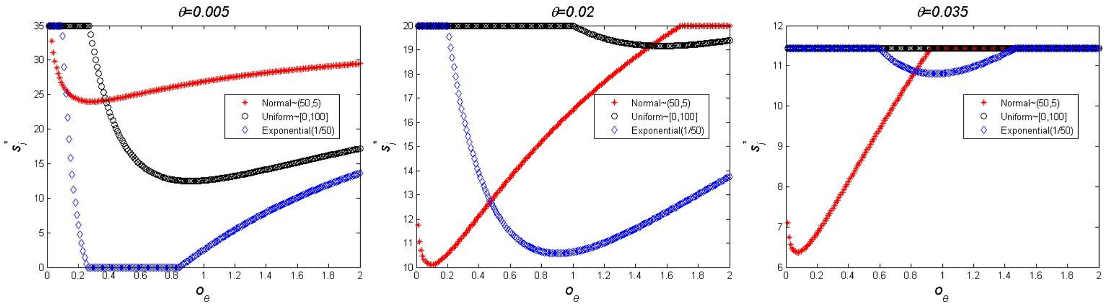
Impact of o_e

As o_e increases, the additional demand created for every dollar invested in the demand–side increases. The higher return on investment suggests that investing additional funds to stimulate demand would be a cost–effective method of increasing coverage, or in other words, s_i^* should decrease with o_e. However, the additional demand would translate into increased coverage only if a sufficient quantity of products are available to serve people who show up, which in turn depends on θ and Q. At a higher level, it appears that the relationship between s_i^* and o_e would depend on supply and demand considerations similar to the one that we discussed earlier when analyzing the impact of θ. Therefore, it is reasonable to expect a threshold–type result connecting s_i^* and o_e. Lemma 6 confirms that this is indeed the case.

Lemma 6. *s_i^* increases with o_e if and only if $o_e \geq \frac{2D_u\theta-(\alpha_0+B\theta)Q\theta}{(\alpha_0+B\theta)}$ and decreases otherwise.*

Comparing Lemmas 5 and 6, it is apparent that the effect of θ and o_e on s_i^* are the opposite from a qualitative perspective. However, the results need to be interpreted with caution since Lemmas 5 and 6 do not imply that the effect of a unit increase in θ and o_e on s_i^* are the opposite for any given set of parameter values. In fact, at a given θ and o_e, s_i^* could increase (or decrease) with both θ and o_e. Hence, the relationship between θ and s_i^* should not be used to make specific inferences regarding how s_i^* would vary with o_e.

We numerically checked to see if the threshold–type result proved in Lemma 6 holds under normal and exponential distributions. Figure 4.2 provides the results for different values of θ. From the figure, we see that for low and medium values of θ (0.005 and 0.02), the threshold



(a) (b) (c)

Figure 4.2: Impact of o_e on s_i^* for different values of θ

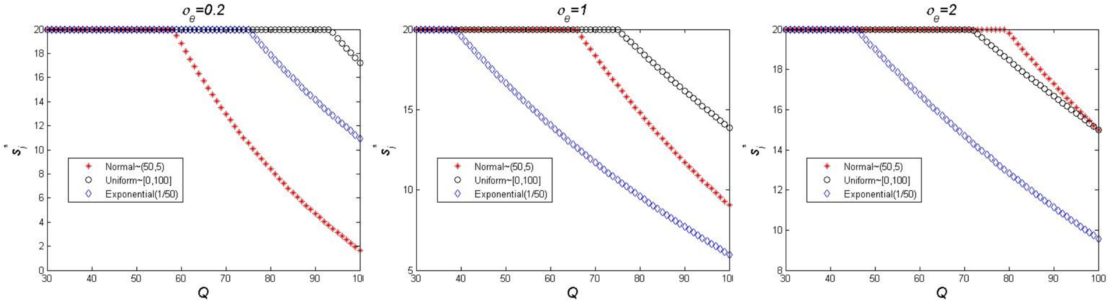
pattern holds for all three distributions. For a higher value of θ (0.035), the threshold pattern holds for the normal and exponential distribution, while in case of the uniform distribution, the threshold pattern weakly holds since s_i^* remains constant at a level where all the procured quantities are available to meet demand.

Impact of Q

Before we present the results regarding the relationship between s_i^* and Q, let us discuss how Q might qualitatively impact s_i^*. Note that Q is the maximum possible quantity available to meet demand. For a given base–demand scenario, i.e., without taking into account the effect of demand mobilization, there would be a "minimum" desired quantity below which supply would be considered too low even to meet base demand. Hence, for very low values of Q, it is beneficial to invest significantly on the supply–side to ensure that as much of the procured quantity is available to meet demand as possible. Once Q is above the "minimum" desired quantity, the cost–effectiveness aspect comes into play: is it better to increase coverage by increasing supply or through increasing demand? The answer to this question naturally depends on the relative effectiveness of the demand– and supply–side investments. As figure 4.3 demonstrates, there is a cutoff point until which it is more cost–effective to increase coverage by increasing supply (by maintaining the supply–side investment level), but beyond the cutoff point, we begin to reduce investments in improving the supply chain efficiency. The idea is easy to visualize for the normal distribution where from figure 4.3, we see that it is better to maintain 100% product availability until Q is less than or equal to the mean expected demand. When Q is greater than expected demand, the supply chain investments begin to reduce. Notice however that investment in the supply–side reduce only gradually (slope less than one) for all demand distributions, implying that, despite the reduction in supply–side investment, product availability continues to increase in conjunction with an increasing demand due to the additional demand–side investment.

The fact that supply chain investments either remain constant or decrease with Q, as seen in figure 4.3, is analytically proven in Lemma 7.

Lemma 7. *s_i^* always decreases with Q.*



(a) (b) (c)

Figure 4.3: Impact of Q on s_i^* for different values of o_e at θ=0.02

4.3.2 Impact of Demand Changes

In this section, we analyze how changes in the stochastic component of the demand impact the investment decisions and also the program coverage, which is the objective of interest. Specifically, we are interested in analyzing the impact of changes in mean demand as well as the variability of the demand. In what follows, we first try to understand how changes in mean demand and demand variance affect the investment decisions.

Impact of demand changes on investment decisions

Impact of mean demand: Let us begin with μ, which is the expected demand in the absence of any demand–side investment. The key question here is: should investments targeted at demand mobilization increase or decrease in anticipation of a higher expected demand? When the demand is expected to be higher, organizations frequently tend to scale back on community mobilization activities since they believe that there would be a substantial demand for the service or product even with little community outreach. Instead, they tend to invest more on the supply–side to increase product availability in anticipation of a higher demand. However, as Lemma 8 demonstrates, this strategy may not be always right and there could be situations where it is optimal to lower the supply–side investment as μ increases. In the following lemma, $D_{|\mu_1}$ and $D_{|\mu_2}$ represent the random variable D given means μ_1 and μ_2 respectively and F_1 and F_2 are the cumulative distribution function associated with $D_{|\mu_1}$ and $D_{|\mu_2}$ respectively.

Lemma 8. *For $\mu_2 \geq \mu_1$, if $D_{|\mu_2} \geq_{st} D_{|\mu_1}$, then $s_i^*(\mu_2) \geq s_i^*(\mu_1)$. Otherwise, $s_i^*(\mu_2)$ could be greater or less than $s_i^*(\mu_1)$ depending on F_1, F_2, α_0, θ, o_e, Q and B.*

In the above lemma, $\geq_{st}$ implies first–order stochastic dominance. Lemma 8 is particulary useful in practice since for common demand distributions like uniform, normal and exponential, $\mu_2 \geq \mu_1$ guarantees that $D_{|\mu_2} \geq_{st} D_{|\mu_1}$. Hence, when demand follows one of these distributions, simply knowing that the *expected demand will be higher* could direct the principal in the right direction in the process of finding the optimal investment mix. However, while first–order stochastic dominance implies $\mu_2 \geq \mu_1$, unfortunately, the converse is not true and hence, s_i^* may not always increase with expected demand. Hence, the specific demand scenario needs to be taken into account when responding to changes in the expected demand.

Impact of demand variability: Next, we look at how changes in demand variability impact the investment decisions. Specifically, we are interested in answering the following question: is it better to invest more or less in the supply–side as demand variability increases? While the question appears tricky, nevertheless, as we will show, the answer is clear and it is easy to use. We require the following definition for our analysis.

Definition 6. *(Song 1994) Consider two random variables X and Y having distributions F and G with densities f and g. Suppose that X and Y are either both continuous or both discrete. We say X is more variable than Y, denoted $X \geq_{var} Y$, if and only if $\mathsf{E}[X] = \mathsf{E}[Y]$ and $S(f-g)=$ 2 with sign sequence +,-,+. That is, f crosses g exactly twice, first from above and then from below.*

The above variability ordering is a natural way of comparing the spread of densities of two random variables. It is stronger than the convex order but it is weaker than "mean preserving spread". In fact, many commonly used distributions like uniform, normal, truncated normal, Weibull and gamma can be compared using the $\geq_{var}$ order. We are now ready to present our main result concerning the impact of demand variability. In the following lemma, $D_{|\sigma_1}$ and $D_{|\sigma_2}$ represent the random variable D given variances σ_1^2 and σ_2^2 respectively. The mean μ is the same for $D_{|\sigma_1}$ and $D_{|\sigma_2}$.

Lemma 9. *Suppose that $D_{|\sigma_2} \geq_{var} D_{|\sigma_1}$ for $\sigma_2 \geq \sigma_1$. Then, there exists a number $\delta, 0 \leq \delta \leq 1$, such that if the ratio $\frac{\theta Q}{\theta Q + o_e} > \delta$, $s_i^*(\sigma_2) \geq s_i^*(\sigma_1)$. Otherwise, $s_i^*(\sigma_2) < s_i^*(\sigma_1)$.*

Lemma 9 offers a clear and concise answer to the question of how investment decisions change with demand variability. We see that the optimal response to a change in variability is not unidirectional. Instead, the response depends critically on the effectiveness ratio $\frac{\theta Q}{\theta Q + o_e}$ that we identified in Section 4.3. From the lemma, we see that when the effectiveness ratio is high (above δ), it is optimal to increase the supply–side investments in response to the increased demand variability while the opposite is true when the effectiveness ratio is below the threshold. The fact that the ratio is easy to calculate and applying the lemma is straightforward makes the above result particularly appealing.

The following corollary, derived from the above lemma, characterizes the role of the different supply– and demand–side parameters namely θ, Q and o_e in determining how the investment decisions change with demand variability.

Corollary 3. *Let $D_{|\sigma_2} \geq_{var} D_{|\sigma_1}$ for $\sigma_2 \geq \sigma_1$. Then the following results hold.*

1. *Given o_e and Q, if $\theta > \frac{\delta o_e}{Q(1-\delta)}$, then $s_i^*(\sigma_2) \geq s_i^*(\sigma_1)$. Otherwise, $s_i^*(\sigma_2) < s_i^*(\sigma_1)$.*
2. *Given θ and Q, if $o_e \leq \frac{Q\theta(1-\delta)}{\delta}$, then $s_i^*(\sigma_2) \geq s_i^*(\sigma_1)$. Otherwise, $s_i^*(\sigma_2) < s_i^*(\sigma_1)$.*
3. *Given o_e and θ, if $Q > \frac{\delta o_e}{\theta(1-\delta)}$, then $s_i^*(\sigma_2) \geq s_i^*(\sigma_1)$. Otherwise, $s_i^*(\sigma_2) < s_i^*(\sigma_1)$.*

The above corollary provides a good understanding of how the different system parameters dictate the optimal response to a change in the demand variability. From the corollary, we see that for low values of θ ($< \frac{\delta o_e}{Q(1-\delta)}$), it is better to scale back the supply–side investment as demand variability increases. However, for higher values of θ, it is beneficial to instead scale back the demand–side investments and invest more on increasing product availability. We can make similar threshold–type conclusions with respect to the other two parameters from the corollary.

Impact of demand changes on program coverage

In this section, we explore how changes to the mean demand and variance impact program coverage. While it is important to understand how demand changes affect investment decisions, it is of considerable interest to humanitarian organizations to understand how the demand changes impact the ultimate outcome of these investment decisions, namely the program coverage.

Impact of mean demand: As the mean demand increases, intuition suggests that the expected program coverage would also increase since more people are expected to access the health service or product. This intuition is correct if the demand distribution belongs to the family of commonly used demand distributions like uniform, exponential and normal but as the following lemma demonstrates, higher expected demand need not necessarily translate into a higher program coverage.

Lemma 10. *For $\mu_2 \geq \mu_1$, if $D_{|\mu_2} \geq_{st} D_{|\mu_1}$, then the program coverage increases with mean demand w.p.1 and hence, also in expectation. Otherwise, the expected coverage could increase or decrease with μ depending on F_1, F_2, α_0, θ, o_e, Q and B.*

As we mentioned before, for uniform, exponential and normal distributions, a higher expected value implies stochastic dominance and hence, the program coverage is guaranteed to increase with the expected demand.

Impact of demand variability: The operations management community has long been interested in understanding the role of variability on operational performance. Several works have attempted to understand the impact of demand variability on the bottom line in a variety of settings, and the general consensus is that higher demand variability results in higher costs/lower profits or poor operational performance, in general. The next result shows that a similar conclusion is also valid in our setting — we find that an increase in demand variability leads to lower program coverage.

Lemma 11. *Suppose that $D_{|\sigma_2} \geq_{var} D_{|\sigma_1}$ for $\sigma_2 \geq \sigma_1$. Then the program coverage always decreases with σ.*

By clearly characterizing the negative impact of variability, Lemma 11 identifies a third way to increase coverage, in addition to the two that we have considered — supply chain strengthening and demand mobilization. While it is out of the scope of this work, nevertheless, an interesting question arises based on this finding: should we invest in creating *additional* demand or should we invest in *reducing* demand variability since either approach could be taken to increase the coverage, but the limited funding availability might preclude doing both. This could certainly be an avenue for future research.

So far, we have focused on the centralized case where the principal makes both supply– and demand–side investment decisions. Next, we consider a decentralized setting where demand mobilization activities are contracted to a third–party. As we mentioned earlier, contracting with third parties to carry out certain services or activities is a frequent practice in the public health sector, reasons for which include lack of efficiency of public health systems, lack of expertise, and human resource constraints. We discuss the decentralized setting in greater detail in the next section.

4.4 Decentralized model

In the decentralized model, the principal invests in the supply chain and contracts with an agent to carry out demand mobilization activities on her behalf. Since supply chain improvements involve advance planning and long implementation times, we assume that the supply–side investments are made well before the agent makes his decision regarding investment in demand–related activities. We also assume that the agent is aware of the total quantity that will be available to satisfy demand even though he may not aware of the principal's specific supply–side investment decision. This is a reasonable assumption since the agent is on the ground and more so, if the agent is directly involved in the program implementation. We denote the effective quantity available by $\tilde{Q}$ where $\tilde{Q} = min\{(\alpha_0 + \theta s_i), 1\}Q$ where s_i is the principal's investment decision. Notice that from the agent's perspective, $\tilde{Q}$ is given and fixed. The agent is interested in maximizing his profits, which is the difference of payments received from the principal less any investments made to stimulate demand. The payment mechanism used by the principal directly influences the investment decisions of the agent and in the decentralized setting, a natural question that arises is: can the principal design a contract to ensure that the program coverage in the decentralized setting is greater than or equal to the coverage in the centralized model?

Traditionally, payments to agents in the humanitarian sector have been fixed price or cost–reimbursement contracts without conditioning on the outcomes. This has resulted in poor aid–effectiveness because under the fixed price contract, agents try to exert as little effort as possible, and under the cost–reimbursement contract, agents do not have an incentive to engage in cost–efficient activities, resulting in wastage of aid dollars. Moreover, oftentimes, the principal who is providing the funding has no way to verify that the agent indeed carried out certain activities and this creates mistrust and accountability issues. To overcome this problem, many humanitarian organizations have started exploring the use of performance–based funding where payments to agents are based on outcomes and outputs rather than inputs. For example, the "Cash on Delivery" method developed and used by the Center for Global Development pays agents a per–unit reward for every unit of output/outcome (Center for Global Development 2011). Another example is a performance–based funding initiative undertaken by USAID at Haiti where the contracted NGOs were paid 94% of a mutually agreed amount upfront to deliver a predefined health services package to the local population (USAID 2010). The remaining 6% funding could be withheld if performance targets are not met and an additional 6% funding could be provided as a bonus if the NGOs exceeded the target levels. Notice that in this case, the payment is structured as a fixed–penalty, fixed–reward payment. The success of these performance–based funding models has caught the attention of the global health community and many organizations are beginning to make at least some portion of funding conditional on outcomes.

