Fleet Management Coordination in Decentralized Humanitarian Operations

We study incentive alignment for the coordination of operations in humanitarian settings. Our research focuses on transportation, the second largest overhead cost to humanitarian organizations after personnel. Motivated by field research, we study the fleet size problem from a managerial perspective. In terms of transportation, an equity focused Humanitarian Program implemented by an International Humanitarian Organization has private information which affects the balance between equity and efficiency intended by the Organization's Headquarters. The incentive alignment issue is complex because traditional instruments based on financial rewards and penalties are not considered to be viable options. This problem is further complicated by information asymmetry in the system due to the dispersed geographical location of the parties. We design a novel mechanism based on an operational lever to coordinate incentives in this setting. This study contributes to two streams of literature, humanitarian logistics, and incentives in operations management.

Key words: Incentives, Humanitarian Logistics, Fleet Management

1. Introduction

The need for humanitarian action has increased dramatically in the last few decades and it is expected to rise significantly in the years to come (Thomas and Kopczak 2005). International Organizations carrying out humanitarian action face serious challenges to deliver the right goods and services to the right people at the right time and at the right cost (Van Wassenhove 2006). Including basic health care provision, nutrition and agriculture, relief and development programs are the primary channels of humanitarian aid delivery carried out by International Humanitarian Organizations (IHO). Every year IHO spend more than $1,5 billions running an international fleet close to 80,000 four wheel drive (4x4) vehicles to support the delivery of humanitarian programs. Motivated by extensive field research, we study coordination issues in a two-party decentralized 4x4 fleet management system supporting IHO's programs.

Our research is motivated by a larger field project to understand field vehicle fleet management in IHO. The field project includes three large IHO using global procurement managed at the Headquarters level: the International Committee of the Red Cross (ICRC), the International Federation of Red Cross and Red Crescent Societies (IFRC), and the World Food Programme (WFP). Staff interviews with various IHO were conducted at the Headquarters in Europe and the Middle East, as well as at national and field program levels in the Middle East and Africa. In the following we will briefly describe the research problem and use our findings from the field research to motivate the theoretical model that is studied in this paper.

The fleet management system has two decision-making parties: the Program and the Headquarters. Often located in remote areas of developing countries (the field), Programs are service oriented. They provide assistance and help alleviating the suffering of people in the aftermath of disasters (relief). Also, Programs implement activities to improve the quality of life of poor communities (development). Transportation requirements for relief and development are different. Relief Programs assign vehicles according to emergency priorities for disaster assessment, or to coordinate search, rescue and emergency aid distribution operations. Development Programs use vehicles for regular visits to villages or refugee camps for health care or to coordinate aid distribution. Urgency in development Programs is lower and vehicles are typically assigned to visits in order of requisition. We focus on fleet management in development Programs with big fleets of twenty or more vehicles in the same geographic location. Development Programs are henceforth referred to as Programs.

Some examples of Programs include health, nutrition, water and sanitation. In terms of transportation, the objective of Programs is to have a vehicle available when it is required by their staff to visit beneficiaries. Although speed in demand fulfillment is not necessarily critical, Programs must meet demand in a reasonable time. Due to the long-term nature of Programs, visits that cannot be performed on time are accumulated. The Program incurs two main costs, the cost of delay and the cost of managing their fleets in the field. Often Program managers are more sensitive

to the cost of delay than to the fleet management cost. The bigger the fleet the lower the cost of delay. During the planning stage, the Program states its transportation needs to the Headquarters.

The Headquarters have the function of procuring the fleet requested by the Program. Typically located in the US or Europe, the Headquarters' objective is balancing the cost of delay of last mile distribution and the operating cost of the fleet, i.e. the fleet management cost plus the running cost. During the planning stage, the Headquarters decide the optimal fleet size to minimize the system's cost. The fleet size includes a fleet buffer to guarantee that a reasonable proportion of visits will not suffer any delay. The fleet buffer determines the service level of the fleet. The bigger the fleet buffer the lower the cost of delay but the higher the operating cost of the fleet.

In summary, the Program states its transportation needs to the Headquarters and manages the fleet in the field. The Headquarters assign operational capacity (fleet) to the Program. Program‘s transportation needs are stochastic and private information. Both the Headquarters and the Program incur a disutility due to the delay in reaching IHO's beneficiaries. But only the Headquarters are accountable for the full operating cost of the fleet. Since the operating cost increases with the fleet size, the Headquarters optimal fleet size is smaller than the calculated by the Program. Hence, the Program may have incentives to distort its needs to get a larger fleet. Due to particular characteristics of humanitarian operations related to earmarked funding, transferring the accountability of the full operating cost to the Program is not feasible. Additionally, standard financial-incentives-based mechanisms are not considered viable in this humanitarian setting. This is primarily due to the organizational culture of the IHO, wherein the employees are driven by their motivation to serve and not by the objectives such as profit maximization (Lindberg 2001, Manell 2010). E.g., it is almost inconceivable for the IHO to incentivize volunteer medical doctors working in the field by using financial penalties and rewards.

The Headquarters monitor the Program's stated needs but unfortunately it does not deter programs from needs distortion. The incentives problem is illustrated by quotes from our interviews. One of the Headquarters staff said:

"*I feel some of our programs have more vehicles than required*"

In fact, one of the Senior Fleet Managers we interviewed in Geneva, Switzerland estimates that their Programs have between 10% and 15% more vehicles than required. In contrast, when we asked about the fulfillment of his transportation needs, a development Program manager in the Zambezia Province, Northern Mozambique said:

"*Often we have to wait too long to have a vehicle available to go to the field*"

These quotes refer to the steady state behavior of the system instead of referring to the punctual – and expected – mismatches between stochastic demand and stochastic supply. In fact, to respond to a stochastic demand, the fleet management system uses the fleet buffer, which is decided by calculating a buffer factor that should balance the service level and the operating cost of the fleet. Combining mechanism design and heavy traffic queueing theory we develop a non-trivial but tractable model to coordinate incentives in this setting. We use the buffer factor to induce truth revelation from the Program. Our model provides insights to overcome current inefficiencies in decentralized humanitarian systems. In some instances our model is able to reduce the fleet excess in more than 10%, matching the intuition of senior fleet managers. We also obtain counterintuitive results regarding the value of private information in a decentralized humanitarian system.

We believe that the unique characteristics of IHO make the incentive misalignment issues an extremely interesting research topic, and not just a mere application of the principal-agent framework from the economics literature. Our work should appeal to the Operations Management community as it showcases the strategic importance of operational design beyond the objective to achieve tactical efficiency in the humanitarian sector. The model can be used to advance the academic understanding of decentralized decision-making in humanitarian operations. To the best of our knowledge, this is the first analytical study of decentralized humanitarian operations completely informed by field research.

2. Literature Review

This paper contributes to the humanitarian logistics literature and to the incentives literature in operations management. There is an increasing interest in studying humanitarian operations

(Altay and Green 2006). Extant literature on humanitarian logistics follows a classical optimization approach. Most of the research examines stochastic relief systems for disaster preparedness or for disaster response. Typically, the literature applies operations research techniques to relief settings assuming central planner coordination. The objective can be equity or cost-efficiency oriented.

Equity-based objective functions have been studied in terms of time of response and demand fulfilment. Research to minimize the time of response can be found in Chiu and Zheng (2007) and Campbell et al (2008). Research exploring demand coverage include Batta and Mannur (1990), Ozdamar et al (2004), Jia et al (2007), De Angelis et al (2007), Yi and Ozdamar (2007), Saadatseresht et al (2009), and Salmeron and Apte (2010).

Cost-based objective functions are often represented either via monetary cost or via travel distance. Cost minimization can be found in the work of Barbarosoglu et al (2002), Barbarosoglu and Arda (2004), Beamon and Kotleba (2006) and Sheu (2007). Distance traveled minimization has been explored by Cova and Johnson (2003), Chang et al (2007), and Stepanov and Smith (2009). In their work, Stepanov and Smith also examine an equity based function of time of response. Regnier (2008) models the trade-off between cost and equity in hurricane evacuation operations, also from a central planner perspective. In contrast to extant literature in humanitarian logistics we analyze incentives in a decentralized system.

The incentives literature on adverse selection discusses settings where there are agents of different type. Agents know their types while the principal does not (Green and Laffont 1977, Dasgupta et al 1979, Myerson, 1979, Harris and Townsend 1981, Maskin and Riley 1984). When offered a menu of contracts of particular characteristics, these agents reveal their type following the revelation principle. The supply chain management literature on incentives has focused on exploring manufacturing and service operations management in "for profit" settings. Typically, decisions in manufacturing supply chains relate to order-quantity of goods while decisions in service supply chains relate to the capacity of the service system. Most of the mechanisms for supply chain coordination in manufacturing and in service operations management are based on financial incentives.

In this paragraph we briefly summarize some commonly studied financial contracts in manufacturing and service supply chains. This is not a comprehensive list of the vast literature on supply chain contracts, but it provides the readers a primer on the nature of contracts that have been studied in such settings. In revenue sharing contracts a retailer pays a supplier a wholesale price for each unit purchased plus a percentage of the revenue generated by the retailer (Cachon and Lariviere 2005). Buy-back contracts have a wholesale price and a buy-back price for unsold goods (Pasternack 1985). In Sales-rebates contracts the supplier charges the retailer a per-unit wholesale price but gives the retailer a rebate per unit of goods sold above a predefined threshold (Krishnan et al 2001, Taylor 2002). In quantity discount contracts the retailer receives a discount either on all units if the purchased quantity exceeds a threshold (all-unit quantity discount) or on every additional unit above a threshold (incremental quantity discount) (Corbett and de Groote 2000, Cachon and Terwiesch 2009). In price-discount contracts wholesale prices are discounted on the basis of annual sales (Bernstein and Federgruen 2003). Our research departs from this stream of literature since we use a non-financial, capacity-based mechanism for system coordination.

In service systems decisions are based on capacity. For instance, Hasija et al (2008) explore pay-per-call and pay-per-time contracts in call-centers. In the first type of contracts the vendor earns a fixed fee from the client for each served phone call. In the second type of contracts the vendor is paid per unit of time spent serving customers. In this service setting as in the previous manufacturing ones coordination is achieved via financial transfer payments. Departing from financial mechanisms, Su and Zenios (2006) explore the efficiency-equity trade-off in kidney transplantation. In their setting financial transfers are not possible. They propose a kidney's allocation rule based on the fact that lower-risk patients are willing to spend more time waiting in order to receive organs of higher quality. Instead of an allocation rule we propose a capacity (fleet buffer) rule for fleet coordination. Our contribution is twofold. First, we develop the first analytical model of decentralized humanitarian operations completely informed by field research. Second, using a combination of heavy traffic models with mechanism design we develop a mathematically tractable

model using operational levers that respect exogenous constraints that exist in the humanitarian context.

3. The Fleet Management System

The system is composed by two parties: the Headquarters and the Program. Based on system cost considerations, the Headquarters decide a service level for the fleet and procure the vehicles from a global source. The Headquarters are also in charge of monitoring the Program's transportation needs whenever it is required.

The Headquarters are accountable for the total cost of the system. The total cost includes the running cost of the fleet, the fleet management cost in the field, and the cost of delay. The running cost of the fleet, r, is the average running cost per vehicle per unit of time. It includes maintenance, repairs, and fuel costs (Pedraza-Martinez and Van Wassenhove 2010). The running cost is an overhead cost to the IHO. The fleet management cost, c, is the average management cost per vehicle per unit of time. Pedraza-Martinez et al (2011) find that the Program is accountable for c. During our field visits we observed that senior humanitarian staff in the field often dedicate a proportion of their time to fleet scheduling and routing. The fleet management cost also includes the cost of vehicle drivers, which are instrumental to Program delivery due to their knowledge of local language and geography. Finally, the fleet management cost includes the cost of information systems and spreadsheets to track fleet scheduling and routing in the field. The field management cost in the field is a direct cost to the Program. We refer to $c+r$ as the operating cost of the fleet.