Motivated by the growing interest in results–based funding, we explore how performance–based funding can be used in a setting like ours to influence the agent's investment decisions. Since the principal is not on ground, she cannot directly observe the level of investment by the agent, or the realized demand. Hence, we use the program coverage as a measure of performance, i.e., payments made to the agent can only be based on the actual number of people served.

Before we provide details of the contracts, let us look at the agent's problem. To differentiate the demand–side investment decisions under the centralized and decentralized models, we denote the agent's investment decision by $\tilde{o}_i$. To allow for the possibility that the agent might be able to increase demand in a more (or less) cost effective way, we assume that for every dollar that the agent invests, demand increases by $\tilde{o}_e\tilde{o}_i$. Then the agent's problem can be formally stated as

$$Max \;\; \mathsf{E}\, TP\left(min\{D + \tilde{o}_e\tilde{o}_i, \tilde{Q}\}\right) - \tilde{o}_i, \quad \tilde{o}_i \geq 0 \tag{4.4}$$

where TP is the transfer payment to the agent that is contingent upon number of people served.

We consider two specific forms of performance–based contracts based on the examples we just provided. The first one is a per–unit reimbursement rate contract similar to the "Cash on Delivery" model that we described earlier. The second one is a penalty–reward contract where the agent incurs a fixed penalty for not meeting a pre–determined target coverage level and gets a fixed reward if the coverage is above the target level. This contract is similar to the contract offered in the USAID example. First, let us look at the per–unit reimbursement rate contract.

4.4.1 Per–unit reimbursement rate contract

Under this contract, the agent receives a payment of C_r for every unit of satisfied demand. Thus, the agent's problem is

$$Max \;\; \mathsf{E}\, C_r min\{D + \tilde{o}_e\tilde{o}_i, \tilde{Q}\} - \tilde{o}_i, \quad \tilde{o}_i \geq 0 \tag{4.5}$$

It is straightforward to show that the objective function in (4.5) is concave in $\tilde{o}_i$. Therefore (assuming that an interior solution exists), the agent's optimal investment decision $\tilde{o}_i^*$ satisfies the following condition.

$$F(\tilde{Q} - \tilde{o}_e\tilde{o}_i^*) = \frac{1}{c_r\tilde{o}_e} \tag{4.6}$$

When deciding the per–unit reimbursement rate, the principal's goal is to ensure that coverage level in the decentralized setting is at least as high as the centralized case. This would happen only if $\tilde{o}_e\tilde{o}_i^* \geq o_e o_i^*$. Notice that the principal is not interested in the agent's actual investment decision — she cares only about the net effect of the agent's investments.

To ensure that $\tilde{o}_e\tilde{o}_i^* \geq o_e o_i^*$ holds, we look at condition describing o_i^*. From equation (4.3), we see that o_i^* is the solution to

$$F(min\{(\alpha_0 + \theta s_i), 1\}Q - o_e o_i^*) = \frac{\theta Q}{\theta Q + o_e} \tag{4.7}$$

If the principal invests s_i^*, the optimal supply–side investment level in the centralized model, then from equations (4.6) and (4.7) (recall that $\tilde{Q} = min\{(\alpha_0 + \theta s_i), 1\}Q$), we see that the coverage under the centralized and decentralized models would be the same if

$$C_r = \frac{\theta Q + o_e}{\tilde{o}_e(\theta Q)} \tag{4.8}$$

We refer to the C_r identified in (4.8) as the coordinating per–unit reimbursement rate. Having identified the coordinating per–unit rate, our next step is to check if offering the per–unit reimbursement specified in (4.8) is actually feasible from the principal's standpoint. Checking the feasibility is critical since the principal has limited funding available ($B - s_i^* = o_i^*$ to be precise) to make payments to the agent. To begin with, notice that for equal expected coverage under the centralized and decentralized models, $\tilde{o}_i^* = \frac{o_e o_i^*}{\tilde{o}_e} > o_i^*$ for $\tilde{o}_e < o_e$. However, the transfer payment from the principal to the agent cannot exceed o_i^*. Thus, when contracting with a less cost–efficient agent ($\tilde{o}_e < o_e$), the coverage level under a decentralized setting would match the coverage in a centralized setting only if the agent's expected profits are negative, which is obviously not practical. In fact, the previous statement is true for any contract irrespective of the contract specifics. Hence, no coordinating contract exists in this case and from now on, we will assume that $\tilde{o}_e \geq o_e$.

Given the coordinating per–unit reimbursement rate, the expected payments from the principal to the agent equals

$$\frac{\theta Q + o_e}{\tilde{o}_e(\theta Q)} \mathsf{E}\ min\{D + o_e o_i^*, (\alpha_0 + \theta s_i^*)Q\}$$

which is less than or equal to o_i^* if and only if $\tilde{o}_e \geq o_e^T$ where $o_e^T = \frac{\frac{\theta Q + o_e}{\theta Q} \mathsf{E}\ min\{D + o_e o_i^*, (\alpha_0 + \theta s_i^*)Q\}}{o_i^*}$. Of course, one could also impose the restriction that the maximum transfer payment should not exceed o_i^* instead of the expected payment. This can be easily handled and will not change our results. Combining all the relevant results in this section so far, we have the following lemma.

Lemma 12. *When $\tilde{o}_e \geq o_e^T$, there exists a per–unit reimbursement rate contract with $C_r = \frac{\theta Q + o_e}{\tilde{o}_e(\theta Q)}$ such that the expected coverage under the centralized and decentralized models are the same. For $\tilde{o}_e < o_e^T$, no such feasible per–unit reimbursement rate contract exists.*

So far, our analysis focused on designing per–unit rate contracts that ensure that the coverage under the centralized and decentralized models remain the same. However, the principal can actually do better — she can make use of the fact that the agent is capable of influencing demand in a more cost-effective way and design a per–unit rate contract that achieves a higher coverage level than the centralized model. To see how, consider $\hat{o}_e \geq o_e$. Then the principal can solve the centralized problem given by equation (4.2) using $\hat{o}_e$ in place of o_e. Naturally, the expected coverage with $\hat{o}_e$ would be higher than the expected coverage with o_e. Let $\hat{s}_i^*$ and $\hat{o}_i^*$ be the optimal supply– and demand–side investment levels for this modified problem. Now if $\tilde{o}_e \geq \hat{o}_e^T$ where $\hat{o}_e^T = \frac{\frac{\theta Q + \hat{o}_e}{\theta Q} \mathsf{E}\ min\{D + \hat{o}_e \hat{o}_i^*, (\alpha_0 + \theta \hat{s}_i^*)Q\}}{\hat{o}_i^*} \geq \hat{o}_e$, then using Lemma 12, we see that it is indeed possible to achieve a coverage level (=$\mathsf{E}\ min\{D + \hat{o}_e \hat{o}_i^*, (\alpha_0 + \theta \hat{s}_i^*)Q\}$) that is higher than the coverage under the centralized model (=$\mathsf{E}\ min\{D + o_e o_i^*, (\alpha_0 + \theta s_i^*)Q\}$).

The only question remaining is: would $\tilde{o}_e \geq \hat{o}_e^T$ hold so that the principal can achieve higher coverage under the decentralized setting? The answer is yes, since for every $\tilde{o}_e > o_e^T$, it is possible to find a $\hat{o}_e \geq o_e$ such that $\tilde{o}_e \geq \hat{o}_e^T$. Hence, the principal can actually use the agent's cost effectiveness to her advantage and achieve a higher coverage level in the decentralized setting by offering a suitable contract.

Before we end the discussion regarding the per–unit reimbursement rate contract, we wish to point out two drawbacks of using such a contract: 1. As stated in Lemma 12, a feasible contract exists if and only if $\tilde{o}_e$ is greater than a threshold. Hence, when contracting with an agent with $\tilde{o}_e < o_e^T$, the principal needs to lower the desired investment level expected of the agent to design a feasible contract. 2. The reimbursement rate C_r depends explicitly on $\tilde{o}_e$. Hence, implementing such a contract would require that the principal have accurate information regarding how the agent's investments influence demand, but obtaining such precise information may not always be feasible in practice. The penalty–reward contract that we discuss next overcomes both the shortcomings.

4.4.2 Penalty–reward contract

Under the penalty–reward contract, the agent incurs a fixed penalty P if coverage is less than a predefined level T^c and gets a fixed reward R if coverage exceeds T^c. The principal sets the penalty P, reward R and the target T^c while the agent decides how much to invest in demand mobilization. The agent's problem under the penalty–reward contract is the following.

$$Max \;\; \mathsf{E} - P \; 1_{\{D + \tilde{o}e\tilde{o}_i < T^c\}} + R \; 1_{\{D + \tilde{o}e\tilde{o}_i \geq T^c\}} - \tilde{o}_i, \quad \tilde{o}_i \geq 0 \tag{4.9}$$

Given that the principal chooses three contract parameters, she has significant flexibility in designing a contract that guarantees that the coverage in the decentralized setting is at least as high as the centralized case. In fact, there could be many such contracts that satisfy the principal's requirements but for our purposes, we are only interested in identifying at least one such contract that can ensure a coverage level at least as high as the centralized setting.

To this end, suppose that the principal chooses the target level $T^c = o_e o_i^*$. Choice of this particular value of T^c is vital since for demand realizations lower than $o_e o_i^*$, the principal can be sure that the low coverage can be attributed solely to the lower demand–side investment by the agent. For any other choice of T^c, the principal cannot be certain that the agent is indeed responsible for the low coverage.

Now let us first consider $\tilde{o}_i$ such that $\tilde{o}_e\tilde{o}_i \geq o_e o_i^*$. Given that $T^c = o_e o_i^*$, the agent's objective function in this region of $\tilde{o}_i$ is

$$Max \ \ \mathsf{E} \ R \ 1_{\{D+\tilde{o}e\tilde{o}_i \geq o_e o_i^*\}} - \tilde{o}_i, \ \ \tilde{o}_i \geq \frac{o_e o_i^*}{\tilde{o}_e}$$

which is clearly decreasing in $\tilde{o}_i$ irrespective of the value of R. Hence the agent's investment would always be less than or equal to $\frac{o_e o_i^*}{\tilde{o}_e}$. For $\tilde{o}_i$ such that $\tilde{o}_e\tilde{o}_i < o_e o_i^*$, the agent's objective function is given by

$$Max \ \ \mathsf{E} - P \ F(o_e o_i^* - \tilde{o}_e\tilde{o}_i) + R \ 1_{\{D+\tilde{o}e\tilde{o}_i \geq o_e o_i^*\}} - \tilde{o}_i, \ \ \tilde{o}_i < \frac{o_e o_i^*}{\tilde{o}_e}$$

Clearly $R \ 1_{\{D+\tilde{o}e\tilde{o}_i \geq o_e o_i^*\}}$ is increasing in $\tilde{o}_i$. Depending on P, the other two terms could increase or decrease with $\tilde{o}_i$. As a next step, we choose a penalty P such that the agent would either be indifferent or strictly prefer investing to paying the penalty. One such penalty is $P = \frac{1}{\tilde{o}_e f_{LB}}$ where $f_{LB} = \min\{f(x) : 0 \leq x \leq o_e o_i^*\}$. With this penalty, the agent is always either indifferent or prefers to invest until $\tilde{o}_e\tilde{o}_i \geq o_e o_i^*$. Combined with our earlier observation that the agent's investment would always be less than or equal to $\frac{o_e o_i^*}{\tilde{o}_e}$, we see that the agent would invest $\tilde{o}_i^*$ such that $\tilde{o}_e\tilde{o}_i^*$ exactly matches $o_e o_i^*$. A final step is to set R in such a way that the agent's expected profits are non–negative so that he would find the contract acceptable. This can be achieved by setting $R = o_i^*$ since $\tilde{o}_i^* = \frac{o_e o_i^*}{\tilde{o}_e} \leq o_i^*$.

As with the per–unit reimbursement contract, if $\tilde{o}_e$ is known to the principal, she can utilize this knowledge to achieve a higher coverage level in the decentralized setting than in the centralized setting. Suppose that $\tilde{o}_e > \hat{o}_e \geq o_e$. Then the principal can solve the centralized problem given by equation (4.2) using $\hat{o}_e$ in place of o_e. Let $\hat{s}_i^*$ and $\hat{o}_i^*$ be the optimal supply– and demand–side investment levels for this modified problem. Now, if the principal sets $T^c = \hat{o}_e\hat{o}_i^*$, $R = \hat{o}_i^*$ and sets $f_{LB} = \min\{f(x) : 0 \leq x \leq \hat{o}_e\hat{o}_i^*\}$, then the agent's investment decision would be such that $\tilde{o}_i^* = \frac{\hat{o}_e \hat{o}_i^*}{\tilde{o}_e} \geq \frac{o_e o_i^*}{\tilde{o}_e}$, i.e., the principal can achieve a higher coverage under the decentralized setting.

4.4.3 Comparison of the contracts

The penalty–reward contract has some important advantages over the per–unit reimbursement contract. First, a feasible per–unit reimbursement contract with $\tilde{o}_e \tilde{o}_i^* = o_e o_i^*$ exists only if $\tilde{o}_e \geq o_e^T > o_e$ where o_e^T is as defined in section 4.4.1. However, with a penalty–reward contract, the same can be achieved as long as $\tilde{o}_e > o_e$. The fact that $\tilde{o}_e$ can be arbitrarily close to o_e also has the important implication that the penalty–reward contract can be used to achieve strictly higher coverage levels than can be obtained using the per–unit reimbursement contract.

Another advantage of the penalty–reward contract is that it can be used even when $\tilde{o}_e$ is unknown to the principal. The only necessary condition is $\tilde{o}_e > o_e$. In this case, the principal can set $P = \frac{1}{o_e f_{LB}}$ instead of $P = \frac{1}{\tilde{o}_e f_{LB}}$. The higher penalty ensures that the agent is again either indifferent or strictly prefers investing to paying the penalty until $\tilde{o}_e \tilde{o}_i \geq o_e o_i^*$.

4.5 Conclusions

In this chapter, we study the problem of determining the optimal level of investments in the supply– and demand–sides with the objective of maximizing the number of people served. Investments in the supply–side increase the fraction of the procured quantity available to serve demand, while community mobilization activities increase the demand for the service or product. Given a limited budget, public health managers are frequently confronted with the dilemma of how much funding to allocate to the two sides respectively. In this chapter, we answer this question and also provide several insights into how the optimal invest mix varies with respect to the procurement budget, the supply–side and demand side investment effectiveness, and the underlying demand distribution. Contrary to common practice, we show that there could be situations where it may be optimal to increase investment in demand–enhancing activities in anticipation of a higher expected demand. In addition, we see that higher expected demand leads to an increase in program coverage for many commonly used demand distributions but that need not always be the case. Our analysis also provides clear insights into the impact of demand variability on the investment mix. We see that if the effectiveness ratio, which is a function of the effectiveness parameters and the procurement quantity, is above a certain threshold, then supply–side investments will go up in response to increased demand variability, while the opposite is true if the critical ratio is below the threshold.

In addition to the results concerning the optimal investment mix, we also study a decentralized setting where the principal is not on ground and contracts with an agent to carry out demand mobilization. The agent is interested in maximizing his own profits while the principal's objective is to increase coverage resulting in an incentive misalignment. Motivated by the recent interest in performance–based contracts among the global health community, we consider two coverage–based contracts that can help achieve equal or better coverage than the centralized setting. The penalty–reward contract is the attractive of the two since it can guarantee higher coverage levels than the per–unit reimbursement contract and it can also be used in situations where the agent's investment effectiveness is unknown to the principal, a situation that could arise in practice.

Chapter 5

Conclusions and Future Research

This dissertation focuses on finding effective and efficient ways to manage humanitarian operations and understand how funding impacts performance in such settings. Uncertain and unpredictable donor funding is a major problem in the humanitarian sector and operations planning in this environment is challenging due to the uncertainty in demand as well as the funding required to satisfy demand. The first two chapters focus on inventory management and resource allocation problems in an uncertain funding environment and contribute to an understanding of how funding timing and uncertainty in funding impact operational performance and health outcomes. The third chapter studies the problem of identifying the optimal mix of investments in supply chain strengthening and demand mobilization activities with the objective of maximizing coverage. Viewed together, this dissertation offers approaches to make efficient operational and investment decisions in humanitarian settings and also provides insights into how different aspects of the operating environment like funding and demand characteristics impact aid effectiveness.