The cost of delay, w, is measured per field visit per unit of time. The Program incurs the cost of delay and it is accountable for it. Although the cost of delay is not a cash cost to the Program, it results in a disutility associated with delay in carrying out Program delivery. To simplify our analysis and capture both the disutility of delay and the tensions of this decentralized system we assume a constant marginal cost.

The Program reports its transportation needs to the Headquarters and uses the fleet to coordinate and execute last mile distribution. The Program uses the fleet for transportation of staff to

visit beneficiaries, transport of materials, and transport of items for distribution to beneficiaries (Pedraza-Martinez et al 2011). Although demand is more stable through time than in relief settings, humanitarian development work has some stochasticity – in both arrivals and service times. This comes from the mobility of beneficiaries and the unpredictability of operating conditions in terms of weather, road conditions and security. In case of unavailability of vehicles to carry out the visits to beneficiaries the demand tends to accumulate but it rarely disappears.

To capture the stochasticity of the system and the nature of work accumulation when the Program faces a fleet shortage, we use a queueing model. The use of queueing models for vehicle fleet management systems is well established in the literature. Queueing models have been used for analyzing police patrol systems (Green 1984, Green and Kolesar 1984a, 1984b, 1989, 2004), fire departments (Kolesar and Blum 1973, Ignall et al 1982), helicopter fire fleets (Bookbinder and Martell 1979), and ambulance fleets (Singer and Donoso 2008). These papers model transportation needs using stochastic inter-arrival times.

We represent the Program's transportation needs with the Greek letter λ. The transportation needs are measured in rate of visits per unit of time. Service times, μ, are also stochastic. For simplicity, we assume $\mu = 1$, which corresponds to measuring time in the scale of mean service times (Whitt 1992).

To balance the system costs, i.e. the operating cost and the cost of delay, the Headquarters decide on the fleet size. In a deterministic setting the minimum fleet size for the system would be λ/μ. However, a fleet buffer is required to maintain stability of the system due to inherent variability in inter-arrival and service times. In other words, the fleet buffer is the extra-number of vehicles needed for protecting the system from stochasticity and achieving a predetermined service level.

We model the system as an Erlang-C system following a first-come first-served queueing discipline. This rule was chosen since, as mentioned before, within the Program all the visits have the same priority. For analytical tractability we use heavy-traffic approximations under the "rationalized regime" to represent the average delay in the system. The Halfin-Whitt (1981) delay function,

$\pi(y)$ is an asymptotically exact approximation to the probability of delay, $\Pr\{\text{wait} > 0\}$. The value of $\pi(y)$ is:

$$\pi(y) = \left[1 + \frac{y\Phi(y)}{\phi(y)}\right]^{-1}$$

$\Phi(y)$ and $\phi(y)$ are the unit normal cdf and pdf, respectively. The service level of the fleet is the average proportion of visits carried out without delay. The Program that is assumed in our system neither carries out emergency response activities nor highly scheduled regular work, but carries out delay sensitive and stochastically arriving developmental activities. Therefore we believe that the rationalized regime is appropriate for this setting.

Borst et al (2004) show that a square-root staffing rule is asymptotically optimal for a system operating in the rationalized regime. Following Halfin and Whitt (1981), Grassman (1988), Whitt (1992), Borst et al (2004), Hasija et al (2005) we use the square-root staffing rule

$$F(\gamma) = \lambda + \gamma(c, r, w)\sqrt{\lambda} \tag{1}$$

to calculate the fleet size. Denoted by γ, the buffer factor is the key decision variable to determine the fleet buffer. The optimal buffer factor can be obtained via optimization methods by balancing the importance of delays in visits with fleet operating costs.

The effectiveness of the square-root rule increases in the size of the fleet and it has been shown that it is a robust approximation for the optimal system size of fleets of 20 or more vehicles. According to Borst et al (2004) and Hasija et al (2005), for a given buffer factor γ the average number of visits in the queue is $Q(\gamma) = \frac{\pi(\gamma)\lambda}{F\mu - \lambda}$. Using $\mu = 1$ and replacing F by its definition in (1) we obtain:

$$Q(\gamma) = \frac{\pi(\gamma)\sqrt{\lambda}}{\gamma} \tag{2}$$

3.1. Centralized Benchmark

In a centralized system the central planner minimizes the total system cost given by the average cost of delay $wQ(\gamma)$ plus the average operating cost of the fleet $(c + r)F(\gamma)$. The central planner's problem is:

$$\min_{0<\gamma} C_{Cent}(\gamma) = wQ(\gamma) + (c+r)F(\gamma) \tag{3}$$

Replacing (1) and (2) in (3) we can rewrite the central planner's problem as:

$$\min_{0<\gamma} C_{Cent}(\gamma) = w\frac{\pi(\gamma)\sqrt{\lambda}}{\gamma} + (c+r)(\lambda+\gamma\sqrt{\lambda}) \tag{4}$$

The cost function (4) is unimodal (Borst et al 2004) and it has a finite and positive minimum value, $\gamma^*(c,r,w)$, which is independent of λ and it only depends on the cost parameters of the system (Hasija et al 2005);

$$\gamma^*(c,r,w) = \arg\min C_{Cent}(\gamma) \tag{5}$$

is the optimal buffer factor for the centralized benchmark.

The key assumption here is that the central planner knows the Programs transportation needs. As we will show next, decentralization and private information issues render the centralized benchmark impossible to implement in the current system.

3.2. Current System Without Monitoring

The Program must balance the cost of delay with the cost of managing the fleet in the field. If the Program reported its true needs the average system cost would be as equation (3). However, the Program is not accountable for the running cost of the fleet. Hence, the relative weight of delay for the Program is greater than the one for the Headquarters. To find its optimal buffer factor the Program solves:

$$\min_{0<\gamma} C_{Prog}(\gamma) = wQ(\gamma) + cF(\gamma) \tag{6}$$

The optimal buffer factor for the Program is:

$$\bar{\gamma}(c,w) = \arg\min C_{Prog}(\gamma) \tag{7}$$

As in equation (5) the optimal buffer factor of the Program only depends on its cost parameters, c, and w. This allows us to state a lemma that clearly explains the misalignment of incentives between the Headquarters and the Program. All the proofs are included in the appendix.

LEMMA 1. $\gamma^*(c,r,w) < \bar{\gamma}(c,w)$.

Equation (1) together with lemma 1 imply that the optimal fleet size for the Headquarters is smaller than the optimal fleet size for the Program. In other words, both the Headquarters and the Program consider the full cost of delay, but the Program considers only a portion of the operating cost, which the Headquarters fully internalize. Since the operating cost increases with the fleet size, then the Headquarters optimal fleet size will be smaller than the one calculated by the Program. This result is consistent with our field observations, in particular with the worries of the Headquarters in terms of oversized fleets. The result is also consistent with the worries of the Program in terms of not having enough vehicles to optimize their service.

The true transportation needs are private information to the Program. The misaligned incentives and private information create an adverse selection issue, and the Program may distort its stated transportation needs. We abstract the Program to two types: low transportation needs, L, and high transportation needs, H, and $\lambda_L < \lambda_H$. This standard assumption helps us to capture the main trade-offs of equity and efficiency while keeping the model analytically tractable. In reality, we observed that λ_H is easy to estimate since the maximum number of vehicles is determined by the total number of humanitarian staff in the Program. For instance, in a Program with 25 medical doctors, the IHO would not procure more than 25 vehicles. Nevertheless, possible coordination in fleet scheduling could results in a decrease in transportation needs, leading to λ_L. By stating its transportation needs, $\hat{\lambda}_i$, the $i \in \{L,H\}$ type program would target an intended buffer factor $\delta(\hat{\lambda}_i)$ such that

$$\delta(\hat{\lambda}_i) = \frac{\hat{\lambda}_i + \gamma\sqrt{\hat{\lambda}_i} - \lambda_i}{\sqrt{\lambda_i}} \tag{8}$$

From (8) follows that $\delta(\hat{\lambda}_i)\,|_{\hat{\lambda}_i=\lambda_i} = \gamma$. Let

$$\delta_L = \delta(\hat{\lambda}_L)\,|_{\hat{\lambda}_L=\lambda_H}, \quad \text{and} \quad \delta_H = \delta(\hat{\lambda}_H)\,|_{\hat{\lambda}_H=\lambda_L} \tag{9}$$

Note that $\delta_L > \gamma$ and $\delta_H < \gamma$ follow from the fact that $\lambda_L < \lambda_H$. For a given i type Program the current system's cost would be:

$$C_{Curr}(\delta(\hat{\lambda}_i)) = wQ_i(\delta(\hat{\lambda}_i)) + (c+r)F_i(\delta(\hat{\lambda}_i)) \tag{10}$$

Also note that for $\gamma = \gamma^*$ in (8) we have $C_{Cent}(\gamma^*) < C_{Curr}(\delta_i)$.

A straightforward way to coordinate the system would be by transferring the accountability of running costs of the fleet to the Program. The running costs are an overhead and the Headquarters, which are considered a support function, are accountable for these costs. Due to accountability reasons particular to the humanitarian sector the internal transfer of running cost's accountability is not possible. Aware of the potential increase in cost coming from transportation needs distortion but unable to transfer the entire cost to the Program, the Headquarters react to transportation needs distortion by monitoring the Program's stated needs.

3.3. Current System With Monitoring

The Headquarters don't know the Program's true transportation needs, but they have some idea –just not a very accurate one. So we assume that the Headquarters have a probabilistic prior belief that:

$$\lambda_i = \begin{cases} \lambda_L, & \text{w.p. } q \\ \lambda_H, & \text{w.p. } 1-q \end{cases}$$

The Headquarters can monitor $\hat{\lambda}_i$ before procuring the fleet. To monitor, the Headquarters have to carefully check the Program's data records. During our field visits we observed that Programs often have detailed data on fleet use and transportation needs in the field. Typically the data is in printed form, ready for auditing purposes but it is not stored in a way that can be easily accessed by the Headquarters. Due to the lack of trustworthy information systems on transportation needs at the national level, the Headquarters may send staff to the field to monitor the Program's estimated workload *in situ*. The Headquarters exert a monitoring effort $p \in [0,1]$ corresponding to the proportion of Programs to monitor. The monitoring cost is $m(p)$, a function of the monitoring

effort. By sending staff to the field, the Headquarters can get an accurate estimation of the Program's transportation needs. We assume that by monitoring the Headquarters can know the true transportation needs of the Program. Hence the Program's problem becomes:

$$\min_{\hat{\lambda}_{i\in\{L,H\}}} \quad E_p[C_{Prog}(\delta(\hat{\lambda}_i))] = p\left(wQ_i(\gamma) + cF_i(\gamma)\right) + (1-p)\left(wQ_i(\delta(\hat{\lambda}_i)) + cF_i(\delta(\hat{\lambda}_i))\right) \tag{11}$$

The first thing to note is that in the current system the high type Program always reveals its true transportation needs. This is formalized in the following proposition.

PROPOSITION 1. *In the current system* $C_{Prog}(\gamma^*) < E_p[C_{Prog}(\delta_H)]$

The intuition behind Proposition 1 comes from the fact that the intended buffer factor from distortion of transportation needs for the high type Program is lower than the buffer factor offered by the Headquarters. Additionally, the buffer factor offered by the Headquarters is lower than the optimal fleet buffer for the Program (Lemma 1). Since the cost function for the Program is unimodal with minimum in $\bar{\gamma}$ (equation 7), it is always cheaper for the high type Program to reveal the truth. Otherwise, the extra-cost of delay would overcome the savings on fleet management.

We are left with the low type Program. The low type Program states its true transportation needs as long as $C_{Prog}(\gamma^*) < C_{Prog}(\delta_L)$. This suggests the existence of a threshold for truth telling, formalized in the following proposition.

PROPOSITION 2. *For any given values of* λ_L *and* λ_H *there exists a truth telling threshold* $\hat{\gamma}_L$ *such that* $\hat{\gamma}_L \neq \gamma^*$ *and:*

$$\textit{if } \gamma < \hat{\gamma}, \textit{then } C_{Prog}(\gamma) > E_p[C_{Prog}(\delta_L)],$$

$$\textit{if } \gamma > \hat{\gamma}, \textit{then } C_{Prog}(\gamma) < E_p[C_{Prog}(\delta_L)]$$

The threshold value satisfies: $E_p[C_{Prog}(\delta_L)] = C_{Prog}(\gamma^*)$

Since δ_L also depends on γ^*, λ_L and λ_H, for a fixed value of γ^* the truth telling threshold depends on the ratio between λ_H and λ_L. The threshold indicates that if the ratio λ_H/λ_L is big enough, the

low type Program's savings in cost of delay are overcome by the extra-cost of fleet management. In summary, the Headquarters know that the high type Program will not distort its needs and the low type Program will only distort its needs for values of $\gamma < \hat{\gamma}_L$. Hence, the Headquarters monitoring effort should only focus on those Programs reporting high type.

Note, however, that there is no penalty for distortion, i.e. the Program's stated transportation needs are independent of the Headquarters monitoring effort. This result is formalized in the following proposition.

PROPOSITION 3. *The Headquarters monitoring effort does not have an influence on whether the Program truly reveal its type*

The lack of effectiveness of monitoring comes from the fact that there is no cost for the Program from being caught distorting its needs. Given the humanitarian nature of Program delivery it is not feasible for the Headquarters to impose a penalty to the Program. The Headquarters choose a monitoring effort solving the following problem:

$$\begin{aligned}\min_{p} \quad & C_{Mon}(\gamma, \delta_i(\hat{\lambda}_i)) = p[q(wQ_L(\gamma) + (c+r)F_L(\gamma)) + (1-q)(wQ_H(\gamma) + (c+r)F_H(\gamma))] \\ & + (1-p)[q(wQ_L(\delta(\hat{\lambda}_L)) + (c+r)F_L(\delta(\hat{\lambda}_L)) + (1-q)(wQ_H(\delta(\hat{\lambda}_H)) + (c+r)F_H(\delta(\hat{\lambda}_H))] + m(p)\end{aligned} \tag{12}$$

The results of propositions 1 and 2 allow us to solve the monitoring problem using backward induction:

$$P = \begin{cases} \min\{m'^{-1}(q[(c+r)[F_L(\delta_L) - F_L(\gamma)] - w[Q_L(\gamma) - Q_L(\delta_L)]]), 1\}, & \text{if } \hat{\lambda}_L = \lambda_H \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

In the next section we will explore a different, capacity-based mechanism to induce truth revelation from the Program.

4. Operational Mechanism

Since financial transfer payments are not implementable in this setting, truth revelation proves to be challenging. In this section we propose a novel mechanism design for truth revelation based on an operational lever, the buffer factor offered by the Headquarters.

The Headquarters (Principal) allocate the Program type i (Agent) an outcome F, i.e. a fleet size, as a function of the reported type. This is operationalized by the Headquarters committing to an offered buffer factor, γ_i, for the stated transportation need of the Program equal to $\hat{\lambda}_i$. The Program reports a type profile $\hat{\lambda}_i$ and the mechanism is executed. The Headquarters' objective is to minimize the system cost for a given distribution of Program types. The incentive compatibility (IC) constraints consist of offering each Program type $i = \{L, H\}$ a buffer factor γ_i to guarantee that $C_{Prog}(\gamma_i) \leq C_{Prog}(\delta_i)$.

During our field visits we observed that Programs do not have an outside option since the Headquarters procure the fleet. Hence, the individual rationality (IR) constraints are defined by the fact that the Headquarters must offer each Program type a buffer factor big enough to protect the system from stochasticity, i.e. $0 < \gamma_i$. The mechanism is formulated as:

$$\begin{aligned} \min_{0<\gamma_L, 0<\gamma_H} E[C_{Mec}(\gamma_L, \gamma_H)] &= q\left[wQ(\gamma_L) + (c+r)F(\gamma_L)\right] + (1-q)\left[wQ(\gamma_H) + (c+r)F(\gamma_H)\right] \\ S.T. & \qquad (14) \\ (IC_L): \quad & wQ(\gamma_L) + cF(\gamma_L) \leq wQ(\delta_L) + cF(\delta_L) \\ (IC_H): \quad & wQ(\gamma_H) + cF(\gamma_H) \leq wQ(\delta_H) + cF\delta_H) \end{aligned}$$

The left hand side of IC constraints in expression (14) is the i type Program cost with a fleet buffer γ_i. The right hand side of IC constraints in (14) represents the cost of the i type Program when distorting its needs. The intended fleet buffer δ_i is defined by (9).

We now show that there exist fleet buffers γ_L and γ_H such that both Program types have incentives to reveal their true transportation needs. We consider the case of both IC constraints loose, one IC constraint active, and both IC constraint active.

Let

$$\tilde{\gamma}_L = \{\gamma \neq \gamma_L : C_{Prog}(\gamma) = C_{Prog}(\gamma_L)\} \qquad (15)$$

There exist two ways of making the low type Program's IC constraint tight. The first way is by forcing $\gamma_L = \delta_L$. The second way of making the low type Program's IC constraint tight is by choosing γ_H such that $\delta_L = \tilde{\gamma}_L$.

PROPOSITION 4. *There exist two regions for induced truth revelation via the operational mechanism as follows:*

1) R1: Equal fleet size region. $\gamma_H < \gamma^* < \gamma_L$ *such that* $F(\gamma_L) = F(\gamma_H)$

2) R2: Different fleet size region. $\gamma^* < \gamma_L$, $\gamma^* < \gamma_H$ *such that* $F(\gamma_L) < (\gamma_H)$ *and the regions for induced truth telling are separated by the threshold* T_1.

The two regions characterized by Proposition 4 can be depicted in the space defined by the transportation needs ratio and probability types (figure 1). In *R1* of Proposition 4 both *IC* constraints in mechanism (14) are binding. The Headquarters offer $\delta_H = \gamma_H < \gamma^* < \gamma_L = \gamma_L$ such that both Program types receive the same fleet size, $F(\gamma_L) = F(\gamma_H))$. This region exists for values of λ_H/λ_L close to 1 and limited by T_1, a threshold implicitly defined by the cost parameters of the system, c, r and w, the lambda ratio, λ_H/λ_L and the low type probability, q. When parameters fall in *R1*, truth revelation is achieved by making the low type Program indifferent between reporting its true needs and distorting these needs, i.e. by making the low type *IC* constraint tight via $\gamma_L = \delta_L$. The extra-cost for low type's truth revelation is mitigated by reducing δ_L via the decrease of γ_H, such that $\gamma_H = \delta_H$. In our numerical experiments we show that the cost mitigation is mediated by q, the low type probability. Increasing q after the threshold T_1 reduces the width of *R1*. Because $\gamma_H = \delta_H$, in *R1* the high type Program is indifferent between revealing its true transportation needs and distorting.

In *R2* of Proposition 4 only the *IC* constraint for the low type Program in (14) is binding. Both γ_L and γ_H are greater than γ^*, which means that the Headquarters give incentives to both Program types. For low values of q the Headquarters keep γ_H closer to γ^* by making $\gamma^* < \gamma_H < \gamma_L$, due to the likelihood of having a larger proportion of high type Programs. By splitting the buffer factor, the Headquarters avoid a big increase in the high type Program cost while making the low type Program *IC* constraint tight, via $\delta_L = \tilde{\gamma}_L$. For large values of q the increase in the likelihood of having low type Programs switches the order of incentives to $\gamma^* < \gamma_L < \gamma_H$. The closer γ_L is to γ^* the lower the cost for the system. As mentioned earlier, we find that as q increases, the region

R2 becomes more favorable than the region $R1$. The intuition is as follows. In *R2*, $F(\gamma_L) < F(\gamma_H)$ while in *R1*, $F(\gamma_L) = F(\gamma_H)$. As q increases, the probability of a Program to be of the low type increases. Hence $F(\gamma_L) < F(\gamma_H)$ becomes a more cost effective mechanism than $F(\gamma_L) = F(\gamma_H)$. In *R2* also characterized in Proposition 4 the high type Program is better of by revealing its true transportation needs since condition $\delta_H < \gamma_H < \bar{\gamma}$ implies $C_{Prog}(\gamma_H) < C_{Prog}(\delta_H)$.

With both constraints loose, neither of the Program types has incentives to distort their transportation needs, i.e. there exists a natural truth telling region. Natural truth revelation follows directly from Proposition 2. The result is formalized in the following corollary.

COROLLARY 1. *There exists a "natural truth telling" region for the operational mechanism.* R3*, the natural truth telling region is defined by* $\gamma^* = \gamma_L = \gamma_H$ *such that* $F_L < F_H$.

Truth revelation In *R3*, characterized by corollary 1, is achieved without the need of incentives. With parameter values falling in this region the low type Program does not have incentives for distorting its needs. If the low type Program distorts its transportation needs, then it gets a fleet big enough to guarantee that the cost of management will overcome the savings in delay. On the other hand, the high type Program does not distort its needs because it would receive a fleet too small for its needs. In this case the cost of delay would overcome the savings in fleet management. It is interesting that T_2 is independent of q. It depends on the cost parameters c, r, w, and the λ_H/λ_L ratio.

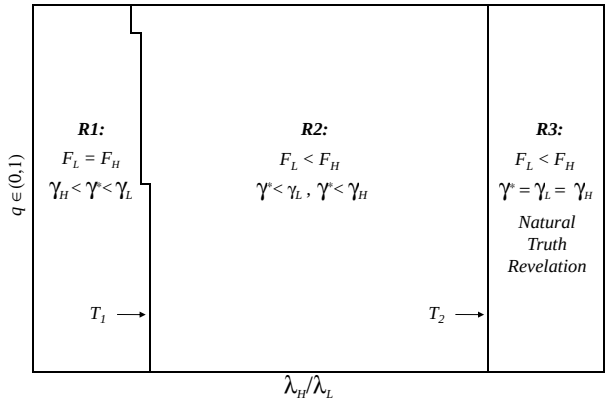


Figure 1 **Truth telling regions in the transportation needs ratio and probability type** $(\gamma_H/\gamma_L, q)$ **space**

To summarize, by being flexible in choosing fleet buffers γ_L and γ_H instead of a unique γ^*, the Headquarters can create operational incentives for truth revelation. There are two ways of achieving truth revelation via the operational mechanism: induced and natural. In the case of induced truth revelation we find that the IC constraint of the low type Program will always be binding. These incentives potentially increase the system's efficiency without the need of monitoring the Program's reported needs. The operational mechanism has the counterintuitive characteristics that the greater the different between types the lower the value of private information for the low type program. The lost of value of information for the low type Program as well as the cost savings for the system in the different models are explored further in the next section.

5. Numerical Study

This section presents a numerical study that complements the analytical insights presented in the previous section. The base case uses weekly planning for a time horizon of 52 weeks. The running cost and the fleet management cost of the fleet are calculated following the research by Pedraza-Martinez and Van Wassenhove (2010) on vehicle replacement in a humanitarian setting. The running cost per vehicle is established at \$17,000 per year. This is equivalent to \$269,23 per week. It includes maintenance, repairs and fuel. The fleet management cost is set to be 15% of the running cost. It includes the time of staff coordinating fleet management, and the salary of vehicle drivers, which depend on the Program (Pedraza-Martinez et al 2011). The normalized demand rate for the low type Program is $\lambda_L = 60$ visits per week. This is equivalent to a fleet of 60 vehicles with a utilization of 100%. We assume that the monitoring cost equals the fleet management cost. In the numerical examples we use a convex monitoring cost function $m(p) = m[qF(\gamma, \lambda_L) + (1-q)F(\gamma, \lambda_H)]p^2$. This signifies that the monitoring cost is proportional to the expected fleet size. We compare the cost of: 1) centralized benchmark; 2) current system without monitoring; 3) current system with monitoring; 4) operational mechanism (figure 2).