In the first chapter of the dissertation, we study the problem of managing inventory of a health commodity in the presence of funding constraints over a finite planning horizon. Funding from donors, which finances the procurement, is received in installments throughout the planning period with uncertainty around both the timing and amount received. We model this problem using a stylized multi–period inventory model with financial constraints. Despite the funding complexities, we show that the optimal replenishment policy is modified base–stock type which is easy to implement. We prove analytically that a higher variability in funding timing and (stochastically) late arrival of funds both drive up operating costs.

Our numerical study provides several insights that would be valuable to humanitarian health managers. Surprisingly, we find that receiving funding in equal installments is not the optimal funding pattern due to its inability to accommodate large demand surges upfront. Our analysis also shows that the benefits of receiving funding early are higher in under–financed systems while avoiding funding delays is critical in fully–financed systems. Our work offers an interesting insight that could be valuable in guiding fundraising efforts by humanitarian organizations. We find that even with less overall funding, performance may be better if the funding is received earlier or more steadily. Hence, humanitarian organizations should also pay attention to getting the funding in a timely fashion rather than focusing solely on raising as much funding as possible. Finally, our work also demonstrates that it is very important to take into account the operating environment (e.g., demand characteristics) when undertaking initiatives to improve the funding situation since the impact of different initiatives vary with the operating environment.

In the second chapter, we extend the work in the first chapter by allowing for the possibility that patients enrolled in a program could be in different health states and they could require treatment over different lengths of time. The treatment duration and the treatment response could be different between the health states. The total available funding is limited and funding inflow is unpredictable. In this setting, a key question is: how to dynamically allocate the limited funding to patients in different health states over a finite horizon so as to minimize the number of disease–adjusted life months lost.

We use a multiperiod stochastic dynamic programming framework with health–state dependent per–period and terminal penalty costs to analyze the allocation problem. We characterize the optimal policy to be a state–dependent allocation policy which makes computing the optimal policy difficult. We prove several monotonicity properties to help reduce the computational burden and we also develop two heuristics to handle large size problems in a reasonable amount of time. Our computational results indicate that the impact of transition rates between the health states on the health outcomes is not straightforward to characterize in such settings since the underlying funding situation plays a critical role in determining how changes in transition rates impact health outcomes. We also demonstrate that the impact of uncertainty in funding timing varies with the funding level and the length of the planning horizon. Finally, we find that in under–financed systems, the low funding availability severely hurts performance and it is beneficial to receive additional funding even if it comes at the cost of added funding uncertainty. In well–funded systems, that is not the case and receiving additional but more uncertain funding would be detrimental.

The third chapter addresses the problem of how much to invest in the supply–side as opposed to the demand–side, given that countries often have a limited budget for interventions aimed at improving health outcomes. Using a simple one–period model with stochastic demand, we identify the optimal investment mix that maximizes coverage. We provide several results regarding how the investment mix changes with respect to different system parameters, and prove some surprising results regarding the impact of expected demand and demand variability on the optimal investment decisions and program coverage. In particular, higher expected demand does not necessarily lead to increased program coverage and it may also not be optimal to lower demand–side investments in anticipation of a higher demand. We also provide a crisp characterization of how the investment decisions change with respect to demand variability and demonstrate that higher demand variability leads to low program coverage. In the second part of the chapter, we consider a decentralized setting where demand mobilization activities are contracted to an agent, who is a profit maximizer. We identify two performance–based contracts that the principal can use to create incentives for the agent to invest in demand–enhancing activities, and we show that both contracts ensure higher coverage levels than the centralized setting.

This dissertation has looked at three problems involving operations and funding that are motivated by humanitarian applications. Overall, I view this dissertation as a starting point to use operations management techniques to address some interesting and important research questions that would be of interest to humanitarian organizations. There are several avenues for future research both within the operations–finance interface and the broader topic of humanitarian operations. For example, in the first two chapters, we look at the problem of managing inventory and making resource allocation decision given an uncertain donor funding stream. However, with the growing popularity of performance–based funding, many donors make subsequent payments contingent on prior performance. In this scenario, the funding stream becomes endogenous to performance and it would be interesting to study how the operational strategies of the recipients would change in the performance–based funding paradigm. Another problem that would interesting to study is to explore the use of “supply chain segmentation” approaches to manage health supply chains. Many humanitarian organizations manage a broad variety of products, and these products exhibit varying characteristics along different dimensions like demand, storage requirements, shelf–life etc. However, currently, they either operate disease/program–specific supply chains or use a one–size–fits–all approach where all the products share the same resources and procedures are standardized across products. This leads to redundancies, poor service and excessive operating costs. An efficient approach to manage this problem is to segment the supply chain, whereby products/health facilities with similar characteristics are segmented and managed in groups. This approach is also currently being explored by USAID but a key challenge is to determine the optimal grouping of products, since the fixed and variable costs of operation depend on the product grouping.

Appendix A

Proofs for results in Chapter 2

Proof of Lemma 1

From the definition of J_0, it is clear that the joint convexity holds for $t = 0$. Let the induction hypothesis be that $J_{t-1}(x_{t-1}, r_{t-1}, O_{t-1})$ is jointly convex in x_{t-1} and r_{t-1} for any fixed value of O_{t-1}. For any (x_t^1, y_t^1, r_t^1), (x_t^2, y_t^2, r_t^2) and $0 \leq \lambda \leq 1$, we have

$$
\begin{aligned}
&J_{t-1}(\lambda y_t^1 + (1-\lambda)y_t^2 - \zeta_t, \lambda\left(r_t^1 - c(y_t^1 - x_t^1)\right) + (1-\lambda)\left(r_t^2 - c(y_t^2 - x_t^2)\right) + \sum_{j=O_{t-1}+1}^{O_t} z_j, O_{t-1}) \\
&\quad \leq \lambda J_{t-1}(y_t^1 - \zeta_t, r_t^1 - cy_t^1 + cx_t^1 + \sum_{j=O_{t-1}+1}^{O_t} z_j, O_{t-1}) \\
&\qquad + (1-\lambda)J_{t-1}(y_t^2 - \zeta_t, r_t^2 - cy_t^2 + cx_t^2 + \sum_{j=O_{t-1}+1}^{O_t} z_j, O_{t-1})
\end{aligned}
$$

The above inequality follows directly from the induction assumption. Since convexity is preserved under expectation $\mathsf{E}_{O_{t-1}}\mathsf{E}_{\zeta_t} J_{t-1}\left(y_t - \zeta_t, r_t - c(y_t - x_t) + \sum_{j=O_{t-1}+1}^{O_t} z_j, O_{t-1}\right)$ is jointly convex in x_t, y_t and r_t. The remaining terms cy_t, $b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+$ and $h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+$ are convex in y_t and hence joint convexity in x_t, y_t and r_t holds. The set $\mathbf{C}=\left\{(x, y, r) : r \geq 0, y \in \left[x, x + \frac{r}{c}\right]\right\}$ is convex. Using proposition B-4 from Heyman and Sobel (1984), we see that $J_t(x_t, r_t, O_t)$ is jointly convex in x_t and r_t for a fixed O_t.

Proof of Theorem 1

The proof proceeds through induction on the number of periods. Recall that O_1=0 always. Since $\tilde{J}_0(x_0, R_0, 0) = 0\ \forall(x_0, R_0)$, we have

$$
\tilde{J}_1(x_1, R_1, 0) = -cx_1 + \min_{y_1 \in \left[x_1, \frac{R_1}{c}\right]} \left\{ \; cy_1 + b\mathsf{E}_{\zeta_1}[\zeta_1 - y_1]^+ + h\mathsf{E}_{\zeta_1}[y_1 - \zeta_1]^+ \; \right\} \tag{A.1}
$$

From $NV_0(x_0) = 0 \; \forall \; x_0$, we have

$$NV_1(x_1) = -cx_1 + \min_{y_1 \geq x_1} \left\{ \; cy_1 + b\mathsf{E}_{\zeta_1}[\zeta_1 - y_1]^+ + h\mathsf{E}_{\zeta_1}[y_1 - \zeta_1]^+ \; \right\} \tag{A.2}$$

Clearly, the function to be minimized in equations (A.1) and (A.2) is the same. Moreover, it is well–known that the function is convex in y_1 and there exists an optimal base stock level y_1^* at which the function is minimized. Therefore, it follows directly that the replenishment policy specified in Theorem 1 is optimal for t=1. Furthermore, $\frac{\partial \tilde{J}_1(x_1,R_1,0)}{\partial x_1} = \frac{\partial \tilde{NV}_1(x_1)}{\partial x_1}$ for any given R_1.

Now let the induction assumption be that replenishment policy specified in Theorem 1 is optimal for t-1. Also, let $\frac{\partial \tilde{J}_{t-1}(x_{t-1},R_{t-1},O_{t-1})}{\partial x_{t-1}} = \frac{\partial \tilde{NV}_{t-1}(x_{t-1})}{\partial x_{t-1}}$ for any given (R_{t-1}, O_{t-1}). We have

$$\tilde{J}_t(x_t, R_t, O_t) = -cx_t + \min_{y_t \in \left[x_t, \frac{R_t}{c}\right]} \left\{ \begin{array}{l} cy_t + b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+ + h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+ \\ \quad +\mathsf{E}_{O_{t-1},\zeta_t} \tilde{J}_{t-1}(y_t - \zeta_t, R_t - c\zeta_t + \displaystyle\sum_{j=O_{t-1}+1}^{O_t} z_j, O_{t-1}) \end{array} \right\} \tag{A.3}$$

and

$$NV_t(x_t) = -cx_t + \min_{y_t \geq x_t} \left\{ \; cy_t + b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+ + h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+ + \mathsf{E}_{\zeta_t} NV_{t-1}(y_t - \zeta_t) \; \right\} \tag{A.4}$$

The function to be minimized in equation (A.3) is $\hat{C}_t(y_t, R_t, O_t)$ (see equation (2.2)). From Lemma 1, we know that for fixed (R_t, O_t), $\hat{C}_t(y_t, R_t, O_t)$ is convex in y_t. It is also well known that the function to be minimized in equation (A.4) is convex in y_t and there exists an optimal base stock level y_t^* at which the function is minimized. Then, the induction assumption $\frac{\partial \tilde{J}_{t-1}(x_{t-1},R_{t-1},O_{t-1})}{\partial x_{t-1}} = \frac{\partial NV_{t-1}(x_{t-1})}{\partial x_{t-1}}$ for every (R_{t-1}, O_{t-1}) implies that the derivatives (with respect to y_t) of the functions to be minimized in equations (A.3) and (A.4) are the same. Therefore, the replenishment policy specified in Theorem 1 is optimal for period t as well.

To prove the other part of the induction, i.e., $\frac{\partial \tilde{J}_t(x_t,R_t,O_t)}{\partial x_t} = \frac{\partial NV_t(x_t)}{\partial x_t}$ for any given (R_t, O_t), consider the following cases.

Case 1: $\frac{R_t}{c} < y_t^*$. Then,

$$\begin{aligned}\tilde{J}_t(x_t, R_t, O_t) = -cx_t + c\frac{R_t}{c} + b\mathsf{E}_{\zeta_t}\left[\zeta_t - \frac{R_t}{c}\right]^+ + h\mathsf{E}_{\zeta_t}\left[\frac{R_t}{c} - \zeta_t\right]^+ \\ + \mathsf{E}_{O_{t-1}}\mathsf{E}_{\zeta_t}\tilde{J}_{t-1}\left(\frac{R_t}{c} - \zeta_t, R_t - c\zeta_t + \sum_{j=O_{t-1}+1}^{O_t} z_j, O_{t-1}\right)\end{aligned}$$

and

$$NV_t(x_t) = -cx_t + cy_t^* + b\mathsf{E}_{\zeta_t}\left[\zeta_t - y_t^*\right]^+ + h\mathsf{E}_{\zeta_t}\left[y_t^* - \zeta_t\right]^+ + \mathsf{E}_{\zeta_t}NV_{t-1}\left(y_t^* - \zeta_t\right)$$

Clearly, $\frac{\partial \tilde{J}_t}{\partial x_t} = \frac{\partial NV_t}{\partial x_t} = -c$.

Case 2: $\frac{R_t}{c} \geq y_t^*$, $x_t < y_t^*$. Then,

$$\begin{aligned}\tilde{J}_t(x_t, R_t, O_t) = -cx_t + cy_t^* + b\mathsf{E}_{\zeta_t}\left[\zeta_t - y_t^*\right]^+ + h\mathsf{E}_{\zeta_t}\left[y_t^* - \zeta_t\right]^+ \\ + \mathsf{E}_{O_{t-1}}\mathsf{E}_{\zeta_t}\tilde{J}_{t-1}\left(y_t^* - \zeta_t, R_t - c\zeta_t + \sum_{j=O_{t-1}+1}^{O_t} z_j, O_{t-1}\right)\end{aligned}$$

and

$$NV_t(x_t) = -cx_t + cy_t^* + b\mathsf{E}_{\zeta_t}\left[\zeta_t - y_t^*\right]^+ + h\mathsf{E}_{\zeta_t}\left[y_t^* - \zeta_t\right]^+ + \mathsf{E}_{\zeta_t}NV_{t-1}\left(y_t^* - \zeta_t\right)$$

Again, $\frac{\partial \tilde{J}_t}{\partial x_t} = \frac{\partial NV_t}{\partial x_t} = -c$.

Case 3: $x_t \geq y_t^*$. Then,

$$\begin{aligned}\tilde{J}_t(x_t, R_t, O_t) = b\mathsf{E}_{\zeta_t}\left[\zeta_t - x_t\right]^+ + h\mathsf{E}_{\zeta_t}\left[x_t - \zeta_t\right]^+ \\ + \mathsf{E}_{O_{t-1}}\mathsf{E}_{\zeta_t}\tilde{J}_{t-1}\left(x_t - \zeta_t, R_t - c\zeta_t + \sum_{j=O_{t-1}+1}^{O_t} z_j, O_{t-1}\right)\end{aligned}$$

and

$$NV_t(x_t) = b\mathsf{E}_{\zeta_t}\left[\zeta_t - x_t\right]^+ + h\mathsf{E}_{\zeta_t}\left[x_t - \zeta_t\right]^+ + \mathsf{E}_{\zeta_t}NV_{t-1}\left(x_t - \zeta_t\right)$$

Since, $\frac{\partial \tilde{J}_{t-1}}{\partial x_{t-1}}$=$\frac{\partial NV_{t-1}}{\partial x_{t-1}}$ for every (R_{t-1}, O_{t-1}), we have $\frac{\partial \tilde{J}_t}{\partial x_t}$=$\frac{\partial NV_t}{\partial x_t}$ in this case as well, completing the proof.

Proof of Theorem 2

We use a sample path approach to prove this result. For expositional clarity, let us define a new variable $\bar{Z}^n = (\bar{z}_1^n, \bar{z}_2^n, ..., \bar{z}_t^n)$, n=1,2, where $\bar{z}_i^n$ is the amount received in period i under funding scenario n, given that O_t=j. Of course, $\bar{Z}^1$ and $\bar{Z}^2$ are different for different sample paths. Let $\zeta = (\zeta_1, \zeta_2, ..., \zeta_t)$ be the vector of realized demands in periods $1, 2, ..., t$ along a particular sample path. Given ζ and $\bar{Z}^n$, let $J^n_{t,\zeta,\bar{Z}^n}(x_t, r_t, j)$ be the cost incurred in periods $1, 2, ..., t$ along a particular sample path under funding scenario n, following the optimal replenishment policy specified in (2.4). If random variables ζ, $\bar{Z}^1$, $\bar{Z}^2$, $\{P^1_t\}$ and $\{P^2_t\}$ are defined on the same probability space, then $Pr(P^2_t(i) \geq P^1_t(i))$=1. Also, $Pr(P^n_t(i') \geq P^n_t(i))$=1 for $i' > i$ and $n = 1, 2$. This implies that vector $\bar{Z}^1$ majorizes $\bar{Z}^2$ w.p. 1. Since replenishment decisions depend only on the current state (x_t, r_t), every replenishment decision feasible under scenario 2 is also feasible under scenario 1 along every sample path. Therefore, $J^2_{t,\zeta,\bar{Z}^n}(x_t, r_t, j) \geq J^1_{t,\zeta,\bar{Z}^n}(x_t, r_t, j)$. w.p.1. Since, this result holds for every sample path, the result also holds in expectation, i.e, $J^2_t(x_t, r_t, j) \geq J^1_t(x_t, r_t, j)$.