The cost of the current system without monitoring suffers from a fleet excess caused by the inflation of transportation needs coming from the low type Program. This inflation holds for parameter

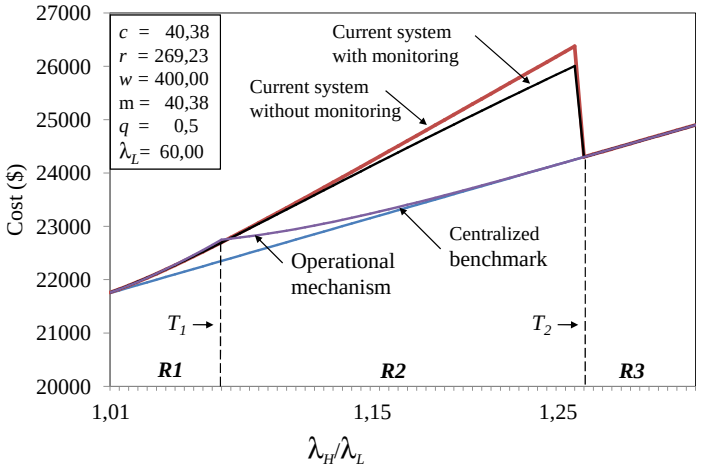


Figure 2 System cost comparison

values below the threshold for "natural truth revelation", T_2. Note that the cost of the current system without monitoring is an upper bound for the cost of the system with monitoring. The greater the monitoring cost, the closer the cost of the current system with monitoring to the upper bound. The centralized benchmark is a lower bound for the current system with monitoring. The lower the monitoring cost, the closer the current system with monitoring will be to the centralized benchmark.

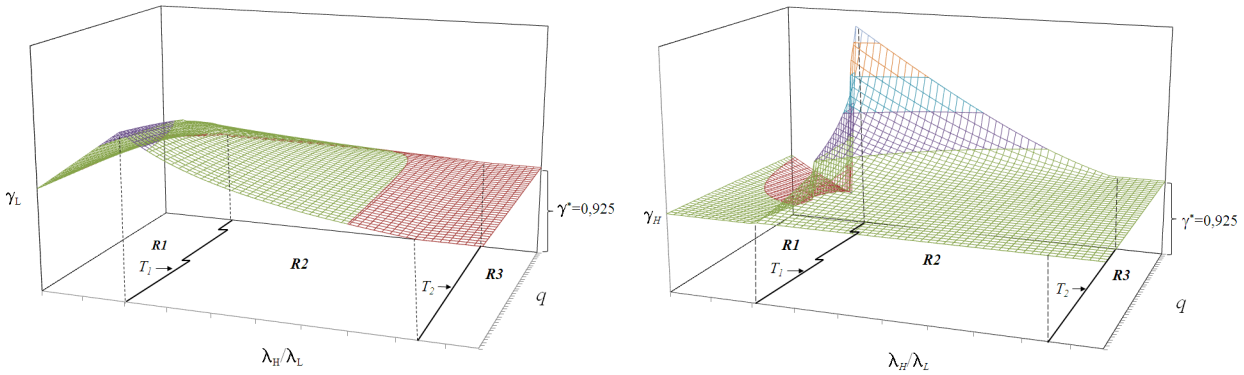
Also note that in *R1* the mechanism does not produce significant savings compared to the current system. In this region the Headquarters offer the same fleet size to both Programs. This strategy makes the low type Program indifferent between revealing the truth and distorting its needs. In *R2* the Headquarters make the low type Program indifferent between revealing its true needs and distorting them by offering $\gamma^* < \gamma_L$. But the system cost is controlled by choosing γ_L according to the proportion of low type Programs. The higher the proportion of low type Programs, the closer γ_L is to γ^*. In *R3* there is no need for incentives since truth revelation is achieved naturally. Results for *R1* and *R3* are quite intuitive. On the other hand, it was surprising to us that the larger the difference in types in *R2*, the lower the value of information for the low type program. We expand the explanation of this result below.

5.1. Lost of Value of Private Information for the Low Type Program

We begin by plotting the buffer factors in the $(\lambda_H/\lambda_L, q, \gamma)$ space (figure 3). The horizontal axis represents the transportation needs ratio λ_H/λ_L. The depth axis represents the low type Program probability q. The vertical axis represents the fleet buffer for the i type Program. Figure 3 is complemented with figure 4 showing the projection of γ_L and γ_H in the $(\lambda_H/\lambda_L, \gamma)$ space, and figure 5 showing the projection of γ_L and γ_H in the (q, γ) space.

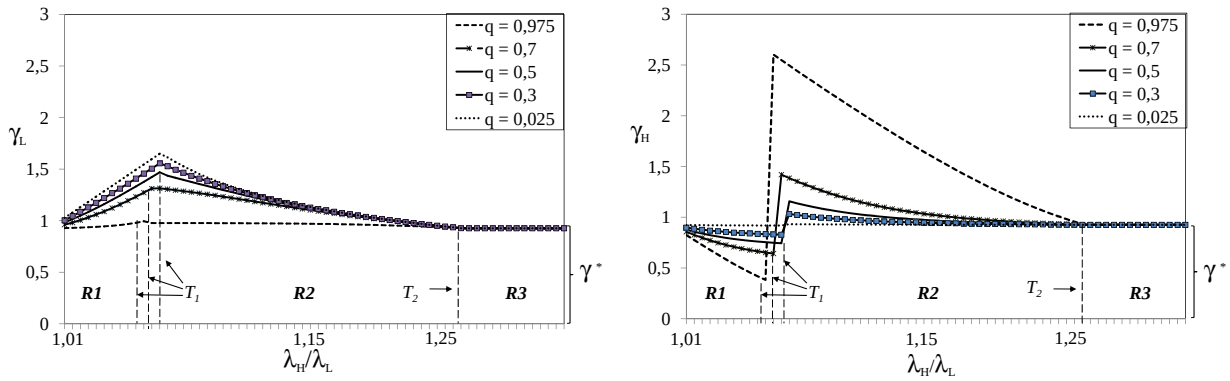
First, note that the Headquarters give incentives to the low type Program in the induced truth region *R1* by choosing $\gamma^* < \gamma_L$ (figures 4a and 5a). On the other hand, to mitigate the extra-cost of this strategy, the Headquarters simultaneously decrease the fleet buffer for the high type Program by choosing $\gamma_H < \gamma^*$ (figure 4b and $\lambda_H/\lambda_L = 1,01$ in figure 5b). This way the Headquarters force the indifference conditions $\gamma_H = \delta_H < \gamma^* < \gamma_L = \delta_L$. For low values of q in $R1$ both γ_L and γ_H are very close to γ^*. When q increases, the Headquarters offer a lower γ_H, which pushes down δ_L. The Headquarters make the low type Program indifferent between revealing the truth and inflating its needs by offering $\gamma_L = \delta_L$. The extra cost of delay for the high type Program is compensated via reduction of operating cost of the low type Programs fleet.

Second, note in figure 4a that in *R2* the greater the ratio between high type and low type needs, λ_H/λ_L, the lower the value of incentives given to the low type program. For low values of q the Headquarter give incentives to both Program types by choosing $\gamma^* < \gamma_H < \gamma_L < \bar{\gamma} < \delta_L = \tilde{\gamma}$. To keep the system cost balanced, the Headquarters offer the low type Program a fleet buffer greater than γ^*, but decreasing in the proportion of low type Programs, q, (figure 5a). In fact, for large values of q it is more efficient for the Headquarters to offer $\gamma^* < \gamma_L < \gamma_H < \bar{\gamma} < \delta_L = \tilde{\gamma}$. Nevertheless, decreasing γ_L increases $\tilde{\gamma}$, such that to have the indifference condition $\delta_L = \tilde{\gamma}$, the Headquarters must offer the high type Program a bigger fleet buffer, up to $\gamma_H = 1,7364$ when $q = 0,975$ (figure 5b). This counterintuitive results is explained by the fact that Program cost curves are u-shaped and do not fulfill the single crossing property. Hence, there is a region, *R2*, in which the difference between types is so big that it becomes costly for the low type Program to distort its type. The distortion would imply a marginal increase of fleet management cost relatively greater than the



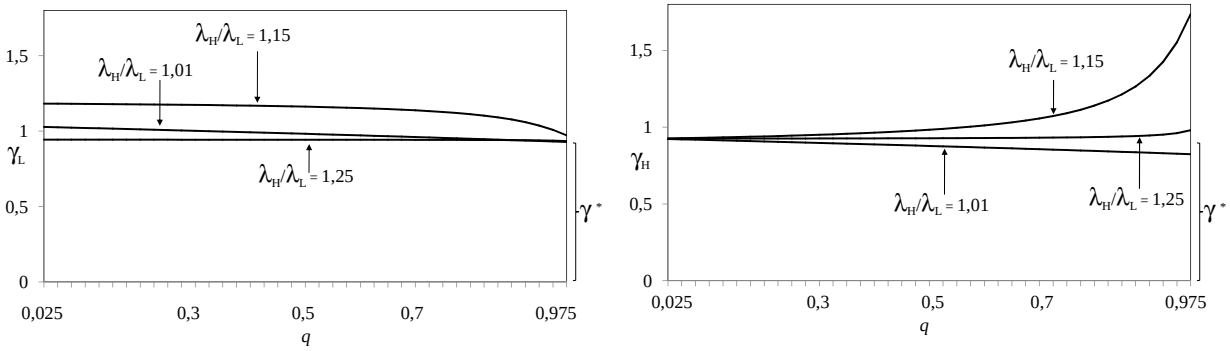
(a) Low type Program (b) High type Program

Figure 3 Fleet buffers in the $(\lambda_H/\lambda_L, q, \gamma)$ space



(a) Low type Program (b) High type Program

Figure 4 Fleet buffers as function of λ_H/λ_L



(a) Low Type Program (b) High Type Program

Figure 5 Fleet buffers as function of q

marginal decrease in the cost of delay in relation to fleet size. The lost of information rent for the low type Program makes the mechanism more valuable in instances where the difference between types is big.

5.2. Sensitivity Analysis on the Waiting Cost

Next we perform a sensitivity analysis with respect to w. Since we do not have data on the possible cost of delay, we study the sensitivity of the system to changes in this parameter by choosing $w = \{100; 200; 400; 600; 800\}$. We also use $q = \{0,025; 0,3; 0,5; 0,7; 0,975\}$ and lambda ratios $\lambda_H/\lambda_L = \{1,01; 1,15; 1,25\}$. The sensitivity analysis summarizes the changes in the Headquarter's optimal fleet buffer, γ^* and the Program's optimal fleet buffer, $\bar{\gamma}$ (table 1).

Note that a decrease in the cost of delay, w, causes a decrease in both the optimal fleet buffer for the Headquarter, γ^* and in the optimal fleet buffer for the Program $\bar{\gamma}$. The lower the cost of delay, the lower the relative weight of the equity component of the objective function for the system. On the other hand, an increase in w increases γ^* and $\bar{\gamma}$. Nevertheless, these changes in the optimal fleet buffer are not linear in w. Decreasing w by 50% starting at the $\lambda_H/\lambda_L = 1,15$ case decreases γ^* by 23,24% while increasing w by 50% only increases γ^* by 15,02%. Note that decreasing w expands *R3*, the natural truth telling region. Lower w implies that the fleet management cost for the Program is relatively higher. Hence the Program has lower incentives to distort transportation needs to secure a bigger fleet. On the other hand, increasing w slightly decreases *R3*.

The sensitivity analysis takes the current system without monitoring as a base case. Cost savings of centralize benchmark, operational mechanism, and current system with monitoring are calculated in comparison with the base case. Note that the cost savings for the 3 systems are increasing in q, the proportion of low type Programs. The savings are also increasing in the ratio between high type and low type needs, λ_H/λ_L. Regarding the mechanism, observe that it produces savings around 10% for values of $q = 0,7$ and $\lambda_H/\lambda_L = 1,25$, simultaneously. The mechanism savings increase to more than 15% for values of $q = 0,975$ and $\lambda_H/\lambda_L = 1,25$, simultaneously. Compared to the current system with monitoring, these two instances produce savings around 8% and 10%, respectively, matching the intuition of the Headquarters staff quoted in the introduction.