Proof of Lemma 2

The proof proceeds through induction. Throughout the proof, we work with the equivalent function $\tilde{J}_t$ instead of J_t, as we find it convenient to do so. Recall that $R_t = r_t + cx_t$. For $t = 2$,

$$\tilde{J}_2(x_2, R_2 + \sum_{k=j+1}^{i} z_k, j) = \min_{y_2 \in \left[x_2, \frac{R_2 + \sum_{k=j+1}^{i} z_k}{c}\right]} \left\{ \begin{array}{l} c(y_2 - x_2) \\ \\ +b\mathsf{E}_{\zeta_2}[\zeta_2 - y_2]^+ + h\mathsf{E}_{\zeta_2}[y_2 - \zeta_2]^+ \\ \\ +\mathsf{E}_{\zeta_2}\tilde{J}_1(y_2 - \zeta_2, R_2 - c\zeta_2 + \sum_{k=1}^{i} z_k, 0) \end{array} \right\} \tag{A.5}$$

Notice that in equation (A.5), the function over which the minimization is done is $-cx_2 +$ $\hat{C}_2(y_2, R_2, i)$ (see equation (2.2)). For a fixed i, we see from Lemma 1 that $\hat{C}_2$ is jointly convex in y_2 and R_2. This implies the convexity of $\hat{C}_2$ in y_2 for a fixed R_2. The desired result, $\tilde{J}_2(x_2, R_2 + \sum_{k=j+1}^{i} z_k, j)$ is increasing convex in j, then follows directly from the fact that $z_{l+1} \geq z_l, l = 1, 2, ..., m-1$.

Now assume that the result holds for $t-1$. Then, for $j \leq i$, we have

$$\tilde{J}_t(x_t, R_t + \sum_{k=j+1}^{i} z_k, j)$$

$$= \min_{y_t \in \left[x_t, \frac{R_t + \sum_{k=j+1}^{i} z_k}{c}\right]} \left\{ \begin{array}{l} c(y_t - x_t) + b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+ + h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+ \\ \\ + \sum_{j'=0}^{i} p_t(j, j')\mathsf{E}_{\zeta_t}\tilde{J}_{t-1}(y_t - \zeta_t, R_t - c\zeta_t + \sum_{k=j'+1}^{i} z_k, j') \end{array} \right\} \tag{A.6}$$

where $p_t(j, j') = 0$ for $j' > j$. For brevity, we denote the function over which the minimization is done in equation (A.6) by $f(j)$. We know that for fixed R_t, $f(j)$ is convex in y_t for each j. In the proof of Theorem 1, we showed that the derivative of $\tilde{J}_{t-1}(y_t - \zeta_t, R_t - c\zeta_t + \sum_{k=j'+1}^{i} z_k, j')$ with respect to y_t is the same irrespective of the value of j'. Hence, functions $f(j), j = 0, 1, 2, ..., i$ are parallel, convex functions in y_t. Now, we make use of the SICX property. Since $\{P_t(j), j \in \{0, 1, 2, ...i\}\} \in$ SICX and $\tilde{J}_{t-1}(x_{t-1}, R_{t-1} + \sum_{k=j+1}^{i} z_k, j)$ is increasing convex in j by our induction assumption, from Definition 3, we see that $\sum_{j'=0}^{i} p_t(j, j')\mathsf{E}_{\zeta_t}\tilde{J}_{t-1}(y_t - \zeta_t, R_t - c\zeta_t + \sum_{k=j'+1}^{i} z_k, j')$ is increasing and convex in j. This implies that $f(j+1) - f(j) \geq f(j) - f(j-1)\ \forall j = 1, 2, ..., i-1$. Combining this with the fact that the funding vector is front-loaded yields the desired result.

Proof of Theorem 3

As with most other proofs, we prove the theorem through induction. We start the induction with t=2. Since $P_2^1(i) \equiv P_2^2(i) \equiv$0, using equation (2.1), we directly get $J_2^2(x_2, r_2, i) =$

$J_2^1(x_2, r_2, i)$. Now, assume that the result holds for $t-1$.

$$\begin{aligned} J_t^1(x_t, r_t, i) &= \min_{y_t \in [x_t, x_t + \frac{r_t}{c}]} \left\{ \begin{array}{l} c(y_t - x_t) + b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+ + h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+ \\ \quad + \sum_{j=0}^{i} p_t^1(i,j) \mathsf{E}_{\zeta_t} J_{t-1}^1(y_t - \zeta_t, r_t - c(y_t - x_t) + \sum_{k=j+1}^{i} z_k, j) \end{array} \right\} \\ &\le \min_{y_t \in [x_t, x_t + \frac{r_t}{c}]} \left\{ \begin{array}{l} c(y_t - x_t) + b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+ + h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+ \\ \quad + \sum_{j=0}^{i} p_t^1(i,j) \mathsf{E}_{\zeta_t} J_{t-1}^2(y_t - \zeta_t, r_t - c(y_t - x_t) + \sum_{k=j+1}^{i} z_k, j) \end{array} \right\} \end{aligned} \tag{A.7}$$

The inequality in the second step follows from the induction assumption. Also,

$$J_t^2(x_t, r_t, i) = \min_{y_t \in [x_t, x_t + \frac{r_t}{c}]} \left\{ \begin{array}{l} c(y_t - x_t) + b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+ + h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+ \\ \quad + \sum_{j=0}^{i} p_t^2(i,j) \mathsf{E}_{\zeta_t} J_{t-1}^2(y_t - \zeta_t, r_t - c(y_t - x_t) + \sum_{k=j+1}^{i} z_k, j) \end{array} \right\} \tag{A.8}$$

Our proof would be complete if we can demonstrate that (A.7)≤(A.8). The convex ordering of $P_t^1(i)$ and $P_t^2(i)$ implies that the desired inequality would hold if $J_{t-1}^2(y_t - \zeta_t, r_t - c(y_t - x_t) + \sum_{k=j+1}^{i} z_k, j)$ is convex in j. We already proved the convexity in Lemma 2. Hence $J_t^2(x_t, r_t, i) \ge J_t^1(x_t, r_t, i)$ holds.

Proof of Proposition 1

We use a sample path approach. We use superscript i when we are dealing with funding vector Z^i. On a fixed sample path, let the vector of realized demands in periods $N, N-1, ..., 1$ be ζ=$(\zeta_N, \zeta_{N-1}, ..., \zeta_1)$. Then,

$$R_{N-t}^i = R_N^i - c\sum_{j=0}^{t-1} \zeta_{N-j} + \sum_{j=1}^{t} z_{N-j}^i + cx_N, \quad t = N-1, N-2, ..., 1.$$

Since $\sum_{j=N}^{N-i} z_j^1 \ge \sum_{j=N}^{N-i} z_j^2,\ i = 0, 1, , ..., N-1,\ R_{N-t}^1 \ge R_{N-t}^2$ for t=1,2,...N. From Theorem 1, we know that the replenishment decision in period t depends only on (x_t, r_t) and not on future funding. Since $R^1_{N-t} \geq R^2_{N-t}$, and the starting inventory x_N is the same, every replenishment decision feasible under Z^2 is also feasible under Z^1. This implies that $\tilde{V}^2_{N,\zeta}(x_N, R^2_N) \geq \tilde{V}^1_{N,\zeta}(x_N, R^1_N)$ where $\tilde{V}^i_{N,\zeta}(x_N, R^1_N)$ is the total realized cost in periods $1, 2, ..., N$ under funding vector Z^2. Taking expectation yields $\tilde{V}^2_N(x_N, R^2_N) \geq \tilde{V}^1_N(x_N, R^1_N)$.

Proof of Proposition 2

To begin with, notice that the amount received until period $t+1$ is the same under all the three funding vectors. Since inventory decisions are made based on only the current state of the system (according to (2.4)), the cost incurred under all three funding vectors is the same until period $t+1$. Let (x_t, r_t) denote the state of the system at the beginning of period t under vector Z. Then, the system state under vectors Z^A and Z^D are $(x_t, r_t+\delta)$ and $(x_t, r_t-\delta)$ respectively. Similar to equation (2.3), we have

$$V_t(x_t, r_t + \delta) = -cx_t + \min_{y_t \in \left[x_t, \frac{R_t+\delta}{c}\right]} \left\{ \ C_t(y_t, R_t) \ \right\}$$

$$V_t(x_t, r_t) = -cx_t + \min_{y_t \in \left[x_t, \frac{R_t}{c}\right]} \left\{ \ C_t(y_t, R_t) \ \right\}$$

and

$$V_t(x_t, r_t - \delta) = -cx_t + \min_{y_t \in \left[x_t, \frac{R_t-\delta}{c}\right]} \left\{ \ C_t(y_t, R_t) \ \right\}$$

where $C_t(y_t, R_t) = cy_t + b\mathsf{E}_{\zeta_t}[\zeta_t - y_t]^+ + h\mathsf{E}_{\zeta_t}[y_t - \zeta_t]^+ + \mathsf{E}_{\zeta_t}V_{t-1}(y_t - \zeta_t, R_t - cy_t + z_{t-1})$. Notice that C_t is a special case of $\hat{C}_t$ (see equation (2.2)). From Lemma 1, it follows that for fixed R_t and O_t, $\hat{C}_t(y_t, R_t, O_t)$ is convex in y_t. Hence, $C_t(y_t, R_t)$ is convex in y_t for fixed R_t. The desired result, $V^D_N(x_N, r_N) - V_N(x_N, r_N) \geq V_N(x_N, r_N) - V^A_N(x_N, r_N)$, follows directly from the convexity.

Proof of Theorem 4

The proof proceeds through induction on the number of periods to go. We know that $NV^\lambda_0(.) = G^\lambda_0(.) = 0$. For $t \leq \lambda$, orders placed in period t will not be received before the end of the horizon and hence, no orders will be placed during these periods. Therefore, it follows that for every (R_t, OF_t), $NV_t^\lambda(x_t, w_t^1, w_t^2, ..., w_t^{\lambda-1})$=$G_t^\lambda(x_t, w_t^1, w_t^2, ..., w_t^{\lambda-1}, R_t, OF_t)$ $\forall\ t \leq \lambda$. Moreover,

$$\begin{aligned} NV_\lambda^\lambda(x_\lambda, w_\lambda^1, w_\lambda^2, ..., w_\lambda^{\lambda-1}) &= \tilde{f}_\lambda(x_\lambda) + \mathsf{E}_{\zeta_\lambda}\tilde{f}_{\lambda-1}(x_\lambda + w_\lambda^1 - \zeta_\lambda) \\ &+ ... + \mathsf{E}_{\zeta_\lambda,\zeta_{\lambda-1},...,\zeta_2}\tilde{f}_1(IP_\lambda - \zeta_\lambda - \zeta_{\lambda-1} - ... - \zeta_2) \end{aligned} \tag{A.9}$$

where $\tilde{f}_i(x) = h\mathsf{E}_{\zeta_i}[x - \zeta_i]^+ + b\mathsf{E}_{\zeta_i}[\zeta_i - x]^+$ and $IP_\lambda = x_\lambda + w_\lambda^1 + ... + w_\lambda^{\lambda-1}$. From equation (A.9), it follows directly that NV_λ^λ is jointly convex in $x_\lambda, w_\lambda^1, w_\lambda^2, ..., w_\lambda^{\lambda-1}$. Now,

$$\begin{aligned} & NV_{\lambda+1}^\lambda(x_{\lambda+1}, w_{\lambda+1}^1, ..., w_{\lambda+1}^{\lambda-1}) \\ &= \min_{z_{\lambda+1}\geq 0} \left\{ \begin{array}{l} cz_{\lambda+1} + b\mathsf{E}_{\zeta_{\lambda+1}}[\zeta_{\lambda+1} - x_{\lambda+1}]^+ + h\mathsf{E}_{\zeta_{\lambda+1}}[x_{\lambda+1} - \zeta_{\lambda+1}]^+ \\ +\mathsf{E}_{\zeta_{\lambda+1}}NV_\lambda^\lambda(x_{\lambda+1} - \zeta_{\lambda+1} + w_{\lambda+1}^1, w_{\lambda+1}^2, ..., w_{\lambda+1}^{\lambda-1}, z_{\lambda+1}) \end{array} \right\} \end{aligned} \tag{A.10}$$

From the convexity of NV_λ^λ, we have that the expression to be minimized in equation (A.10) is jointly convex in $x_{\lambda+1}, w_{\lambda+1}^1, ..., w_{\lambda+1}^{\lambda-1}$ and $z_{\lambda+1}$. Two important results follow: (i) $NV_{\lambda+1}^\lambda$ is jointly convex in $x_{\lambda+1}, w_{\lambda+1}^1, ..., w_{\lambda+1}^{\lambda-1}$. (ii) The specific form of equation (A.9) implies that there exists a base stock level $y_{\lambda+1}^{\lambda*}$, independent of $x_{\lambda+1}, w_{\lambda+1}^1, ..., w_{\lambda+1}^{\lambda-1}$, such that $z_{\lambda+1}^*$=max $(y_{\lambda+1}^{\lambda*} - IP_{\lambda+1}$,0) minimizes (A.10). In general, we can show through induction that NV_t^λ is jointly convex in $x_t, w_t^1, ..., w_t^{\lambda-1}$ and that there exists a base stock level y_t^*, independent of $x_t, w_t^1, ..., w_t^{\lambda-1}$, such that z_t^*=max $(y_t^{\lambda*} - IP_t$,0) is the optimal order quantity in period t.

$$\begin{aligned} & G_{\lambda+1}^\lambda(x_{\lambda+1}, w_{\lambda+1}^1, ..., w_{\lambda+1}^{\lambda-1}, R_{\lambda+1}, OF_{\lambda+1}) \\ &= \min_{0\leq z_{\lambda+1}\leq \frac{R_{\lambda+1}}{c} - IP_{\lambda+1}} \left\{ \begin{array}{l} cz_{\lambda+1} + b\mathsf{E}_{\zeta_{\lambda+1}}[\zeta_{\lambda+1} - x_{\lambda+1}]^+ + h\mathsf{E}_{\zeta_{\lambda+1}}[x_{\lambda+1} - \zeta_{\lambda+1}]^+ \\ +\mathsf{E}G_\lambda^\lambda(x_{\lambda+1} - \zeta_{\lambda+1} + w_{\lambda+1}^1, \\ \qquad ..., w_{\lambda+1}^{\lambda-1}, z_{\lambda+1}, R_{\lambda+1} - c\zeta_{\lambda+1} + OF_{\lambda+1} - OF_\lambda, OF_\lambda) \end{array} \right\} \\ &= \min_{0\leq z_{\lambda+1}\leq \frac{R_{\lambda+1}}{c} - IP_{\lambda+1}} \left\{ \begin{array}{l} cz_{\lambda+1} + b\mathsf{E}_{\zeta_{\lambda+1}}[\zeta_{\lambda+1} - x_{\lambda+1}]^+ + h\mathsf{E}_{\zeta_{\lambda+1}}[x_{\lambda+1} - \zeta_{\lambda+1}]^+ \\ +\mathsf{E}NV_\lambda^\lambda(x_{\lambda+1} - \zeta_{\lambda+1} + w_{\lambda+1}^1, w_{\lambda+1}^2, ..., w_{\lambda+1}^{\lambda-1}, z_{\lambda+1}) \end{array} \right\} \end{aligned} \tag{A.11}$$

where the second equality follows from the fact that $NV_\lambda^\lambda(.) = G_\lambda^\lambda(., R_\lambda, OF_\lambda)$ $\forall$ (R_λ, OF_λ).

The expressions to be minimized in (A.10) and (A.11) are the same, implying that $z^*_{\lambda+1}$= max(min($y^{\lambda*}_{\lambda+1} - IP_{\lambda+1}, \frac{R_{\lambda+1}}{c} - IP_{\lambda+1}$),0) minimizes equation (A.11). Furthermore, for fixed $R_{\lambda+1}$, C=$\{(x_{\lambda+1}, w^1_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, z_{\lambda+1}) : 0 \leq z_{\lambda+1} \leq \frac{R_{\lambda+1}}{c} - IP_{\lambda+1}\}$ is a convex set. Using proposition B-4 from Heyman and Sobel (1984), we see that $G^{\lambda}_{\lambda+1}(x_{\lambda+1}, w^1_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, R_{\lambda+1}, OF_{\lambda+1})$ is jointly convex in $x_{\lambda+1}, w^1_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}$ for fixed $(R_{\lambda+1}, OF_{\lambda+1})$.

Now consider the following cases.

Case 1.1: $IP_{\lambda+1} \geq y^{\lambda*}_{\lambda+1}$

$$\begin{aligned} NV^{\lambda}_{\lambda+1}(x_{\lambda+1}, w^1_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}) &= G^{\lambda}_{\lambda+1}(x_{\lambda+1}, w^1_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, R_{\lambda+1}, OF_{\lambda+1}) \\ &= h\mathsf{E}_{\zeta_{\lambda+1}}[x_{\lambda+1} - \zeta_{\lambda+1}]^+ + b\mathsf{E}_{\zeta_{\lambda+1}}[\zeta_{\lambda+1} - x_{\lambda+1}]^+ \\ &\quad + \mathsf{E}_{\zeta_{\lambda+1}} NV^{\lambda}_{\lambda}(x_{\lambda+1} + w^1_{\lambda+1} - \zeta_{\lambda+1}, w^2_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, 0) \end{aligned}$$

where we have again used the fact that $NV^{\lambda}_{\lambda}(.) = G^{\lambda}_{\lambda}(., R_{\lambda}, OF_{\lambda})$.

Case 1.2: $IP_{\lambda+1} < y^{\lambda*}_{\lambda+1}$, $\frac{R_{\lambda+1}}{c} \geq y^{\lambda*}_{\lambda+1}$

$$\begin{aligned} NV^{\lambda}_{\lambda+1}(x_{\lambda+1}, w^1_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}) &= G^{\lambda}_{\lambda+1}(x_{\lambda+1}, w^1_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, R_{\lambda+1}, OF_{\lambda+1}) \\ &= h\mathsf{E}_{\zeta_{\lambda+1}}[x_{\lambda+1} - \zeta_{\lambda+1}]^+ + b\mathsf{E}_{\zeta_{\lambda+1}}[\zeta_{\lambda+1} - x_{\lambda+1}]^+ + c(y^{\lambda*}_{\lambda+1} - IP_{\lambda+1}) \\ &\quad + \mathsf{E}_{\zeta_{\lambda+1}} NV^{\lambda}_{\lambda}(x_{\lambda+1} + w^1_{\lambda+1} - \zeta_{\lambda+1}, w^2_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, y^{\lambda*}_{\lambda+1} - IP_{\lambda+1}) \end{aligned}$$

In cases 1.1 and 1.2, $NV^{\lambda}_{\lambda+1}(.) = G^{\lambda}_{\lambda+1}(., R_{\lambda+1}, OF_{\lambda+1})$. Therefore, $\frac{\partial NV^{\lambda}_{\lambda+1}}{\partial x_{\lambda+1}} = \frac{\partial G^{\lambda}_{\lambda+1}}{\partial x_{\lambda+1}}$ and $\frac{\partial NV^{\lambda}_{\lambda+1}}{\partial w^i_{\lambda+1}} = \frac{\partial G^{\lambda}_{\lambda+1}}{\partial w^i_{\lambda+1}}, i = 1, 2, ..., \lambda - 1$.

Case 1.3: $\frac{R_{\lambda+1}}{c} < y^{\lambda*}_{\lambda+1}$

$$\begin{aligned} NV^{\lambda}_{\lambda+1}(x_{\lambda+1}, w^1_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}) &= h\mathsf{E}_{\zeta_{\lambda+1}}[x_{\lambda+1} - \zeta_{\lambda+1}]^+ + b\mathsf{E}_{\zeta_{\lambda+1}}[\zeta_{\lambda+1} - x_{\lambda+1}]^+ \\ &\quad + c(y^{\lambda*}_{\lambda+1} - IP_{\lambda+1}) + \mathsf{E}_{\zeta_{\lambda+1}} NV^{\lambda}_{\lambda}(x_{\lambda+1} + w^1_{\lambda+1} - \zeta_{\lambda+1}, w^2_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, y^{\lambda*}_{\lambda+1} - IP_{\lambda+1}) \end{aligned}$$

$$\begin{aligned} G^{\lambda}_{\lambda+1}(x_{\lambda+1}, w^1_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, R_{\lambda+1}, OF_{\lambda+1}) &= h\mathsf{E}_{\zeta_{\lambda+1}}[x_{\lambda+1} - \zeta_{\lambda+1}]^+ + b\mathsf{E}_{\zeta_{\lambda+1}}[\zeta_{\lambda+1} - x_{\lambda+1}]^+ \\ &+ R_{\lambda+1} - cIP_{\lambda+1} + \mathsf{E}_{\zeta_{\lambda+1}} NV^{\lambda}_{\lambda}(x_{\lambda+1} + w^1_{\lambda+1} - \zeta_{\lambda+1}, w^2_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, \frac{R_{\lambda+1}}{c} - IP_{\lambda+1}) \end{aligned}$$

Using expression (A.9), we see that

$$\begin{aligned} \frac{\partial NV^{\lambda}_{\lambda+1}}{\partial x_{\lambda+1}} &= -c + hF_{\lambda+1}(x_{\lambda+1}) - b\bar{F}_{\lambda+1}(x_{\lambda+1}) + \mathsf{E}_{\zeta_{\lambda+1}} \frac{\partial \tilde{f}_{\lambda}(x_{\lambda+1} + w^1_{\lambda+1} - \zeta_{\lambda+1})}{\partial x_{\lambda+1}} \\ &+ ... + \mathsf{E}_{\zeta_{\lambda+1},...,\zeta_2} \frac{\partial \tilde{f}_1(y^{\lambda *}_{\lambda+1} - \zeta_{\lambda+1} - ... - \zeta_2)}{\partial x_{\lambda+1}} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial G^{\lambda}_{\lambda+1}}{\partial x_{\lambda+1}} &= -c + hF_{\lambda+1}(x_{\lambda+1}) - b\bar{F}_{\lambda+1}(x_{\lambda+1}) + \mathsf{E}_{\zeta_{\lambda+1}} \frac{\partial \tilde{f}_{\lambda}(x_{\lambda+1} + w^1_{\lambda+1} - \zeta_{\lambda+1})}{\partial x_{\lambda+1}} \\ &+ ... + \mathsf{E}_{\zeta_{\lambda+1},...,\zeta_2} \frac{\partial \tilde{f}_1(\frac{R_{\lambda+1}}{c} - \zeta_{\lambda+1} - ... - \zeta_2)}{\partial x_{\lambda+1}} \end{aligned}$$

The last term is 0 in both the above equations. Therefore $\frac{\partial NV^{\lambda}_{\lambda+1}(.)}{\partial x_{\lambda+1}} = \frac{\partial G^{\lambda}_{\lambda+1}(.,R_{\lambda+1},OF_{\lambda+1})}{\partial x_{\lambda+1}}$. Similarly,

$$\frac{\partial NV^{\lambda}_{\lambda+1}}{\partial w^1_{\lambda+1}} = -c + \mathsf{E}_{\zeta_{\lambda+1}} \frac{\partial \tilde{f}_{\lambda}(x_{\lambda+1} + w^1_{\lambda+1} - \zeta_{\lambda+1})}{\partial w^1_{\lambda+1}} + ... + \mathsf{E}_{\zeta_{\lambda+1},...,\zeta_2} \frac{\partial \tilde{f}_1(y^{\lambda *}_{\lambda+1} - \zeta_{\lambda+1} - ... - \zeta_2)}{\partial w^1_{\lambda+1}}$$

and

$$\frac{\partial G^{\lambda}_{\lambda+1}}{\partial w^1_{\lambda+1}} = -c + \mathsf{E}_{\zeta_{\lambda+1}} \frac{\partial \tilde{f}_{\lambda}(x_{\lambda+1} + w^1_{\lambda+1} - \zeta_{\lambda+1})}{\partial w^1_{\lambda+1}} + ... + \mathsf{E}_{\zeta_{\lambda+1},...,\zeta_2} \frac{\partial \tilde{f}_1(\frac{R_{\lambda+1}}{c} - \zeta_{\lambda+1} - ... - \zeta_2)}{\partial w^1_{\lambda+1}}$$

Again, the last term is 0 in both the above equations. Therefore, $\frac{\partial NV^{\lambda}_{\lambda+1}}{\partial w^1_{\lambda+1}} = \frac{\partial G^{\lambda}_{\lambda+1}}{\partial w^1_{\lambda+1}}$. For any given $(R_{\lambda+1}, OF_{\lambda+1})$, following similar steps, it is easy to show that $\frac{\partial NV^{\lambda}_{\lambda+1}(.)}{\partial w^i_{\lambda+1}} = \frac{\partial G^{\lambda}_{\lambda+1}(.,R_{\lambda+1},OF_{\lambda+1})}{\partial w^i_{\lambda+1}}$ $\forall\ i = 1, ..., \lambda - 1$.

Let the induction hypothesis be that G^{λ}_t is jointly convex in $(x_t, w^1_t, ..., w^{\lambda-1}_t)$, given R_t and OF_t. Furthermore, let $\frac{\partial NV^{\lambda}_t(.)}{\partial x_t} = \frac{\partial G^{\lambda}_t(.,R_t,OF_t)}{\partial x_t}$ and $\frac{\partial NV^{\lambda}_t(.)}{\partial w^i_t} = \frac{\partial G^{\lambda}_t(.,R_t,OF_t)}{\partial w^i_t}$ for any given (R_t, OF_t) and $i = 1, 2, ..., \lambda - 1$. We will show that they hold true for $t + 1$ as well.

Fix (R_{t+1}, OF_{t+1}). We know that

$$G^{\lambda}_{t+1}(x_{t+1}, w^1_{t+1}, ..., w^{\lambda-1}_{t+1}, R_{t+1}, OF_{t+1})$$

$$= \min_{0 \le z_{t+1} \le \frac{R_{t+1}}{c} - IP_{t+1}} \left\{ \begin{array}{l} cz_{t+1} + b\mathsf{E}_{\zeta_{t+1}}[\zeta_{t+1} - x_{t+1}]^+ + h\mathsf{E}_{\zeta_{t+1}}[x_{t+1} - \zeta_{t+1}]^+ \\ +\mathsf{E}_{OF_t|OF_{t+1}}\mathsf{E}_{\zeta_{t+1}} G^{\lambda}_t(x_{t+1} - \zeta_{t+1} + w^1_{t+1}, \\ \qquad ..., w^{\lambda-1}_{t+1}, z_{t+1}, R_{t+1} - c\zeta_{t+1} + OF_{t+1} - OF_t, OF_t) \end{array} \right\} \tag{A.12}$$

Since G^{λ}_t is jointly convex in $(x_t, w^1_t, ..., w^{\lambda-1}_t)$, it follows that the function to be minimized in equation (A.12) is jointly convex in $(x_{t+1}, w^1_{t+1}, ..., w^{\lambda-1}_{t+1}, z_{t+1})$. Again, using proposition B-4 from Heyman and Sobel (1984), we see that G^{λ}_{t+1} is jointly convex in $x_{t+1}, w^1_{t+1}, ..., w^{\lambda-1}_{t+1}$ for fixed (R_{t+1}, OF_{t+1}). This completes the first part of the induction.

From the convexity of G^{λ}_{t+1} and the induction assumptions $\frac{\partial NV^{\lambda}_t}{\partial x_t} = \frac{\partial G^{\lambda}_t}{\partial x_t}$ and $\frac{\partial NV^{\lambda}_t}{\partial w^i_t} = \frac{\partial G^{\lambda}_t}{\partial w^i_t}$, it follows that z^*_{t+1}=max(min($y^{\lambda *}_{t+1} - IP_{t+1}, \frac{R_{t+1}}{c} - IP_{t+1}$),0) minimizes equation (A.12).

To prove the other part of the induction for t+1, the following recursive equation for NV^{λ}_t, $t > \lambda + 1$, would be useful.

$$\begin{aligned} &NV^{\lambda}_t(x_{t+1} - \zeta_{t+1} + w^1_{t+1}, w^2_{t+1}, ..., w^{\lambda-1}_{t+1}, z_{t+1}) \\ &\qquad = c(max(y^{\lambda *}_t - IP_t, 0)) + \tilde{f}_t(x_{t+1} - \zeta_{t+1} + w^1_{t+1}) \\ &\qquad\qquad + \mathsf{E}_{\zeta_t}\tilde{f}_{t-1}(x_{t+1} + w^1_{t+1} + w^2_{t+1} - \zeta_{t+1} - \zeta_t) \\ &\qquad\qquad + ... + \mathsf{E}_{\zeta_t,...,\zeta_{t-\lambda+2}}\tilde{f}_{t-\lambda+1}(IP_{t+1} + z_{t+1} - \zeta_{t+1} - \zeta_t - ... - \zeta_{t-\lambda+2}) \\ &\qquad\qquad + \mathsf{E}_{\zeta_t,...,\zeta_{t-\lambda+1}} NV^{\lambda}_{t-\lambda}(x_{t-\lambda}, w^1_{t-\lambda}, ..., w^{\lambda-1}_{t-\lambda}) \end{aligned} \tag{A.13}$$

Now consider three cases similar to the ones we considered earlier for $t = \lambda + 1$.

Case 2.1: $IP_{t+1} \ge y^{\lambda *}_{t+1}$.

$$\begin{aligned} G^{\lambda}_{t+1}(x_{t+1}, w^1_{t+1}, ..., w^{\lambda-1}_{t+1}, R_{t+1}, OF_{t+1}) &= b\mathsf{E}_{\zeta_{t+1}}[\zeta_{t+1} - x_{t+1}]^+ + h\mathsf{E}_{\zeta_{t+1}}[x_{t+1} - \zeta_{t+1}]^+ \\ + \mathsf{E}_{OF_t|OF_{t+1}}\mathsf{E}_{\zeta_{t+1}} G^{\lambda}_t(x_{t+1} - \zeta_{t+1} + w^1_{t+1}, &..., w^{\lambda-1}_{t+1}, 0, R_{t+1} - c\zeta_{t+1} + OF_{t+1} - OF_t, OF_t) \end{aligned}$$

and

$$\begin{aligned} NV^{\lambda}_{t+1}(x_{t+1}, w^{1}_{t+1}, ..., w^{\lambda-1}_{t+1}) &= b\mathsf{E}_{\zeta_{\lambda+1}}[\zeta_{\lambda+1} - x_{\lambda+1}]^{+} + h\mathsf{E}_{\zeta_{\lambda+1}}[x_{\lambda+1} - \zeta_{\lambda+1}]^{+} \\ &\quad + \mathsf{E}_{\zeta_{\lambda+1}} NV^{\lambda}_{t}(x_{\lambda+1} - \zeta_{\lambda+1} + w^{1}_{\lambda+1}, w^{2}_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, 0) \end{aligned}$$

Since $\frac{\partial NV^{\lambda}_{t}(.)}{\partial x_t} = \frac{\partial G^{\lambda}_{t}(.,R_t,OF_t)}{\partial x_t}$ and $\frac{\partial NV^{\lambda}_{t}(.)}{\partial w^{i}_{t}} = \frac{\partial G^{\lambda}_{t}(.,R_t,OF_t)}{\partial w^{i}_{t}}$ for $i = 1, 2, ..., \lambda - 1$, it follows that, for fixed (R_{t+1}, OF_{t+1}), $\frac{\partial NV^{\lambda}_{t+1}(.)}{\partial x_{t+1}} = \frac{\partial G^{\lambda}_{t+1}(.,R_{t+1},OF_{t+1})}{\partial x_{t+1}}$ and $\frac{\partial NV^{\lambda}_{t+1}(.)}{\partial w^{i}_{t+1}} = \frac{\partial G^{\lambda}_{t+1}(.,R_{t+1},OF_{t+1})}{\partial w^{i}_{t+1}}$ $\forall\ i = 1, 2, ..., \lambda - 1$.