This section presented a cost comparison between the different regimes analyzed in the paper. It also illustrated graphically the Headquarters‘ strategies to achieve truth revelation using the operational mechanism. These strategies vary depending on parameter values. While in *R1* the

	$\lambda_H/\lambda_L = 1,01$					$\lambda_H/\lambda_L = 1,15$					$\lambda_H/\lambda_L = 1,25$				
w =	100	200	400	600	800	100	200	400	600	800	100	200	400	600	800
						Fleet Buffer Measures†									
γ^*	0,529	0,710	0,925	1,064	1,166	0,529	0,710	0,925	1,064	1,166	0,529	0,710	0,925	1,064	1,166
$\bar{\gamma}$	1,246	1,468	1,701	1,840	1,939	1,246	1,468	1,701	1,840	1,939	1,246	1,468	1,701	1,840	1,939
$T_1^{\ddagger}$	1,055	1,065	1,070	1,070	1,070	1,055	1,065	1,070	1,070	1,070	1,055	1,065	1,070	1,070	1,070
T_2	1,220	1,250	1,265	1,265	1,270	1,220	1,250	1,265	1,265	1,270	1,220	1,250	1,265	1,265	1,270
q						**Low Type Program Fleet Buffer,** γ_L									
0,025	0,630	0,811	1,028	1,167	1,270	0,673	0,924	1,182	1,331	1,434	0,529	0,710	0,943	1,088	1,191
0,300	0,607	0,787	1,002	1,141	1,244	0,670	0,918	1,174	1,322	1,425	0,529	0,710	0,943	1,088	1,191
0,500	0,588	0,767	0,982	1,121	1,223	0,665	0,909	1,162	1,310	1,413	0,529	0,710	0,943	1,087	1,191
0,700	0,567	0,746	0,961	1,099	1,202	0,655	0,890	1,138	1,284	1,388	0,529	0,710	0,942	1,087	1,190
0,975	0,533	0,713	0,928	1,067	1,170	0,560	0,750	0,972	1,114	1,217	0,529	0,710	0,934	1,075	1,178
q						**High Type Program Fleet Buffer,** γ_H									
0,025	0,527	0,708	0,923	1,062	1,164	0,530	0,711	0,927	1,066	1,168	0,529	0,710	0,925	1,064	1,167
0,300	0,504	0,683	0,898	1,036	1,138	0,541	0,729	0,950	1,091	1,193	0,529	0,710	0,926	1,066	1,168
0,500	0,486	0,664	0,878	1,016	1,118	0,557	0,755	0,983	1,125	1,228	0,529	0,710	0,928	1,068	1,170
0,700	0,465	0,643	0,856	0,994	1,096	0,594	0,815	1,057	1,201	1,304	0,529	0,710	0,931	1,072	1,175
0,975	0,431	0,610	0,824	0,962	1,064	1,018	1,404	1,736	1,901	2,006	0,529	0,710	0,980	1,138	1,243
						SYSTEM COSTS									
q						**Current System without Monitoring ($)**										
0,025	20745	21270	21846	22197	22450	23465	24022	24636	25010	25280	25308	25890	26632	27023	27303
0,300	20699	21222	21797	22147	22399	23263	23795	24391	24761	25028	23967	24532	26370	26756	27035
0,500	20666	21187	21761	22111	22363	23116	23629	24214	24579	24845	22991	23544	26180	26561	26839
0,700	20632	21152	21725	22075	22326	22969	23464	24036	24397	24662	22016	22556	25990	26367	26644
0,975	20586	21105	21676	22025	22276	22767	23236	23792	24147	24410	20674	21198	25728	26100	26375
						SYSTEM SAVINGS (%) VS. CURRENT SYSTEM WITHOUT MONITORING									
q						**Centralized Benchmark**									
0.025	0.01	0.01	0.01	0.00	0.01	0.24	0.22	0.22	0.22	0.22	0.00	0.00	0.38	0.38	0.38
0.300	0.05	0.04	0.04	0.03	0.03	2.84	2.70	2.61	2.58	2.57	0.00	0.00	4.62	4.57	4.56
0.500	0.08	0.06	0.06	0.05	0.05	4.75	4.53	4.38	4.33	4.31	0.00	0.00	7.75	7.68	7.64
0.700	0.10	0.09	0.08	0.07	0.07	6.70	6.39	6.18	6.10	6.07	0.00	0.00	10.93	10.83	10.78
0.975	0.14	0.12	0.11	0.10	0.10	9.40	8.98	8.70	8.58	8.54	0.00	0.00	15.37	15.23	15.16
q						**Operational Mechanism**									
0.025	0.00	0.00	0.00	0.00	0.00	0.23	0.21	0.20	0.20	0.20	0.00	0.00	0.38	0.38	0.38
0.300	0.00	0.00	0.00	0.00	0.00	2.73	2.53	2.39	2.36	2.35	0.00	0.00	4.62	4.57	4.55
0.500	0.01	0.01	0.01	0.01	0.01	4.58	4.24	4.03	3.97	3.96	0.00	0.00	7.75	7.67	7.64
0.700	0.04	0.04	0.04	0.04	0.04	6.47	6.01	5.73	5.64	5.62	0.00	0.00	10.92	10.82	10.77
0.975	0.13	0.11	0.10	0.10	0.09	9.31	8.84	8.53	8.42	8.37	0.00	0.00	15.37	15.23	15.16
q						**Current System With Monitoring**									
0.025	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.00
0.300	0.00	0.00	0.00	0.00	0.00	0.17	0.15	0.14	0.13	0.13	0.00	0.00	0.44	0.44	0.44
0.500	0.00	0.00	0.00	0.00	0.00	0.47	0.41	0.39	0.37	0.37	0.00	0.00	1.29	1.27	1.26
0.700	0.00	0.00	0.00	0.00	0.00	0.94	0.84	0.78	0.76	0.75	0.00	0.00	2.67	2.61	2.59
0.975	0.00	0.00	0.00	0.00	0.00	1.90	1.71	1.58	1.54	1.52	0.00	0.00	5.55	5.45	5.39

† The fleet buffer measures remain constant to changes in λ_H/λ_L
‡$q = 0,5$

Table 1 Sensitivity of results to changes in w

Headquarters offer the same fleet size to both Program types, in *R2* the strategy consists of offering incentives to both Program types and keeping the buffer for the Program with the biggest expected proportion closer to γ^*. Finally, this section included a sensitivity analysis with respect to w. The effectiveness of the mechanism compared to the current system increases with w as long as

the parameters fall into the induced truth regions. The next section presents the conclusions and possible avenues for future research.

6. Conclusions and Further Research

Completely informed by field research, this paper studies a decentralized two party fleet management system in a humanitarian setting. Located in the field, Programs are service oriented and have private information on their transportation needs. Transportation needs are fulfilled using vehicle fleets. Located in the US or Europe, the Headquarters main objective is balancing the service level and the operating cost of the fleet. The Headquarters can monitor the Program's stated transportation needs to avoid this distortion. Differences in objectives, distant geographic locations and the private information of Programs about their transportation needs give rise to an adverse selection problem. The Headquarters are concerned about the excess of fleet size. The Headquarters are also concerned about the lack of effectiveness of monitoring tools. Finally, the Programs are concerned about the lack of vehicles to respond to their transportation needs. We develop a mathematically tractable model to analyze this problem while respecting exogenous constraints that exist in the humanitarian context regarding internal budget allocation.

We find that the concerns of the two parties in the system are rational from each party's perspective. In the current fleet management system the Headquarters monitoring effort does not dissuade the low type Program from inflating transportation needs. This is because of the impossibility that the Headquarters punish the distorted needs stated by the Program. We also find that the optimal buffer factor offered by the Headquarters is lower than the optimal fleet buffer intended by the Program. Hence, the lack of effectiveness of monitoring tools added to the incentive of the low type Program to distort transportation needs produce a fleet excess that justifies the Headquarters concerns. Finally, the Program's concerns about not having enough vehicles to respond to its needs are explained by the fact that its optimal fleet buffer is greater than the one offered by the Headquarters.

Nevertheless, we find that for appropriate parameter combinations the current system has a "natural" truth telling threshold in which the centralized benchmark solution can be achieved. In

the current fleet management system the high type Program always reveals its true transportation needs. Otherwise, the high type Program would receive a lower fleet buffer compared to the one offered by the Headquarters, increasing even more this Program's costs of delay. The threshold for natural truth telling arises when the extra fleet management cost for the fleet excess intended by the low type Program dominates the savings from the reduction in the cost of delay. The coordination of incentives is particularly challenging since financial transfer payments are not a viable way to induce Program's truth revelation in this humanitarian setting.

We propose a novel operational capacity based mechanism to coordinate incentives in this system. In this mechanism the Headquarters offer different fleet buffer factors to the different Program types. Additionally, the monitoring role is suppressed. We show the existence of three mutually exclusive and collectively exhaustive regions for truth revelation under the proposed mechanism. *R1* is called the equal fleet size region. *R2* is called the different fleet size region. Finally, *R3* is called the natural truth revelation region.

In the equal fleet size region the Headquarters achieve truth revelation by offering both Program types the same fleet size. This strategy makes the low type Program indifferent between inflating its transportation needs and revealing the truth. In the different fleet size region both Program types are offered fleet buffers greater than optimal fleet buffer for the current system. When small proportions of low type Programs are expected, the low type Program is offered a fleet buffer such that this Program's cost when inflating its needs equals the cost of revealing its true needs. When the low type probability is high enough, it is cheaper for the Headquarters to incentivize the high type Program. By offering the high type a bigger fleet buffer, the Headquarters increase the low type Program's intended fleet buffer enough to make this Program indifferent between revealing its needs and inflating them. Finally, under some parameter combination, which is independent of the low type probability, the system reaches a natural truth telling region. In this region, as in the current system, the extra-cost of fleet management deters the low type Program from inflating its transportation needs. Additionally, we show that under the mechanism the high type Program

always reveals his true transportation needs. This result is similar to the one we obtained for the current system.

Our numerical section complements the analysis. First, it allows us to explain the lost of value of private information for the low type program as a function of the increase in the difference with the high type program. Second, the numerical experiments show the behavior of the system due to changes in the cost of delay. As expected, a decrease in the cost of delay decreases the optimal fleet buffers for the system. An increase in the cost of delay increases the fleet buffers for the system. But the change in the optimal fleet buffers is not linear to changes in the cost of delay. Equally, the thresholds defining the regions for truth telling are much more sensitive to decreases in the cost of delay than they are to the increases in that cost. This is because the decrease of the cost of delay makes the Program's cost function flatter rapidly, increasing the size of the natural truth revelation region.

This paper introduces a tractable mathematical model to analyze incentive alignment in decentralized humanitarian settings. Some interesting extensions of this work include fleet pooling and the joint analysis of relief and development transportation needs, which are issues faced by humanitarian fleet managers in practice.