Case 2.2: $IP_{t+1} < y^{\lambda *}_{t+1}$, $\frac{R_{t+1}}{c} \geq y^{\lambda *}_{t+1}$.

$$\begin{aligned} &G^{\lambda}_{t+1}(x_{t+1}, w^{1}_{t+1}, ..., w^{\lambda-1}_{t+1}, R_{t+1}, OF_{t+1}) \\ &= c(y^{\lambda *}_{t+1} - IP_{t+1}) + b\mathsf{E}_{\zeta_{t+1}}[\zeta_{t+1} - x_{t+1}]^{+} + h\mathsf{E}_{\zeta_{t+1}}[x_{t+1} - \zeta_{t+1}]^{+} \\ &+ \mathsf{E}_{OF_t|OF_{t+1}} \mathsf{E}_{\zeta_{t+1}} G^{\lambda}_{t}(x_{t+1} - \zeta_{t+1} + w^{1}_{t+1}, ..., y^{\lambda *}_{t+1} - IP_{t+1}, R_{t+1} - c\zeta_{t+1} + OF_{t+1} - OF_t, OF_t) \end{aligned}$$

and

$$\begin{aligned} &NV^{\lambda}_{t+1}(x_{t+1}, w^{1}_{t+1}, ..., w^{\lambda-1}_{t+1}) \\ &\quad = c(y^{\lambda *}_{t+1} - IP_{t+1}) + b\mathsf{E}_{\zeta_{\lambda+1}}[\zeta_{\lambda+1} - x_{\lambda+1}]^{+} + h\mathsf{E}_{\zeta_{\lambda+1}}[x_{\lambda+1} - \zeta_{\lambda+1}]^{+} \\ &\qquad + \mathsf{E}_{\zeta_{\lambda+1}} NV^{\lambda}_{t}(x_{\lambda+1} - \zeta_{\lambda+1} + w^{1}_{\lambda+1}, w^{2}_{\lambda+1}, ..., w^{\lambda-1}_{\lambda+1}, y^{\lambda *}_{t+1} - IP_{t+1}) \end{aligned}$$

Again, the induction hypothesis concerning the derivatives of NV^{λ}_{t} and G^{λ}_{t} directly yields $\frac{\partial NV^{\lambda}_{t+1}(.)}{\partial x_{t+1}} = \frac{\partial G^{\lambda}_{t+1}(.,R_{t+1},OF_{t+1})}{\partial x_{t+1}}$ and $\frac{\partial NV^{\lambda}_{t+1}(.)}{\partial w^{i}_{t+1}} = \frac{\partial G^{\lambda}_{t+1}(.,R_{t+1},OF_{t+1})}{\partial w^{i}_{t+1}}$ $\forall\ i = 1, 2, ..., \lambda - 1$.

Case 2.3: $\frac{R_{t+1}}{c} < y^{\lambda *}_{t+1}$

$$\begin{aligned} &NV^{\lambda}_{t+1}(x_{t+1}, w^{1}_{t+1}, ..., w^{\lambda-1}_{t+1}) \\ &\quad = c(y^{\lambda *}_{t+1} - IP_{t+1}) + b\mathsf{E}_{\zeta_{t+1}}[\zeta_{t+1} - x_{t+1}]^{+} + h\mathsf{E}_{\zeta_{t+1}}[x_{t+1} - \zeta_{t+1}]^{+} \\ &\qquad + \mathsf{E}_{\zeta_{t+1}} NV^{\lambda}_{t}(x_{t+1} - \zeta_{t+1} + w^{1}_{t+1}, ..., w^{\lambda-1}_{t+1}, y^{\lambda *}_{t+1} - IP_{t+1}) \end{aligned}$$

From equation (A.13), we get

$$
\begin{aligned}
NV_t^\lambda&(x_{t+1}-\zeta_{t+1}+w_{t+1}^1,w_{t+1}^2,...,w_{t+1}^{\lambda-1},y_{t+1}^{\lambda*}-IP_{t+1})\\
&=c(max(y_t^{\lambda*}-(y_{t+1}^{\lambda*}-\zeta_{t+1}),0))+\tilde{f}_t(x_{t+1}-\zeta_{t+1}+w_{t+1}^1)\\
&\quad+\mathsf{E}_{\zeta_t}\tilde{f}_{t-1}(x_{t+1}+w_{t+1}^1+w_{t+1}^2-\zeta_{t+1}-\zeta_t)\\
&\quad+...+\mathsf{E}_{\zeta_t,...,\zeta_{t-\lambda+2}}\tilde{f}_{t-\lambda+1}(y_{t+1}^{\lambda*}-\zeta_{t+1}-\zeta_t-...-\zeta_{t-\lambda+2})\\
&\quad+\mathsf{E}_{\zeta_t,...,\zeta_{t-\lambda+1}}NV_{t-\lambda}^\lambda(x_{t-\lambda},w_{t-\lambda}^1,...,w_{t-\lambda}^{\lambda-1}) \qquad \text{(A.14)}
\end{aligned}
$$

In equation (A.14), $x_{t-\lambda}{=}max(y_t^{\lambda*}-(y_{t+1}^{\lambda*}-\zeta_{t+1}),0)+y_{t+1}^{\lambda*}-\zeta_{t+1}-\zeta_t-...-\zeta_{t-\lambda+1}$, i.e., it is a function only of $y_{t+1}^{\lambda*},y_t^{\lambda*},\zeta_{t+1},\zeta_t,...,\zeta_{t-\lambda+1}$. $w_{t-\lambda}^{\lambda-1}$, which is the order placed in $t-1$, is a function only of $y_{t+1}^{\lambda*},y_t^{\lambda*},y_{t-1}^{\lambda*},\zeta_{t+1}$, and ζ_t. Following similar logic, it is easy to see that the state variables $(x_{t-\lambda},w_{t-\lambda}^1,...,w_{t-\lambda}^{\lambda-1})$ are independent of $(x_{t+1},w_{t+1}^1,...,w_{t+1}^{\lambda-1})$.

$$
\begin{aligned}
&G_{t+1}^\lambda(x_{t+1},w_{t+1}^1,...,w_{t+1}^{\lambda-1},R_{t+1},OF_{t+1})\\
&=R_{t+1}-cIP_{t+1}+b\mathsf{E}_{\zeta_{t+1}}[\zeta_{t+1}-x_{t+1}]^++h\mathsf{E}_{\zeta_{t+1}}[x_{t+1}-\zeta_{t+1}]^+\\
&+\mathsf{E}_{OF_t|OF_{t+1}}\mathsf{E}_{\zeta_{t+1}}G_t^\lambda(x_{t+1}-\zeta_{t+1}+w_{t+1}^1,...,\frac{R_{t+1}}{c}-IP_{t+1},R_{t+1}-c\zeta_{t+1}+OF_{t+1}-OF_t,OF_t)
\end{aligned}
$$

Since the derivatives of G_t^λ and NV_t^λ are equal (for a given (R_t,OF_t)) by the induction hypothesis, to analyze the derivative of $G_t^\lambda(x_{t+1}-\zeta_{t+1}+w_{t+1}^1,...,w_{t+1}^{\lambda-1},\frac{R_{t+1}}{c}-IP_{t+1},R_{t+1}-c\zeta_{t+1}+OF_{t+1}-OF_t,OF_t)$ with respect to state variables $x_{t+1},w_{t+1}^1,...,w_{t+1}^{\lambda-1}$, we focus on the following expression.

$$
\begin{aligned}
NV_t^\lambda&(x_{t+1}-\zeta_{t+1}+w_{t+1}^1,w_{t+1}^2,...,w_{t+1}^{\lambda-1},\frac{R_{t+1}}{c}-IP_{t+1})\\
&=c(max(y_t^{\lambda*}-(\frac{R_{t+1}}{c}-\zeta_{t+1}),0))+\tilde{f}_t(x_{t+1}-\zeta_{t+1}+w_{t+1}^1)\\
&\quad+\mathsf{E}_{\zeta_t}\tilde{f}_{t-1}(x_{t+1}+w_{t+1}^1+w_{t+1}^2-\zeta_{t+1}-\zeta_t)\\
&\quad+...+\mathsf{E}_{\zeta_t,...,\zeta_{t-\lambda+2}}\tilde{f}_{t-\lambda+1}(\frac{R_{t+1}}{c}-\zeta_{t+1}-\zeta_t-...-\zeta_{t-\lambda+2})\\
&\quad+\mathsf{E}_{\zeta_t,...,\zeta_{t-\lambda+1}}NV_{t-\lambda}^\lambda(x_{t-\lambda},w_{t-\lambda}^1,...,w_{t-\lambda}^{\lambda-1}) \qquad \text{(A.15)}
\end{aligned}
$$

In equation (A.15), $x_{t-\lambda}{=}max(y_t^{\lambda*}-(\frac{R_{t+1}}{c}-\zeta_{t+1}),0)+\frac{R_{t+1}}{c}-\zeta_{t+1}-\zeta_t-...-\zeta_{t-\lambda+1}$, i.e., it is a function only of $\frac{R_{t+1}}{c}, y_t^{\lambda *}, \zeta_{t+1}, \zeta_t, ..., \zeta_{t-\lambda+1}$. $w_{t-\lambda}^{\lambda-1}$ is a function only of $\frac{R_{t+1}}{c}, y_t^{\lambda *}, y_{t-1}^{\lambda *}, \zeta_{t+1}$ and $\zeta_{t-\lambda}$. Using similar logic, we see that the state variables $(x_{t-\lambda}, w_{t-\lambda}^1, ..., w_{t-\lambda}^{\lambda-1})$ are independent of $(x_{t+1}, w_{t+1}^1, ..., w_{t+1}^{\lambda-1})$ in this case as well. Therefore it follows that the derivatives of $NV_t^\lambda(x_{t+1} - \zeta_{t+1} + w_{t+1}^1, w_{t+1}^2, ..., w_{t+1}^{\lambda-1}, y_{t+1}^{\lambda *} - IP_{t+1})$ and $NV_t^\lambda(x_{t+1} - \zeta_{t+1} + w_{t+1}^1, w_{t+1}^2, ..., w_{t+1}^{\lambda-1}, \frac{R_{t+1}}{c} - IP_{t+1})$ with respect to x_{t+1} and $w_{t+1}^i, i = 1, 2, ..., \lambda - 1$ are equal. This in turn implies that $\frac{\partial NV_{t+1}^\lambda(.)}{\partial x_{t+1}} = \frac{\partial G_{t+1}^\lambda(., R_{t+1}, OF_{t+1})}{\partial x_{t+1}}$ and $\frac{\partial NV_{t+1}^\lambda(.)}{\partial w_{t+1}^i} = \frac{\partial G_{t+1}^\lambda(., R_{t+1}, OF_{t+1})}{\partial w_{t+1}^i}$ for any given (R_{t+1}, OF_{t+1}) and $i = 1, 2, ..., \lambda - 1$. Hence the induction is complete.

Appendix B

Proofs for results in Chapter 3

Proof of Lemma 3

The proof proceeds through induction. Let us begin with t=0. When $r_0 < n_0^2$, it is easy to see from equation (3.2) that V_0 is linear in S_0. The linearity holds true for the other two cases as well, i.e., $n_0^2 \leq r_0 < n_0^1 + n_0^2$ and $r_0 \geq n_0^1 + n_0^2$. Hence the joint convexity holds for V_0.

Now assume that V_{t-1} is jointly convex in S_{t-1} for fixed OF_{t-1}. Since S_{t-1} is a linear function of S_t, a^1 and a^2, and convexity is preserved under expectations, it follows that the function to be minimized in equation (3.1) is jointly convex in (S_t, a^1, a^2). Now consider the constraint set $\mathbf{C}$=$\{(a^1, a^2) : 0 \leq a^1 \leq n_t^1, 0 \leq a^2 \leq n_t^2, a^1 + a^2 \leq r_t\}$. Clearly, $\mathbf{C}$ is a convex set. Therefore, using proposition B-4 from Heyman and Sobel (1984), we see that V_t is jointly convex in S_t for fixed OF_t.

Proof of Theorem 5

When $n_t^2 \geq r_t$, the result follows directly from the fact that $a^{2*} = \min\{n_t^2, r_t\}$ and $a^1 + a^2 \leq r_t$. If $n_t^2 < r_t$, then, clearly $a^{2*} = n_t^2$. a^{1*} is the solution to the following problem.

$$\min_{0 \leq a^1 \leq \min\{n_t^1, r_t - n_t^2\}} \quad \left\{C_t\left(a^1, S_t, OF_t\right)\right\}$$

From Lemma 3, we have that C_t is jointly convex in a^1 and S_t for fixed OF_t. Then it follows directly that $a^{1*} = \max\{a^1 : \frac{\partial C_t}{\partial a^1} \leq 0, 0 \leq a^1 \leq \min\{n_t^1, r_t - n_t^2\}\}$ is indeed the optimal allocation in period t.

Proof of Proposition 3

To prove the monotonicity of a^{1*} with respect to n_t^1 and r_t, we vary them one at a time. First fix n_t^1 and let $\hat{r}_t > r_t$. Consider the following three cases.

Case 1.1: $n_t^2 > \hat{r}_t$. In this case, $a^{1*}(\hat{r}_t)$=$a^{1*}(r_t)$=0.

Case 1.2: $r_t \leq n_t^2 \leq \hat{r}_t$. Then, $a^{1*}(r_t) = 0 \leq a^{1*}(\hat{r}_t)$.

Case 1.3: $n_t^2 < r_t$. Then,

$$V_t\left(n_t^1, n_t^2, r_t, OF_t\right) = \min_{0 \leq a^1 \leq \min\{n_t^1, r_t - n_t^2\}} \left\{C_t\left(a^1, n_t^1, n_t^2, r_t, OF_t\right)\right\} \tag{B.1}$$

and $V_t\left(n_t^1, n_t^2, \hat{r}_t, OF_t\right)$ is obtained by replacing r_t with $\hat{r}_t$ in equation (B.1). From equation (3.3), we see that

$$\frac{\partial C_t}{\partial a^1} = -\hat{b}^1 + \mathsf{E}_{n_{t-1}^N}\mathsf{E}_{OF_{t-1}|OF_t}(-\alpha_{11})\frac{\partial V_{t-1}}{\partial n_{t-1}^1} + (-\alpha_{12})\frac{\partial V_{t-1}}{\partial n_{t-1}^2} + (-1)\frac{\partial V_{t-1}}{\partial r_{t-1}}$$

The convexity of V_{t-1} implies that $(\alpha_{11})\frac{\partial V_{t-1}}{\partial n_{t-1}^1} + (\alpha_{12})\frac{\partial V_{t-1}}{\partial n_{t-1}^2} + (1)\frac{\partial V_{t-1}}{\partial r_{t-1}}$ is increasing in r_{t-1} for fixed n_{t-1}^1, n_{t-1}^2 and OF_{t-1}. This yields $\frac{\partial C_t\left(a^1, n_t^1, n_t^2, r_t, OF_t\right)}{\partial a^1} \geq \frac{\partial C_t\left(a^1, n_t^1, n_t^2, \hat{r}_t, OF_t\right)}{\partial a^1}$, which then directly implies that $a^{1*}(\hat{r}_t) \geq a^{1*}(r_t)$.

We use a similar approach to prove the monotonicity of a^{1*} with respect to n_t^1. Fix r_t and let $\hat{n}_t^1 > n_t^1$. Consider the following cases.

Case 2.1: $n_t^2 > r_t$. In this case, $a^{1*}(\hat{n}_t^1)$=$a^{1*}(n_t^1)$=0.

Case 2.2: $n_t^2 \leq r_t$. Then, $V_t\left(n_t^1, n_t^2, r_t, OF_t\right)$ is given by (B.1) while $V_t\left(\hat{n}_t^1, n_t^2, r_t, OF_t\right)$ is obtained by replacing n_t^1 with $\hat{n}_t^1$ in equation (B.1). From the convexity of V_{t-1}, we have that $(\alpha_{11})\frac{\partial V_{t-1}}{\partial n_{t-1}^1} + (\alpha_{12})\frac{\partial V_{t-1}}{\partial n_{t-1}^2} + (1)\frac{\partial V_{t-1}}{\partial r_{t-1}}$ is increasing in n_{t-1}^1 and n_{t-1}^2 for fixed r_{t-1} and OF_{t-1}. Therefore, $\frac{\partial C_t\left(a^1, n_t^1, n_t^2, r_t, OF_t\right)}{\partial a^1} \geq \frac{\partial C_t\left(a^1, \hat{n}_t^1, n_t^2, r_t, OF_t\right)}{\partial a^1}$, and hence $a^{1*}(\hat{n}_t^1) \geq a^{1*}(n_t^1)$, completing the proof.