Appendix. Proofs

Proof of Lemma 1: Note that:

$$\frac{\partial C_{Cent}(\gamma,)}{\partial \gamma} = \left[\frac{w\pi'(\gamma)}{\gamma} - \frac{w\pi(\gamma)}{\gamma^2} + c + r\right]\sqrt{\lambda} \tag{16}$$

and:

$$\frac{\partial C_{Prog}(\gamma)}{\partial \gamma} = \left[\frac{w\pi'(\gamma)}{\gamma} - \frac{w\pi(\gamma)}{\gamma^2} + c\right]\sqrt{\lambda} \tag{17}$$

Hence

$$\frac{\partial C_{Prog}(\gamma)}{\partial \gamma} = \frac{\partial C_{Cent}(\gamma,\lambda)}{\partial \gamma} - r\sqrt{\lambda}$$
$$\frac{\partial C_{Prog}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma^*(c,r,w)} = \frac{\partial C_{Cent}(\gamma,\lambda)}{\partial \gamma}\Big|_{\gamma=\gamma^*(c,r,w)} - r\sqrt{\lambda}$$
$$\frac{\partial C_{Prog}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma^*(c,r,w)} = 0 - r\sqrt{\lambda} < 0$$

Because $C_{Prog}(\gamma)$ is unimodal with a finite minimum, $C_{Prog}(c,w)$ decreases for values of γ such that $\gamma^*(c,r,w) < \gamma$, reaching its minimum at $\gamma = \bar{\gamma}(c,w)$. This implies $\gamma^*(c,r,w) < \bar{\gamma}(c,w)$. □

Proof of Proposition 1: Given the central benchmark solution γ^*, we have $\delta_H = \frac{\lambda_L + \gamma^*\sqrt{\lambda_L} - \lambda_H}{\sqrt{\lambda_H}}$. $\delta_H < \gamma^*$ follows from $\lambda_L < \lambda_H$. Since $\frac{\partial C_{Prog}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma^*} < 0$ (from lemma 1), and due to the fact that the cost function $C_{Prog}(\gamma)$ is unimodal with minimum at $\bar{\gamma}$, and given that from lemma 1 we know that $\gamma^* < \bar{\gamma}$, it follows that $C_{Prog}(\gamma^*) < C_{Prog}(\delta_H)$. □

Proof of Proposition 2: Note that $\bar{\gamma} < \hat{\gamma}_L$. If $\delta_L < \hat{\gamma}_L$, then $C_{Prog}(\delta_L) < C_{Prog}(\gamma^*)$. If $\delta_L = \hat{\gamma}_L$, then $C_{Prog}(\delta_L) = C_{Prog}(\gamma^*)$. Finally, if $\hat{\gamma}_L < \delta_L$, then $C_{Prog}(\gamma^*) < C_{Prog}(\delta_L)$. Hence, $\hat{\gamma}_L$ defines the truth telling region for the low type Program. □

Proof of Proposition 3: A Program of type i would report the true transportation needs if $wQ(\gamma) + cF(\gamma) \leq p(wQ(\gamma) + cF(\gamma)) + (1-p)(wQ(\delta_i) + cF(\delta_i))$. Since $wQ(\gamma) + cF(\gamma) = p\,(wQ(\gamma) + cF(\gamma)) + (1-p)\,(wQ(\gamma) + cF(\gamma))$. Using this fact we get the truth telling condition $wQ(\gamma) + cF(\gamma) \leq wQ(\delta_i) + cF(\delta_i)$. The proof follows from the fact that the condition for truth telling does not depend on p, the Headquarters monitoring effort. □

Proof of Proposition 4: First, we rewrite the mechanism in extended form as:

$$\begin{aligned}\min_{0<\gamma_L,0<\gamma_H} E[C_{Mec}] = {} & q\left[\frac{w\pi(\gamma_L)\sqrt{\lambda_L}}{\gamma_L} + (c+r)(\lambda_L + \gamma_L\sqrt{\lambda_L})\right] \\ & + (1-q)\left[\frac{w\pi(\gamma_H)\sqrt{\lambda_H}}{\gamma_H} + (c+r)(\lambda_H + \gamma_H\sqrt{\lambda_H})\right] \\ S.T.& \\ (IC_L): \quad & \frac{w\pi(\gamma_L)\sqrt{\lambda_L}}{\gamma_L} + c(\lambda_L + \gamma_L\sqrt{\lambda_L}) \leq \frac{w\pi(\delta_L)\sqrt{\lambda_L}}{\delta_L} + c(\lambda_L + \delta_L\sqrt{\lambda_L}) \\ (IC_H): \quad & \frac{w\pi(\gamma_H)\sqrt{\lambda_H}}{\gamma_H} + c(\lambda_H + \gamma_H\sqrt{\lambda_H}) \leq \frac{w\pi(\delta_H)\sqrt{\lambda_H}}{\delta_H} + c(\lambda_H + \delta_H\sqrt{\lambda_H})\end{aligned} \tag{18}$$

Second, we build the mechanism's lagrangian function $\mathcal{L}$.

$$\begin{aligned}\mathcal{L}(\gamma_L,\gamma_H,\alpha_1,\alpha_2) = {} & q\left[\frac{w\pi(\gamma_L)\sqrt{\lambda_L}}{\gamma_L} + (c+r)(\lambda_L + \gamma_L\sqrt{\lambda_L})\right] \\ & + (1-q)\left[\frac{w\pi(\gamma_H)\sqrt{\lambda_H}}{\gamma_H} + (c+r)(\lambda_H + \gamma_H\sqrt{\lambda_H})\right] \\ & + \alpha_1\left[\frac{w\pi(\gamma_L)\sqrt{\lambda_L}}{\gamma_L} + c(\lambda_L + \gamma_L\sqrt{\lambda_L}) - \left(\frac{w\pi(\delta_L)\sqrt{\lambda_L}}{\delta_L} + c(\lambda_L + \delta_L\sqrt{\lambda_L})\right)\right] \\ & + \alpha_2\left[\frac{w\pi(\gamma_H)\sqrt{\lambda_H}}{\gamma_H} + c(\lambda_H + \gamma_H\sqrt{\lambda_H}) - \left(\frac{w\pi(\delta_H)\sqrt{\lambda_H}}{\delta_H} + c(\lambda_H + \delta_H\sqrt{\lambda_H})\right)\right]\end{aligned} \tag{19}$$

Where α_1 and α_2 are the lagrange multipliers for IC_L and IC_H, respectively. Third, we derive the first order conditions (FOC) of (19).

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\gamma_L} &= (q+\alpha_1)\left[\frac{w\pi'(\gamma_L)}{\gamma_L} - \frac{w\pi(\gamma_L)}{\gamma_L^2} + c\right]\sqrt{\lambda_L} + qr\sqrt{\lambda_L} - \alpha_2\left[\frac{w\pi'(\delta_H)}{\delta_H} - \frac{w\pi(\delta_H)}{\delta_H^2} + c\right]\sqrt{\lambda_L} = 0 \\ \frac{\partial\mathcal{L}}{\partial\gamma_H} &= (1-q+\alpha_2)\left[\frac{w\pi'(\gamma_H)}{\gamma_H} - \frac{w\pi(\gamma_H)}{\gamma_H^2} + c\right]\sqrt{\lambda_H} + (1-q)r\sqrt{\lambda_H} - \alpha_1\left[\frac{w\pi'(\delta_L)}{\delta_L} - \frac{w\pi(\delta_L)}{\delta_L^2} + c\right]\sqrt{\lambda_H} = 0\end{aligned}$$

Letting $f(\gamma) = \frac{w\pi'(\gamma)}{\gamma} - \frac{w\pi(\gamma)}{\gamma^2} + c$ and dividing by $\sqrt{\lambda_L}$ in the first equation above and dividing by $\sqrt{\lambda_H}$ in the second equation above, we can re-write the FOC as:

$$\frac{\partial\mathcal{L}}{\partial\gamma_L} = (q+\alpha_1)f(\gamma_L) - \alpha_2 f(\delta_H) + qr = 0 \tag{20}$$

$$\frac{\partial\mathcal{L}}{\partial\gamma_H} = (1-q+\alpha_2)f(\gamma_H) - \alpha_1 f(\delta_L) + (1-q)r = 0 \tag{21}$$

Note that

$$\frac{\partial C_{Prog}(\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma_L} = f(\gamma_L)\sqrt{\lambda_L} \tag{22}$$

$$\frac{\partial C_{Cent}(\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma_L} = (f(\gamma_L) + r)\sqrt{\lambda_L} \tag{23}$$

Replacing (22) and (23) in (20) and (21) we can re-write the FOC as:

$$q\frac{\partial C_{Cent}(\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma_L} + \alpha_1\frac{\partial C_{Prog}(\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma_L} - \alpha_2\frac{\partial C_{Prog}(\gamma)}{\partial\gamma}\Big|_{\gamma=\delta_H} = 0 \tag{24}$$

$$(1-q)\frac{\partial C_{Cent}(\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma_H} + \alpha_2\frac{\partial C_{Prog}(\gamma)}{\partial\gamma}\Big|_{\gamma=\gamma_H} - \alpha_1\frac{\partial C_{Prog}(\gamma)}{\partial\gamma}\Big|_{\gamma=\delta_L} = 0 \tag{25}$$

Equations (9) and the fact that $\lambda_L < \lambda_H$ imply that

$$\gamma_H < \delta_L \tag{26}$$

$$\delta_H < \gamma_L \tag{27}$$

Combining the definitions of δ_L and δ_H we get the useful relation:

$$(\delta_L - \gamma_L)\sqrt{\lambda_L} = (\gamma_H - \delta_H)\sqrt{\lambda_H} \tag{28}$$

Equation (28) implies the set of conditions:

$$\gamma_L = \delta_L \text{ if and only if } \gamma_H = \delta_H; \quad \gamma_L < \delta_L \text{ if and only if } \delta_H < \gamma_H; \delta_L < \gamma_L \text{ if and only if } \gamma_H < \delta_H \tag{29}$$

Fourth, we characterize the induced truth telling region stated in the proposition.

Characterization of *R1*: Suppose $0 < \alpha_1$ and $0 < \alpha_2$. Note that $0 < \alpha_1$ implies $C_{Prog}(\gamma_L) = C_{Prog}(\delta_L)$ and $0 < \alpha_2$ implies $C_{Prog}(\gamma_H) = C_{Prog}(\delta_H)$.

There are three mutually exclusive and collectively exhaustive possibilities: either 1) $\delta_L < \bar{\gamma}$, or 2) $\delta_L = \bar{\gamma}$, or 3) $\bar{\gamma} < \delta_L$.

Subcase 1: $\delta_L < \bar{\gamma}$.

Note that $\delta_L < \bar{\gamma}$ implies that either $\delta_L < \bar{\gamma} < \gamma_L$ or $\delta_L = \gamma_L$. First, suppose that $\delta_L < \bar{\gamma} < \gamma_L$. Then, $\gamma_H < \delta_H$ (due to condition (29)). It also implies $\gamma_H < \delta_L < \bar{\gamma}$, and $C_{Prog}(\delta_L) < C_{Prog}(\gamma_H)$ follows from the fact that the Program's cost function is well behaved.

For $C_{Prog}(\gamma_H) = C_{Prog}(\lambda_H)$ to be true it must be that $\bar{\gamma} < \delta_H$. Also, $\delta_H < \gamma_L$ (from condition (26)), which implies that $C_{Prog}(\delta_H) < C_{Prog}(\gamma_L)$. But $C_{Prog}(\gamma_L) = C_{Prog}(\delta_L)$ (because we supposed $0 < \alpha_1$). It follows that $C_{Prog}(\delta_H) < C_{Prog}(\gamma_L) = C_{Prog}(\delta_L) < C_{Prog}(\gamma_H)$, which is a contradiction since $\alpha_2 > 0$ implies $C_{Prog}(\delta_H) = C_{Prog}(\gamma_H)$.

Second, suppose that $\delta_L = \gamma_L$. This implies $\gamma_H = \delta_H$ (from condition (29)), and $\gamma_H < \gamma_L$ follows (from condition (27)) and the fact that $\delta_L = \gamma_L$). This also implies $\delta_L = \gamma_L < \bar{\gamma}$ (from condition $0 < \alpha_1$). Using these facts in the first order conditions (24) and (25) and adding them we get:

$$q\frac{\partial C_{Cent}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma_L} + (1-q)\frac{\partial C_{Cent}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma_H} = 0 \tag{30}$$

Equation (30) only holds in three possible cases: 1) when $\gamma_L = \gamma_H = \gamma^*$, contradicting the fact $\gamma_H < \gamma_L$; 2) when $\gamma_L < \gamma^* < \gamma_H$, contradicting the fact that $\gamma_H < \gamma_L$, and 3) when $\gamma_H < \gamma^* < \gamma_L$. This case is feasible and leads to the condition

$$\delta_H = \gamma_H < \gamma^* < \gamma_L = \delta_L < \bar{\gamma} \tag{31}$$

Subcase 2: $\delta_L = \bar{\gamma}$.