Proof of Corollary 1

The first result follows directly from the following facts: a^{1*} is monotone increasing with respect to r_t for fixed N_t (Proposition (3)) and $0 \leq a^1 \leq n_t^1$. We move on to the second result. The convexity of C_t, combined with our assumption that $a^2 = n_t^2$ and $a^1 = r_t - n_t^2$ are optimal for r_t, lead to the following series of inequalities:

$$\begin{aligned}\frac{\partial C_t\left(a^1, N_t, \hat{r}_t, OF_t\right)}{\partial a^1}\Big|_{a^1=\hat{r}_t-n_t^2} &\leq \frac{\partial C_t\left(a^1, N_t, r_t, OF_t\right)}{\partial a^1}\Big|_{a^1=\hat{r}_t-n_t^2} \\ &\leq \frac{\partial C_t\left(a^1, N_t, r_t, OF_t\right)}{\partial a^1}\Big|_{a^1=r_t-n_t^2} \\ &\leq 0.\end{aligned}$$

Hence, $a^1 = \hat{r}_t - n_t^2$ is optimal for $\hat{r}_t$. The third result follows from Proposition (3), which demonstrates that $a^{1*}(r_t) \geq a^{1*}(\hat{r}_t)$. The fourth result is obtained by combining the following facts: $a^{1*}(r_t) \geq a^{1*}(\hat{r}_t)$ and $0 \leq \hat{a}^1 \leq \hat{r}_t - n_t^2$, since $\hat{r}_t \geq n_t^2$.

Proof of Lemma 4

The proof proceeds through induction and a sample path argument. We start with t=0. Notice that TF_0=r_0 and $J_0\left(S_0, TF_0\right)$=$V_0\left(S_0, 0\right)$. Then, the convexity of V_0 in S_0 directly implies the convexity of J_0 in (S_0, TF_0).

Now, suppose that $J_t\left(S_{t-1}, TF_{t-1}\right)$ is jointly convex in S_{t-1} and TF_{t-1}. Pick any $S_t = (n_t^1, n_t^2, r_t)$, $\hat{S}_t = (\hat{n}_t^1, \hat{n}_t^2, \hat{r}_t)$, TF_t, $T\hat{F}_t$, $a^1, \hat{a}^1, a^2, \hat{a}^2$ and $0 \leq \lambda \leq 1$. Analogous to $n_{t-1}^1, n_{t-1}^2, r_{t-1}$ and TF_{t-1}, define $\hat{n}_{t-1}^1, \hat{n}_{t-1}^2, \hat{r}_{t-1}$ and $T\hat{F}_{t-1}$ in terms of $\hat{n}_t^1, \hat{n}_t^2, \hat{r}_t, \hat{a}^1, \hat{a}^2$. For brevity, also define $\bar{n}_{t-1}^1 = \lambda n_{t-1}^1 + (1-\lambda)\hat{n}_{t-1}^1$, $\bar{n}_{t-1}^2 = \lambda n_{t-1}^2 + (1-\lambda)\hat{n}_{t-1}^2$, $\bar{r}_{t-1} = \lambda r_{t-1} + (1-\lambda)\hat{r}_{t-1}$, $T\bar{F}_{t-1} = \lambda TF_{t-1} + (1-\lambda)T\hat{F}_{t-1}$.

The first step in proving the convexity of J_t makes use of the SSCV property. Since $f_{t-1}\mid_{TF_t-r_t}$ is SSCV, we have that $f_{t-1}\mid_{\lambda(TF_t-r_t)+(1-\lambda)(T\hat{F}_t-\hat{r}_t)} \geq \lambda f_{t-1}\mid_{TF_t-r_t} + (1-\lambda) f_{t-1}\mid_{T\hat{F}_t-\hat{r}_t}$ w.p.1, provided all the three random variables are defined on a common probability space. This results in the following series of inequalities.

$$\begin{aligned}&J_{t-1}(\bar{n}_{t-1}^1, \bar{n}_{t-1}^2, \lambda(r_t - a^1 - a^2) + (1-\lambda)(\hat{r}_t - \hat{a}^1 - \hat{a}^2) + f_{t-1}\mid_{\lambda(TF_t-r_t)+(1-\lambda)(T\hat{F}_t-\hat{r}_t)}, T\bar{F}_{t-1}) \\ &\quad\leq J_{t-1}(\bar{n}_{t-1}^1, \bar{n}_{t-1}^2, \bar{r}_{t-1}, T\bar{F}_{t-1}) \\ &\quad\leq \lambda J_{t-1}\left(n_{t-1}^1, n_{t-1}^2, r_{t-1}, TF_{t-1}\right) + (1-\lambda) J_{t-1}(\hat{n}_{t-1}^1, \hat{n}_{t-1}^2, \hat{r}_{t-1}, T\hat{F}_{t-1}) \quad \text{w.p.1.}\end{aligned}$$

The first inequality stems from combining the following facts: (i) $f_{t-1}\mid_{TF_t-r_t}$ is SSCV and (ii) J_{t-1} is decreasing in r_{t-1} for fixed $n_{t-1}^1, n_{t-1}^2, n_{t-1}^D$ and TF_{t-1}. The second inequality comes from the induction assumption regarding the joint convexity of J_{t-1}. Since the above inequality holds almost surely, we have that

$$b^1 a^2 + \hat{b}^1(n_t^1 - a^1) + \hat{b}_t^2(n_t^2 - a^2) + + \mathsf{E}_{n_{t-1}^N} \mathsf{E}_{TF_t - r_t - f_{t-1}|TF_t - r_t} J_{t-1}\left(n_{t-1}^1, n_{t-1}^2, r_{t-1}, TF_{t-1}\right)$$

is jointly convex in $\left(a^1, a^2, S_t, TF_t\right)$. Then, using proposition B-4 from Heyman and Sobel (1984), we see that J_t is jointly convex in (S_t, TF_t).

Proof of Proposition 4

To show that a^{1*} is increasing with respect to n_t^1 and r_t, we vary them one at a time. First fix n_t^1. Let $\hat{r}_t > r_t$. We consider the following three cases.

Case 1.1: $n_t^2 > \hat{r}_t$. In this case, $a^{1*}(\hat{r}_t)$=$a^{1*}(r_t)$=0.

Case 1.2: $r_t \leq n_t^2 \leq \hat{r}_t$. Then, $a^{1*}(r_t) = 0 \leq a^{1*}(\hat{r}_t)$.

Case 1.3: $n_t^2 < r_t$. Then,

$$J_t\left(n_t^1, n_t^2, r_t, TF_t\right) = \min_{0 \leq a^1 \leq \min\{n_t^1, r_t - n_t^2\}} \left\{\tilde{C}_t\left(a^1, n_t^1, n_t^2, r_t, TF_t\right)\right\} \tag{B.2}$$

and $J_t\left(n_t^1, n_t^2, \hat{r}_t, TF_t\right)$ is obtained by replacing r_t with $\hat{r}_t$ in equation (B.2). From expression (3.6), we get

$$\frac{\partial \tilde{C}_t}{\partial a^1} = -\hat{b}^1 + \mathsf{E}_{n_{t-1}^N, TF_t - r_t - f_{t-1}|TF_t - r_t}(-\alpha_{11})\frac{\partial J_{t-1}}{\partial n_{t-1}^1} + (-\alpha_{12})\frac{\partial J_{t-1}}{\partial n_{t-1}^2} + (-1)\left[\frac{\partial J_{t-1}}{\partial r_{t-1}} + \frac{\partial J_{t-1}}{\partial TF_{t-1}}\right]$$

From the convexity of J_{t-1}, we have that $(-\alpha_{11})\frac{\partial J_{t-1}}{\partial n_{t-1}^1} + (-\alpha_{12})\frac{\partial J_{t-1}}{\partial n_{t-1}^2} + (-1)\left[\frac{\partial J_{t-1}}{\partial r_{t-1}} + \frac{\partial J_{t-1}}{\partial TF_{t-1}}\right]$ is decreasing in r_{t-1} for fixed n_{t-1}^1, n_{t-1}^2 and TF_{t-1}. Now, we make use of the assumption that $r_t + f_{t-1}$ is stochastically increasing in r_t for fixed TF_t. Given this assumption, $\hat{r}_t + f_{t-1} |_{TF_t - \hat{r}_t} \geq r_t + f_{t-1} |_{TF_t - r_t}$ w.p.1, provided the random variables $f_{t-1} |_{TF_t - \hat{r}_t}$ and $f_{t-1} |_{TF_t - r_t}$ are defined on the same probability space. This implies that

$$(-\alpha_{11})\frac{\partial J_{t-1}}{\partial n^1_{t-1}} + (-\alpha_{12})\frac{\partial J_{t-1}}{\partial n^2_{t-1}} + (-1)\left[\frac{\partial J_{t-1}}{\partial r_{t-1}} + \frac{\partial J_{t-1}}{\partial TF_{t-1}}\right] |_{(n^1_{t-1}, n^2_{t-1}, r_{t-1}, TF_{t-1})}$$

$$\geq (-\alpha_{11})\frac{\partial J_{t-1}}{\partial n^1_{t-1}} + (-\alpha_{12})\frac{\partial J_{t-1}}{\partial n^2_{t-1}} + (-1)\left[\frac{\partial J_{t-1}}{\partial r_{t-1}} + \frac{\partial J_{t-1}}{\partial TF_{t-1}}\right] |_{(n^1_{t-1}, n^2_{t-1}, \hat{r}_{t-1}, TF_{t-1})} \quad \text{w.p.1.}$$

Since the above inequality holds almost surely, we see that $\frac{\partial \tilde{C}_t\left(a^1, N_t, r_t, OF_t\right)}{\partial a^1} \geq \frac{\partial \tilde{C}_t\left(a^1, N_t, \hat{r}_t, OF_t\right)}{\partial a^1}$. Hence, $a^{1*}(\hat{r}_t) \geq a^{1*}(r_t)$. To prove the second part of the result, fix r_t and consider $\hat{n}^1_t > n^1_t$. In this case, $a^{1*}(\hat{n}^1_t) \geq a^{1*}(n^1_t)$ follows using an approach identical to the one used in Proposition 3.

Proof of Corollary 2

Identical to proof of Corollary 1.

Proof of Proposition 5

We use a sample path approach to prove this result. To make the exposition clear, let us define vector $\bar{R^n_A} = (\bar{r}^n_1, \bar{r}^n_2, ..., \bar{r}^n_t)$, n=1,2, where $\bar{r}^n_i$ is the funding available on–hand (for a particular sample path) at the beginning of period i under funding scenario n. Here, $A = ((a^1_1, a^2_1), (a^1_2, a^2_2), ..., (a^1_t, a^2_t))$ represents the vector of allocations (a^1_t, a^2_t) made in period t. Given $\bar{R^n_A}$, let $V^n_{t, \bar{R^n_A}}(S_t, OF_t)$ be the cost incurred in periods $1, 2, ..., t$ along a particular sample path under funding scenario n, following the allocation vector A. Let A^{n*} represent the optimal allocation vector along a particular sample path for funding scenario n. Now, given conditions (3.7) and (3.13), notice that A^{2*} is also feasible under scenario 1 along every sample path, but it may not be optimal. This implies that $V^2_{t, \bar{R^n}_{A^{2*}}}(S_t, OF_t) = V^1_{t, \bar{R^n}_{A^{2*}}}(S_t, OF_t) \geq V^1_{t, \bar{R^n}_{A^{1*}}}(S_t, OF_t)$ w.p.1. Since, this result holds for every sample path, the result also holds in expectation, i.e, $V^2_t(S_t, OF_t) \geq V^1_t(S_t, OF_t)$.

Proof of Proposition 6

We use the equivalent function J_t to prove the result. Since OF^1_1=OF^2_1=0, from equation (3.4), it follows that $J^2_2(n^1_2, n^2_2, r_2, TF_2) \geq J^1_2(n^1_2, n^2_2, r_2, TF_2)$. Let the induction assumption be $J^2_{t-1}(n^1_{t-1}, n^2_{t-1}, r_{t-1}, TF_{t-1}) \geq J^1_{t-1}(n^1_{t-1}, n^2_{t-1}, r_{t-1}, TF_{t-1})$. Now we make use of conditions

(3.7) and (3.8). Since the conditions hold for both funding scenarios, using Lemma 4, we see that J^1_{t-1} and J^2_{t-1} are jointly convex in (S_{t-1}, OF_{t-1}). Then,

$$\begin{aligned}
J^1_t(n^1_t, n^2_t, r_t, TF_t) = & \min_{\substack{0 \le a^1 \le n^1_t \\ 0 \le a^2 \le n^2_t \\ a^1 + a^2 \le r_t}} \left\{ \begin{array}{l} b^1 a^2 + \hat{b}^1(n^1_t - a^1) + \hat{b}^2_t(n^2_t - a^2) \\ \qquad + \mathsf{E}_{n^N_{t-1}} \mathsf{E}_{OF^1_{t-1}|OF_t} J^1_{t-1}(n^1_{t-1}, n^2_{t-1}, r_{t-1}, TF_{t-1}) \end{array} \right\} \\
\le & \min_{\substack{0 \le a^1 \le n^1_t \\ 0 \le a^2 \le n^2_t \\ a^1 + a^2 \le r_t}} \left\{ \begin{array}{l} b^1 a^2 + \hat{b}^1(n^1_t - a^1) + \hat{b}^2_t(n^2_t - a^2) \\ \qquad + \mathsf{E}_{n^N_{t-1}} \mathsf{E}_{OF^1_{t-1}|OF_t} J^2_{t-1}(n^1_{t-1}, n^2_{t-1}, r_{t-1}, TF_{t-1}) \end{array} \right\} \\
\le & \min_{\substack{0 \le a^1 \le n^1_t \\ 0 \le a^2 \le n^2_t \\ a^1 + a^2 \le r_t}} \left\{ \begin{array}{l} b^1 a^2 + \hat{b}^1(n^1_t - a^1) + \hat{b}^2_t(n^2_t - a^2) \\ \qquad + \mathsf{E}_{n^N_{t-1}} \mathsf{E}_{OF^2_{t-1}|OF_t} J^2_{t-1}(n^1_{t-1}, n^2_{t-1}, r_{t-1}, TF_{t-1}) \end{array} \right\} \\
= & J^2_t(n^1_t, n^2_t, r_t, TF_t)
\end{aligned}$$

The first inequality follows from the induction assumption and the second inequality is obtained by combining the facts that J^2_{t-1} is convex and $OF^2_{t-1}|OF_t \ge_{cvx} OF^1_{t-1}|OF_t$.

Proof of Proposition 7

Using Property 2, we see that conditions (3.15) and (3.16), combined, imply condition (3.13). The result then follows directly from Proposition 5.

Proof of Proposition 8

Using Property 2, we see that conditions (3.16) and (3.17), combined, imply $OF^1_{t-1} \mid_{OF_t = i}$ $\ge_{st} OF^2_{t-1} \mid_{OF_t = i}$. The result then follows from Proposition 5.

Appendix C

Proofs for results in Chapter 4

Proof of Lemma 5

Assuming an interior solution exists, s_i^* is the solution to

$$-o_e + (\theta Q + o_e)\bar{F}\left((\alpha_0 + \theta s_i)Q - o_e(B - s_i)\right) = 0.$$

When demand is $U \sim [0, D_u]$, s_i^* is given by

$$s_i^* = \frac{\theta D_u Q + (Bo_e - \alpha_0 Q)(o_e + Q)}{(o_e + Q\theta)^2}$$

Taking the derivative of s_i^* with respect to θ, we have

$$\frac{\partial s_i^*}{\partial \theta} = \frac{Q(D_u o_e - Bo_e^2 + \alpha_0 o_e Q) - Q\theta(D_u Q - \alpha_0 Q^2 + Bo_e Q)}{(o_e + Q\theta)^3}$$

which is positive if and only if $\theta < \frac{o_e(D_u - Bo_e + \alpha_0 Q)}{Q(D_u + Bo_e - \alpha_0 Q)}$. Hence the result.

Proof of Lemma 6

The expression for s_i^* is given in the proof of Lemma 5. Taking the derivative of s_i^* with respect to o_e, we have

$$\frac{\partial s_i^*}{\partial o_e} = \frac{-\left(Q(2D_u\theta - o_e(\alpha_0 + B\theta)) - Q^2(B\theta^2 + \alpha_0\theta)\right)}{(o_e + Q\theta)^3}$$

The above expression is positive if and only if $o_e \geq \frac{2D_u\theta - (\alpha_0 + B\theta)Q\theta}{(\alpha_0 + B\theta)}$, yielding the desired result.

Proof of Lemma 7

Taking the derivative of s_i^* with respect to Q, we have

$$\frac{\partial s_i^*}{\partial Q} = \frac{-\left(o_e^2(\alpha_0 + B\theta) - o_e(\theta(D_u - \alpha_0 Q) - BQ\theta^2) + D_u Q\theta^2\right)}{(o_e + Q\theta)^3}$$

The derivative is positive only if $Q \leq \frac{o_e(D_u\theta - o_e(\alpha_0 + B\theta))}{\theta(D_u\theta + o_e(\alpha_0 + B\theta))}$. However, $o_i^* = B - s_i^* \geq 0$ implies that this condition can never hold. Hence, s_i^* always decreases with Q.

Proof of Lemma 8

From the concavity of the objective function in s_i, we have that s_i^* is the maximum possible s_i, $0 \leq s_i \leq min\{B, (1-\alpha_0)/\theta\}$, such that

$$-o_e + (\theta Q + o_e)\bar{F}\left((\alpha_0 + \theta s_i)Q - o_e(B - s_i)\right) \geq 0.$$

This implies that for s_i^* to increase with μ, $\bar{F}_{\mu_2} \geq \bar{F}_{\mu_1}$ for $\mu_2 \geq \mu_1$. From the definition of first–order stochastic dominance, it is easy to see that $\bar{F}_{\mu_2} \geq \bar{F}_{\mu_1}$ if $D_{|\mu_2} \geq_{st} D_{|\mu_1}$. Hence when $D_{|\mu_2} \geq_{st} D_{|\mu_1}$ for $\mu_2 \geq \mu_1$, s_i^* always increases with μ. However, when $D_{|\mu_2} \geq_{st} D_{|\mu_1}$ does not hold, then whether or not s_i^* increases depends on the specific values of F_1, F_2,α_0,θ, o_e, B and Q. Hence s_i^* is not guaranteed to increase in general.

Proof of Lemma 9

The proof uses the following property of the $\geq_{var}$ ordering.

Property 3. *(Song 1994) Let X and Y be two random variables with distribution functions F and G respectively. If $X \geq_{var} Y$, then S(F-G)=1 with sign sequence +,-, i.e., F crosses G exactly once and the cross is from above.*

Let F_1 and F_2 be the CDFs corresponding to random variables $D_{|\sigma_1}$ and $D_{|\sigma_2}$ respectively. From the above result, we have that F_1 and F_2 intersect exactly once with sign sequence +,-. This implies that there exists a point z^* such that $F_2(z) >= F_1(z)$ for $z < z^*$ and $F_2(z) <= F_1(z)$ for $z >= z^*$. Let us denote $F_1(z^*)$ as δ. Then the result immediately follows from the fact that s_i^* is the maximum possible s_i, $0 \leq s_i \leq min\{B, (1-\alpha_0)/\theta\}$, such that

$$F\left((\alpha_0 + \theta s_i)Q - o_e(B - s_i)\right) \leq \frac{\theta Q}{\theta Q + o_e}.$$

Proof of Lemma 10

The expected program coverage is given by

$$o_e o_i + \int_0^{(\alpha_0+\theta s_i)Q - o_e o_i} \epsilon f(\epsilon) d\epsilon + ((\alpha_0 + \theta s_i)Q - o_e o_i)\bar{F}((\alpha_0 + \theta s_i)Q - o_e o_i)$$

which can be rewritten as

$$(\alpha_0 + \theta s_i)Q - \int_0^{(\alpha_0+\theta s_i)Q - o_e o_i} F(\epsilon) d\epsilon$$

If we denote the optimal supply–and demand–side investments by s_i^{1*}, s_i^{2*} and o_i^{1*}, o_i^{2*} for the demand scenarios with means μ_1 and μ_2 respectively, then coverage under scenario 2 minus coverage under scenario 1 equals

$$\begin{aligned} &(\theta s_i^{2*} - \theta s_i^{1*})Q + \int_0^{(\alpha_0+\theta s_i^{1*})Q - o_e o_i^{1*}} F_1(\epsilon) d\epsilon - \int_0^{(\alpha_0+\theta s_i^{2*})Q - o_e o_i^{2*}} F_2(\epsilon) d\epsilon \\ \geq &\int_0^{(\alpha_0+\theta s_i^{1*})Q - o_e o_i^{1*}} F_1(\epsilon) d\epsilon - \int_0^{(\alpha_0+\theta s_i^{1*})Q - o_e o_i^{1*}} F_2(\epsilon) d\epsilon \end{aligned}$$

The above inequality holds because in the second expression, we have replaced s_i^{2*} and o_i^{2*} by s_i^{1*} and o_i^{1*} in the coverage term for scenario 2. Now, if $D_{\mu_2} \geq_{st} D_{\mu_1}$, then $F_1 \geq F_2$ always and hence the coverage under scenario 2 would be higher than the coverage under scenario 1. However, when $D_{|\mu_2} \geq_{st} D_{|\mu_1}$ does not hold, there does not appear to be a condition that can guarantee that the coverage under scenario 2 would be higher than the coverage under scenario 1. Hence the result.

Proof of Lemma 11

From the proof of Lemma 10, we have that coverage under scenario 2 minus coverage under scenario 1 is greater than or equal to

$$\int_0^{(\alpha_0+\theta s_i^{1*})Q-o_e o_i^{1*}} F_1(\epsilon)d\epsilon - \int_0^{(\alpha_0+\theta s_i^{1*})Q-o_e o_i^{1*}} F_2(\epsilon)d\epsilon$$

where F_1 and F_2 correspond to the CDFs under demand scenarios 1 and 2 with mean μ and variances σ_1 and σ_2 respectively. Ridder et al. (1998) show that when $D_{|\sigma_2} \geq_{var} D_{|\sigma_1}$,

$$H_n(x) = \int_0^x (F_1(x) - F_2(x))dx \geq 0 \ \forall \ x \geq 0$$

A direct application of this result proves the lemma.

Bibliography

Archibald, T.W., L.C. Thomas, J.M. Betts, R.B. Johnston. 2002. Should start–up companies be cautious? Inventory policies which maximise survival probabilities. *Management Science.* **48**(9) 1161-1174.

Arslan, H., S. C. Graves, T. A. Roemer. 2007. A Single–product inventory model for multiple demand classes. *Management Science.* **53**(9) 1486–1500.

Aviv, Y., A. Federgruen. 1997. Stochastic inventory models with limited production capacity and periodically varying parameters. *Probability in the Engineering and Informational Sciences.* **11**(1) 107-135.

Babich, V., M.J. Sobel. 2004. Pre–IPO operational and financial decisions. *Management Science.* **50**(7) 935-948.

Bachmann, M. O. 2009. Costeffectiveness of community–based treatment of severe acute malnutrition in children. *Expert Review of Pharmacoeconomics & Outcomes Research.* **10**(5) 605—612.

Bassok, Y., R. Anupindi. 2008. Analysis of supply contracts with commitments and flexibility. *Naval Research Logistics.* **55**(5) 459–477.

Basu, A. K., R. Lal, V. Srinivasan, R. Staelin. 1985. Salesforce compensation plans: an agency theoretic perspective. *Marketing Science.* **4**(4) 267–291.

Beamon, B.M., S.A. Kotleba. 2006. Inventory modeling for complex emergencies in humanitarian relief operations. *International Journal of Logistics: Research and Applications.* **9** (1) 1-18.

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

Camdereli, A. Z., Swaminathan, J. M. 2010. Misplaced inventory and radio-frequency identification (RFID) technology: information and coordination. *Production and Operations Management.* **19** 1–18.

Celasun, O., J. Walliser. 2007. Predictability of aid: do fickle donors undermine economic development? *46th Panel Meeting of Economic Policy, Lisbon.*

Birdsall, N., W. D. Savedoff, A. Mahgoub, K. Vyborny. 2011. Cash on Delivery: a new approach to foreign aid. Available online: http://www.cgdev.org/publication/9781933286600-cash-delivery-new-approach-foreign-aid

Ciarallo, F.W., R. Akella, T.E. Morton. 1994. A periodic review, production planning model with uncertain capacity and uncertain demand – optimality of extended myopic policies. *Management Science.* **40**(3) 320-332.

Chao, X., J. Chen, S. Wang. 2008. Dynamic inventory management with cash flow constraints. *Naval Research Logistics.* **55**(8) 758-768.

Chen, L. C., A. K. M. A. Chowdhury, S. L. Huffman. 1980. Anthropometric assessment of energy–protein malnutrition and subsequent risk of mortality among preschool aged children. *The American Journal of Clinical Nutrition* **33** 1836–1845.

Cohen, M. A., N. Agrawal. 1998. An analytical comparison of long and short termcontracts. *IIE Transactions.* **31**(8) 783–796.

Duran, S., M.A. Gutierrez, P. Keskinocak. 2011. Pre-Positioning of Emergency Items for CARE International. *Interfaces.* **41** (3) 223-237.

Deo, S., C. J. Corbett. 2010. Dynamic Allocation of Scarce Resources Under Supply Uncertainty. *Indian School of Business.* **Working Paper**.

Deo, S., S. Iravani, T. Jiang, K. Smilowitz, S. Samuelson. 2012. Improving Access to Community-Based Chronic Care Through Improved Capacity Allocation. *Indian School of Business.* **Working Paper**.

Deo, S., M. Sohoni. 2011. Decentralization of diagnostic networks: access vs. accuracy tradeoff and network externality. *Indian School of Business.* **Working Paper**.

Deshpande, V., M. A. Cohen, K. Donohue. 2003. A threshold inventory rationing policy for service-differentiated demand classes. *Management Science.* **49**(6) 683–703.

Evans, R. V. 1968. Sales and restocking policies in a single inventory system. *Management Science.* **14**(7) 463–473.

Federgruen, A., P. Zipkin. 1986a. An inventory model with limited production capacity and uncertain demands I: the average–cost criterion. *Mathematics of Operations Research.* **11**(2) 193-207.

Federgruen, A., P. Zipkin. 1986b. An inventory model with limited production capacity and uncertain demands II: the discounted–cost criterion. *Mathematics of Operations Research.* **11**(2) 208-215.

Fininnov.org. 2011. The Pledge Guarantee for Health (PGH): a new platform to improve efficiency and impact of donor funding for health commodities. Available online: http://en.fininnov.org/img/pdf/ateliers/The Pledge Guarantee for Health.pdf

Frank, K. C., R. Q. Zhang, I. Duenyas. 2003. Optimal policies for inventory systems with priority demand classes. *Operations Research.* **51**(6) 993–1002.

Garcia, V. 2012. Children malnutrition and horizontal inequalities in Sub-Saharan Africa: a focus on contrasting domestic trajectories. *United Nations Development Programme.* **Working Paper**.

Gaur, V., S. Seshadri. 2005. Hedging inventory risk through market instruments. *Manufacturing and Service Operations Management.* **7**(2) 103-120.

Glasserman, P., S. Tayur. 1995. Sensitivity analysis for base–stock levels in multi–echelon production–inventory systems. *Management Science.* **41**(2) 263-281.

Hart, O., B. Holmstrom. 1987. The theory of contracts. T. F. Bewley, ed. *Advances in Economic Theory Fifth World Congress.* Cambridge University Press, Cambridge, UK. 97–103.

Heese, H. S., J. M. Swaminathan. 2010. Inventory and sales effort management under unobservable lost sales. *European Journal of Operational Research.* **207**(3) 1263–1268.

Heyman, D.P., M.J. Sobel. 1984. Stochastic models in operations research, vol. II: stochastic optimization. McGraw-Hill, NY.

Inter Press Service. 2010. Sierra Leone: unfulfilled promise of free maternal health care for mothers. Available online: http://www.ipsnews.net/2010/10/sierra-leone-unfulfilled-promise-of-free-maternal-health-care-for-mothers/

Joseph, K., A. Thevaranjan. 1998. Monitoring and incentives in sales organizations: an agency-theoretic perspective. *Marketing Science*. **17**(2) 107-123.

Song, J. 1994. The effect of leadtime uncertainty in a simple stochastic inventory model. *Management Science*. **40**(5) 603–613.

Kaplan, A. 1969. Stock rationing. *Management Science*. **15**(5) 260–267.

Kapuscinski, R., S. Tayur. 1998. A capacitated production–inventory model with periodic demand. *Operations Research*. **46**(6) 899-911.

Kleijn, M. J., R. Dekker. 1998. An overview of inventory systems with several demand classes. Econometric Institute Report 9838/A, Erasmus University, Rotterdam, The Netherlands.

Khouja, M., S. S. Robbins. 2003. Linking advertising and quantity decisions in the single-period inventory model. *International Journal of Production Economics*. **86**(2) 93–105.

Lane, C., A. Glassman. 2008. Smooth and predictable aid for health: a role for innovative financing? *Brookings Institution*. **Working paper 1**, Global Health Financing Initiative.

Mathers, C. D., R. Sadana, J. A. Salomon, C. JL. Murray, A. D. Lopez. 2000. Estimates of DALE for 191 countries: methods and results. *World Health Organization*. **Working Paper**.

Nahmias, S., S. Demmy. 1981. Operating characteristics of an inventory system with rationing. *Management Science*. **27**(11) 1236–1245.

Porteus, E. L. 1990. Stochastic inventory theory. D.P. Heyman and M.J. Sobel (Eds.). *Handbooks in Operations Research and Management Science*. **2** 605-652.

Rashkova, I., J. Gallien, P. Yadav. 2011. A data-driven model of drug stockouts in global fund grant recipient countries. *INFORMS Annual Meeting, Charlotte.*

Rekik, Y., Z. Jemai, E. Sahin, Y. Dallery. 2007. Improving the performance of retail stores subject to execution errors: Coordination versus RFID technology. *OR Spectrum*. **29**(4) 597–626.

Rekik, Y., E. Sahin, Y. Dallery. 2008. Analysis of the impact of the RFID technology on reducing product misplacement errors at retail stores. *International Journal of Production Economics*. **112** 264–278.

Ridder, A., E.v.d. Laan, M. Salomon. 1998. How larger demand variability may lead to lower costs in the newsvendor problem. *Operations Research*. **46**(6) 934–936.

Shaked, M., G. Shanthikumar. 2007. Stochastic orders. Springer, NY.

Stokes, T. 2011. US budget quagmire leaves global health funding in the lurch. *Nature Medicine*. **17**(9) 1028.

Swaminathan, J. 2009. UNICEF Plumpy'Nut Supply Chain. *UNC Kenan–Flagler Business School*. **Case Study**.

Swaminathan, J. 2010. Case study: Getting food to disaster victims. *Financial Times.*

Taylor, T., P. Yadav. 2011. Subsidizing the distribution channel: donor funding to improve the availability of products with positive externalities. **Working Paper**.

Topkis, D. 1968. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. *Management Science.* **15**(3) 160–176.

Tsay, A. A., W. S. Lovejoy. 1999. Quantity exibility contracts and supply chain performance. *Manufacturing & Service Operations Management.* **1**(2) 89–111.

UNICEF. 2009. A supply chain analysis of ready–to–use therapeutic foods for the horn of Africa: the nutrition articulation project. Available online: http://ghta-nutrition.org/

USAID. 2010. Performance–based incentives primer for USAID missions. Available online: http://pdf.usaid.gov/pdf_docs/PNADX747.pdf

Veinott, A. F. 1965. Optimal policy in a dynamic, single product, nonstationary inventory model with several demand classes. *Operations Research.* **13**(5) 761–778.

Wakolbinger, T., F. Toyasaki. 2011. Impact of funding systems on humanitarian operations. M. Christopher and P. Tatham (Eds.). *Meeting the challenge of preparing for and responding to disasters.* Kogan Page.

Wang, Y., Y. Gerchak. 1996. Periodic review production models with variable capacity, random yield, and uncertain demand. *Management Science.* **42**(1) 130–137.

Wei, Y., Y. Chen. 2011. Joint determination of inventory replenishment and sales effort with uncertain market responses. *International Journal of Production Economics.* **134**(2) 368–374.

World Health Organization. 2004. Global burden of disease 2004 update: disability weight for diseases and conditions. *World Health Organization.*

World Health Organization. 2007. Aid effectiveness and health. WHO/HSS/healthsystems/2007.2 working paper No.9.

Xu, Y., J.R. Birge. 2004. Joint production and financing decisions: modeling and analysis. Working paper. Northwestern Univeristy, Evanston, IL.

Xue, W., X. Xu, R. Wang. 2013. Combined Sales Effort and Inventory Control under Demand Uncertainty. it Discrete Dynamics in Nature and Society. **2013** 1–8.

Yang, Y., J. V. d. Broeck, L. M. Wein. 2013. Ready-to-use food-allocation policy to reduce the effects of childhood undernutrition in developing countries. *PNAS.* **110**(12) 4545–4550.