This implies $\gamma_L = \delta_L = \bar{\gamma}$ (since the Program cost function is unimodal). It also implies that $\gamma_H = \delta_H < \bar{\gamma}$. Using these facts in the first order conditions (24) and (25) and adding them we get again equation (30). As in the previous subcase, this equation only holds in three possible cases: 1) when $\gamma_L = \gamma_H = \gamma^*$, contradicting the fact $\gamma_H < \gamma_L$; 2) when $\gamma_L < \gamma^* < \gamma_H$, contradicting the fact that $\gamma_H < \gamma_L$; and 3) when $\gamma_H < \gamma^* < \gamma_L$. This leads to the condition:

$$\delta_H = \gamma_H < \gamma^* < \gamma_L = \delta_L = \bar{\gamma} \tag{32}$$

Subcase 3: $\bar{\gamma} < \delta_L$. The proof for this case follows the same logic that the one for subcase 1. Note that $\bar{\gamma} < \delta_L$ implies either that $\gamma_L < \bar{\gamma} < \delta_L$ or $\bar{\gamma} < \delta_L = \gamma_L$. First, suppose that $\gamma_L < \bar{\gamma} < \delta_L$. This implies $\delta_H < \gamma_H$ (from condition 29) and $\delta_H < \gamma_L < \bar{\gamma}$. It follows that $C_{Prog}(\gamma_L) < C_{Prog}(\delta_H)$ (given that the Program cost function is well behaved). Since the cost function of the program is unimodal, for $C_{Prog}(\delta_H) = C_{Prog}(\gamma_H)$ to hold it must be that $\bar{\gamma} < \gamma_H < \delta_L$. This implies $C_{Prog}(\gamma_H) < C_{Prog}(\delta_L)$. Hence we have $C_{Prog}(\gamma_H) < C_{Prog}(\delta_L) = C_{Prog}(\gamma_L) < C_{Prog}(\delta_H)$, which contradicts the condition $C_{Prog}(\gamma_H) = C_{Prog}(\delta_H)$ (because we supposed $0 < \alpha_2$). Second, suppose that $\bar{\gamma} < \delta_L = \gamma_L$. This implies $\delta_H = \gamma_H$. By using those relations in first order conditions (24) and (25) and adding them, we get equation (30). Following the same reasoning that we used in the two previous subcases, we find the following condition:

$$\delta_H = \gamma_H < \gamma^* < \bar{\gamma} < \gamma_L = \delta_L \tag{33}$$

We can combine conditions (31), (32) and (33) in the following condition, **which characterizes *R1* in Proposition 4**:

$$\delta_H = \gamma_H < \gamma^* < \gamma_L = \delta_L \tag{34}$$

Note that condition (34) implies that $F_L = F_H$ in *R1*. Depending on parameter values, to make the low type Program indifferent between telling the truth and lying the Headquarters have to sacrifice some cost for the high type Program.

Characterization of *R2* Suppose that $0 < \alpha_1$ and $\alpha_2 = 0$. FOC (24) and (25) become:

$$q\frac{\partial C_{Cent}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma_L} + \alpha_1\frac{\partial C_{Prog}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma_L} = 0 \tag{35}$$

$$(1-q)\frac{\partial C_{Cent}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma_H} - \alpha_1\frac{\partial C_{Prog}(\gamma)}{\partial \gamma}\Big|_{\gamma=\delta_L} = 0 \tag{36}$$

Condition (35) holds when $\frac{\partial C_{Cent}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma_L}$ and $\frac{\partial C_{Prog}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma_L}$ have opposite signs. This only happens when:

$$\gamma^* < \gamma_L < \bar{\gamma} \tag{37}$$

Condition (36) holds when both $\frac{\partial C_{Cent}(\gamma)}{\partial \gamma}\Big|_{\gamma=\gamma_H}$ and $\frac{\partial C_{Prog}(\gamma)}{\partial \gamma}\Big|_{\gamma=\delta_L}$ have the same sign. This happens in two mutually exclusive cases: 1) $\gamma_H < \gamma^*$ and $\delta_L < \bar{\gamma}$ simultaneously or, 2) $\gamma^* < \gamma_H$ and $\bar{\gamma} < \delta_L$ simultaneously.

First, combining condition (37) with $\gamma_H < \gamma^*$ and $\delta_L < \bar{\gamma}$ simultaneously we get $\gamma^* < \gamma_L < \bar{\gamma}$, $\gamma_H < \gamma^*$ and $\delta_L < \bar{\gamma}$. We supposed $0 < \alpha_1$. This implies $C_{Prog}(\gamma_L) = C_{Prog}(\delta_L)$. We also supposed $\alpha_2 = 0$ which implies $C_{Prog}(\gamma_H) < C_{Prog}(\delta_H)$. Since

$\gamma_L < \bar{\gamma}$ and $\delta_L < \bar{\gamma}$, it must be that $\gamma_L = \delta_L$. This implies $\gamma_H = \delta_H$ (from one of the conditions 29) and $C_{Prog}(\gamma_H) = C_{Prog}(\delta_H)$ follows, contradicting $C_{Prog}(\gamma_H) < C_{Prog}(\delta_H)$.

Second, combining condition (37) with $\gamma^* < \gamma_H$ and $\bar{\gamma} < \delta_L$ simultaneously we get $\gamma^* < \gamma_L < \bar{\gamma}$, $\gamma^* < \gamma_H$, and $\bar{\gamma} < \delta_L$. Remember the definition of $\tilde{\gamma}_L$ presented in (15). It must be that $\bar{\gamma} < \tilde{\gamma}_L = \delta_L$. Otherwise, the low type Program would claim high transportation needs violating the revelation principle. It follows that:

$\gamma^* < \gamma_L < \bar{\gamma} < \tilde{\gamma}_L = \delta_L$, $\gamma^* < \gamma_H$ and $\delta_H < \gamma_H$. These conditions can be divided in three sub-cases. The first Sub-case is $\gamma^* < \gamma_H < \gamma_L < \bar{\gamma} < \tilde{\gamma}_L = \delta_L$, $\delta_H < \gamma_H$. For low values of q, these two conditions **characterize *R2* in Proposition 4.**

The second sub-case is $\gamma^* < \gamma_L < \gamma_H < \bar{\gamma} < \tilde{\gamma}_L = \delta_L$, $\delta_H < \gamma_L$. The third sub-case is $\gamma^* < \gamma_L < \bar{\gamma} < \gamma_H < \tilde{\gamma}_L = \delta_L$, $\delta_H < \gamma_L$. For high values of q **sub-cases two and three characterize *R2* in Proposition 4**.

Next, we show the existence of the threshold T_1 in Proposition 4. Let $g(\gamma) = \frac{w\pi'(\gamma)}{\gamma} - \frac{w\pi(\gamma)}{\gamma^2} + (c+r)$. Note that:

$$qg(\gamma_L)\sqrt{\lambda_L} + (1-q)g(\gamma_H)\sqrt{\lambda_H}\left(\frac{\partial \gamma_H}{\partial \gamma_L}\right) = 0 \tag{38}$$

is a required condition for the FOC of the mechanism. The justification is as follows.

First, for *R1* we know that $\gamma_L = \delta_L$ and $\gamma_H = \delta_H$. Replacing these values in FOC (20) and (21) we get:

$$(q+\alpha_1)f(\gamma_L) - \alpha_2 f(\gamma_H) + qr = 0$$
$$(1-q+\alpha_2)f(\gamma_H) - \alpha_1 f(\gamma_L) + (1-q)r = 0$$

Solving the system for α_1 and α_2 we get:

$$q(f(\gamma_L)+r) + (1-q)(f(\gamma_H)+r) = 0 \tag{39}$$

The equivalence between (38) and (39) follows from the fact that in (39) $\frac{\partial \gamma_H}{\partial \gamma_L} = \frac{\sqrt{\lambda_L}}{\sqrt{\lambda_H}}$.

Second, for *R2* we know that $\alpha_2 = 0$. Hence, FOC (20) and (21) become:

$$\frac{\partial \mathcal{L}}{\partial \gamma_L} = (q+\alpha_1)f(\gamma_L) + qr = 0$$
$$\frac{\partial \mathcal{L}}{\partial \gamma_H} = (1-q)f(\gamma_H) - \alpha_1 f(\delta_L) + (1-q)r = 0$$

Then we get $\alpha_1 = -\frac{q(f(\gamma_L)+r)}{f(\gamma_L)}$ such that:

$$(1-q)f(\gamma_H) + \frac{q(f(\gamma_L)+r)}{f(\gamma_L)}f(\delta_L) + (1-q)r = 0$$
$$q(f(\gamma_L)+r) + (1-q)(f(\gamma_H)+r)\frac{f(\gamma_L)}{f(\delta_L)} = 0 \tag{40}$$

Remember the relation $\delta_L = \tilde{\gamma}_L$ in *R2*. Using this we get $\gamma_H = \frac{\lambda_L + \tilde{\gamma}_L\sqrt{\lambda_L} - \lambda_H}{\sqrt{\lambda_H}}$. We can write $\frac{\partial \gamma_H}{\partial \gamma_L} = \frac{\partial \gamma_H}{\partial \tilde{\gamma}_L}\frac{\partial \tilde{\gamma}_L}{\partial \gamma_L}$. Note that $\frac{\partial \gamma_H}{\partial \tilde{\gamma}_L} = \frac{\sqrt{\lambda_L}}{\sqrt{\lambda_H}}$ and $\frac{\partial \tilde{\gamma}_L}{\partial \gamma_L} = \frac{f(\gamma_L)}{f(\tilde{\gamma}_L)}$. The equivalence between (38) and (40) follows.

Keeping λ_L fixed, the next step is obtaining $\frac{dC}{d\lambda_H}$.

$$\frac{dC}{d\lambda_H} = (1-q)\left[\frac{w\pi(\gamma_H)}{2\gamma_H\sqrt{\lambda_H}} + (c+r)\left(1 + \frac{\gamma_H}{2\sqrt{\gamma_H}}\right) + g(\gamma_H)\sqrt{\lambda_H}\left(\frac{\partial \gamma_H}{\partial \lambda_H}\right)\right]$$
$$+ \frac{\partial \gamma_L}{\partial \lambda_H}\left[qg(\gamma_L)\sqrt{\lambda_L} + (1-q)g(\gamma_H)\sqrt{\lambda_H}\left(\frac{\partial \gamma_H}{\partial \gamma_L}\right)\right]$$

Note that from condition (38) the second part of $\frac{dC}{d\lambda_H}$ equals zero. Hence,

$$\frac{dC}{d\lambda_H} = (1-q)\left[\frac{w\pi(\gamma_H)}{2\gamma_H\sqrt{\lambda_H}} + (c+r)\left(1 + \frac{\gamma_H}{2\sqrt{\lambda_H}}\right) + g(\gamma_H)\sqrt{\lambda_H}\left(\frac{\partial \gamma_H}{\partial \lambda_H}\right)\right] \tag{41}$$

For *R1* let $\gamma_H^1 = \frac{\lambda_L + \gamma_L^1\sqrt{\lambda_L} - \lambda_H}{\sqrt{\lambda_H}}$. Then

$$\frac{\partial \gamma_H^1}{\partial \lambda_H} = -\frac{\lambda_L + \gamma_L\sqrt{\lambda_L}}{2(\lambda_H)^{3/2}} - \frac{1}{2\sqrt{\lambda_H}}$$
$$= -\frac{\lambda_L + \gamma_L^1\sqrt{\lambda_L} + \lambda_H}{2\lambda_H\sqrt{\lambda_H}}$$

Such that

$$g(\gamma_H)\sqrt{\lambda_H}\left(\frac{\partial \gamma_H}{\partial \lambda_H}\right) = \left(\frac{w\pi'(\gamma_H)}{\gamma_H} - \frac{w\pi(\gamma_H)}{\gamma_H^2} + c + r\right)(-\sqrt{\lambda_H})\left(\frac{\lambda_L\gamma_L^1\sqrt{\lambda_L} + \lambda_H}{2\lambda_H\sqrt{\lambda_H}}\right)$$

$$= \frac{\lambda_L + \gamma_L^1\sqrt{\lambda_L} + \lambda_H}{2\lambda_H}\left(\frac{w\pi(\gamma_H)}{\gamma_H^2} - \frac{w\pi'(\gamma_H)}{\gamma_H} - (c+r)\right)$$

Replacing for $\frac{dC}{d\lambda_H}$ and simplifying we get:

$$\frac{dC}{d\lambda_H} = (1-q)\left[\frac{w}{\sqrt{\lambda_H}}\left(\frac{\pi'(\gamma_H^1)}{\gamma_H^1} - \frac{\pi'(\gamma_H^1)}{2}\right) + \left(\frac{w\pi(\gamma_H^1)}{(\gamma_H^1)^2} - \frac{w\pi'(\gamma_H^1}{\gamma_H^1}\right)\right] \tag{42}$$

On the other hand, for *R2* we have $\frac{\partial\gamma_H^2}{\partial\lambda_H} = -\frac{\lambda_L+\bar{\gamma}_L^2\sqrt{\lambda_L}}{2(\lambda_H)^{3/2}} - \frac{1}{2\sqrt{\lambda_H}}$.
Following the same reasoning we used for *R1* and replacing γ_H^1 with γ_H^2 for *R2* we get:

$$\frac{dC}{d\lambda_H} = (1-q)\left[\frac{w}{\sqrt{\lambda_H}}\left(\frac{\pi'(\gamma_H^2)}{\gamma_H^2} - \frac{\pi'(\gamma_H^2)}{2}\right) + \left(\frac{w\pi(\gamma_H^2)}{(\gamma_H^2)^2} - \frac{w\pi'(\gamma_H^2}{\gamma_H^2}\right)\right] \tag{43}$$

Note that $\gamma_H^1 < \gamma^* < \gamma_H^2$. If we show that $\frac{dC}{d\lambda_H}$ is decreasing in γ_H^1 this is equivalent to show that $\frac{d}{d\lambda_H}(C_1 - C_2) \geq 0$. The second part of $\frac{dC}{d\lambda_H}$ in (42) is decreasing in γ_H^1 because it is the slope of a convex function. Next we show that $\frac{\pi(\gamma)}{\gamma} - \frac{\pi'(\gamma)}{2}$ is decreasing in γ. Note that $\pi'(\gamma) = \frac{\pi(\gamma)^2}{\gamma} - \gamma\pi(\gamma) - \frac{\pi(\gamma)}{\gamma}$. Replacing we get $\frac{\pi(\gamma)}{\gamma} - \frac{\pi'(\gamma)}{2} = \frac{1}{2}\left[\frac{3\pi(\gamma)}{\gamma} + \gamma\pi(\gamma) - \frac{\pi(\gamma)^2}{\gamma}\right]$. Taking the derivative with respect to γ and simplifying we get:

$$-\left(\gamma^2 + \frac{6}{\gamma^2} + 3\right)\pi(\gamma) + \left(\frac{6}{\gamma^2} + 3\right)[\pi(\gamma)]^2 - \frac{2}{\gamma^2}[\pi(\gamma)]^3 \leq 0$$

So $\frac{d}{d\lambda_H}(C_1 - C_2) \geq 0$ is monotonous. Now while at $\lambda_H = \lambda_L$ we know that $C_1 < C_2$, at $\lambda_H = \lambda_L + \bar{\gamma}\sqrt{\lambda_L}$ we have $C_1 > C_2$. This implies that there exists a threshold T_1 such that

$$C_1 < C_2 \text{ for } \lambda_H < T_1$$
$$C_1 \geq C_2 \text{ for } \lambda_H \geq T_1 \quad \square$$

Proof of Corollary 1: **Characterization of *R3***. Suppose $\alpha_1 = 0$ and $\alpha_2 = 0$. From condition (24) we get: $q\frac{\partial C_{Cent}(\gamma)}{\partial\gamma}\big|_{\gamma=\gamma_L} = 0$. Due to the fact that $C_{Cent}(\gamma)$ is unimodal we conclude that $\frac{\partial C_{Cent}(\gamma)}{\partial\gamma}\big|_{\gamma=\gamma_L} = 0$ only holds for $\gamma_L = \gamma^*$, the centralized solution found in equation (5). A similar argument for condition (25) leads to $(1-q)\frac{\partial C_{Cent}(\gamma)}{\partial\gamma}\big|_{\gamma=\gamma_H} = 0$, which only holds for $\gamma_H = \gamma^*$. Then the solution is $\gamma_L = \gamma_H = \gamma^*$. The Headquarters propose $\gamma_L = \gamma_H = \gamma^*$ when $\tilde{\gamma}_L < \delta_L$. The explanation is as follows. Since $C_{Prog}(\gamma)$ is unimodal with minimum in $\bar{\gamma}$, and $\gamma^* < \bar{\gamma}$ it must be that $\bar{\gamma} < \tilde{\gamma}_L$. Hence $0 < \frac{\partial C_{Prog}(\gamma)}{\partial\gamma}\big|_{\gamma=\tilde{\gamma}}$. Therefore, $\tilde{\gamma} < \delta_L$ implies that $C_{Prog}(\gamma^*) = C_{Prog}(\tilde{\gamma}) < C_{Prog}(\delta_L)$. On the other hand, the high type Program would report its true needs since $\delta_H < \gamma_L = \gamma^*$ and $\frac{\partial C_{Prog}(\gamma)}{\partial\gamma}\big|_{\gamma=\gamma^*} < 0$ implying $C_{Prog}(\gamma^*) < C_{Prog}(\delta_H)$. $\square$

References

Altay, N., Green III, W.G. 2006. OR/MS research in disaster operations management *European Journal of Operational Research* 175, 475–493.

Barbarosoglu, G., L. Ozdamar, A. Cevik. 2002. An interactive approach for hierarchical analysis of helicpter logistics in disaster relief operations. *European Journal of Operational Research*, 140 (1), 118–133.

Barbarosoglu, G., Y. Arda. 2004. A two-stage stochastic Programming framework for transportation planning in disaster response. *Journal of the Operational Research Society.* 55 (1), 43–53.

Batta, R., N.R. Mannur. 1990. Covering-location models for emergency situations that require multiple response units. *Management Sci.* 36 (1), 16–23.

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

Bernstein, F., A. Federgruen. 2003. Pricing and replenishement strategies in a distribution system with competitive retailers. *Operations Research.* 51(3) 409–426.

Bookbinder, J.H., D.L. Martell. 1979. Time-dependent queueing approach to helicopter allocation for forest fire initial-attack *INFOR* 17(1) 58–70.

Borst, S. Mandelbaum, A. Reiman. 2004. Dimensioning large call centers. *Operations Research.* 52(1) 17–34.

Campbell A.M., D. Vanderbussche, W. Hermann. 2008. Routing for relief efforts. *Transportation Sci.* 42(2) 127–145.

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

Cachon, G.P., Tearwiesch, C. 2009. Matching supply with demand. An introduction to operations management. *McGraw-Hill.* Second Edition - International Edition.

Chang, M-S, Y-L. Tseng, J-W Cheng. 2007. A scenario planning approach for the flood emergency logistics preparation problem under uncertainty. *Transportation Research Part E.* 43 737-754

Chiu, Y.C., H. Zheng. 2007. Real-time mobilization decisions for multi-priority emergency response resources and evacuation groups: Model formulation and solution *Transportation Research Part E.* 43 710-736

Corbett, C.J., X. de Groote. 2000. A supplier's optimal quantity discount policy under asymmetric information. *Management Sci.* 46(3) 444–450.

Cova, T.J., J.P. Johnson. 2003. A network flow model for lane-based evacuation routing *Transportation Research Part A.* 37 579–604

Dasgupta, P., P. Hammon, E. Maskin. 1979. The Implementation of social choice rules: some general results on incentive compatibility *The Review of Economic Studies*, 46(2) 185–216

De Angelis, V. M. Mecoli, C. Nikoi, G. Storchi. 2007. Multiperiod integrated routing and scheduling of World Food Programme cargo planes in Angola. *Computers & Operations Research.* 34 1601–1615

Grassmann, W.K. 1988. Finding the right number of servers in real-world queuing systems. *Interfaces.* 18(2) 94–104.

Green, L.V.. 1984. A multiple dispatch queueing model of police patrol operations. *Management Sci.* 30(6) 653–664.

Green, L.V., P. Kolesar. 1984a. The feasibility of one-officer patrol in New York City. *Management Sci.* 30(8) 964–981.

Green, L.V., P. Kolesar. 1984b. A comparison of the multiple dispatch and M/M/c priority queueing models of police patrol. *Management Sci.* 30(6) 665–670.

Green, L.V., P. Kolesar. 1989. Testing the validity of a queueing model of a police patrol. *Management Sci.* 35(2) 127–148.

Green, L.V., P. Kolesar. 2004. Improving emergency responsiveness with management science. *Management Sci.* 50(8) 1001–1014.

Green, J., J.J. Laffont. 1977. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica.* 45(2) 427–438.

Halfin, S. W. Whitt. 1981. Heavy-traffic limits of queues with many exponential servers. *Operations Research.* 29(3) 567–588

Harris, M., R.M. Townsend. 1981. Resource Allocation Under Asymmetric Information *Econometrica.* 49(1) 33–64

Hasija, S., E.J. Pinker, R.A. Shumsky. 2005. Staffing and routing in a two-tier call centre. *Int. J. Operational Research.* 1(1) 8–29 .

Hasija, S., E.J. Pinker, R.A. Shumsky. 2008. Call center outsourcing contracts under information asymmetry. *Management Sci.* 54(4) 793–807.

Ignall, E., Carter, G., Rider, K. 1982. An algorithm for the unitial dispatch of fire companies. *Management Sci.* 28(4) 366-378

Jia, H., F. Ordonez, M. Dessouky. 2007. A modeling framework for facility location of medical services for large-scale emergencies *IIE Transactions.* 39 41-55

Kolesar, P., E. W. Blum. 1973. Square root laws for fire engine response distances. *Management Sci.* 19(2) 1368–1378.

Krishnan, H., R. Kapuscinski, D.A. Butz. 2007. Coordinating contracts for decentralized supply chains with retailer promotional effort. *Management Sci.* 50(1) 48–63.

Lindenberg, M. 2001. Are we at the cutting edge of the blunt blunt edge? Improving NGO organizatinal performance with private and public sector strategic management frameworks. *Nonprofit Management & Leadership.* 11(3) 247–270.

Mannell, J. 2010. Are the sectors compatible? International development work and lessons for a business–profit partnership framework. *Journal of Applied Social Psychology*. 40(5) 1106–1122.

Maskin, E., Riley, J. 1984. Monopoly with Incomplete Information. *The RAND Journal of Economics*. 15(2) 171–19

Myerson, R. B. 1979. Incentive Compatibility and the Bargaining Problem. *Econometrica*. 47(1) 61–73

Ozdamar, L., E. Ekinci, B. Kucukyazici. 2004. Emergency logistics planning in natural disasters *Annals of Operations Research*. 129 217–245

Pedraza Martinez, A.J., O. Stapleton, L.N. Van Wassenhove. 2010. Field vehicle fleet management in humanitarian operations: a case-based approach. *Journal of Operations Management* Jul 2011, 29 (5) 404–421

Pedraza Martinez, A.J., L.N. Van Wassenhove. 2010. Vehicle replacement in the International Committee of the Red Cross. *Production and Operations Management* (Forthcoming).

Pasternack, B.A. 1985. Optimal pricing and return policies for perishable commodities. *Marketing Sci*. 4(2) 166–176.

Regnier, E. 2008. Public evacuation decisions and hurricane track uncertainty. *Management Sci*. 54 (1) 16–28

Saadatseresht, M., A. Mansourian, M. Taleai. 2009. Evacuation planning using multiobjective evolutionary optimization approach. *European Journal of Operational Research*. 198 305–314

Salmeron, J., A. Apte. 2009. Stochastic optimization for natural disaster asset prepositioning. *Production and Operations Management*. 19(5) 561–574

Sheu, J.B. 2007. An emergency logistics distribution approach for quick response to urgent relief demand in disasters. *Transportation Research E*. 43 687 – 709

Singer, M., P. Donoso. 2008. Assessing an ambulance service with queuing theory. *Computers & Operations Research*. 35 2549–2560

Stepanov. A., J.M. Smith. 2009. Multi-objective evacuation routing in transportation networks. *European Journal of Operational Research*. 198 435–446

Su, X., S.A. Zenios. 2006. Recipient choice can address the efficiency-equity trade-off in kidney transplantation: a mechanism design model. *Management Sci*. 52(11) 1647–1660.

Taylor, T. 2002. Supply chain coordination under channel rebates with sales effort effects. *Management Sci*. 48(8) 992–1007.

Thomas, A., L. Kopczak. 2005. From logistics to supply chain management: the path forward in the humanitarian sector. *Fritz Institute White Paper*

Van Wassenhove, L. N. 2006. Humanitarian aid logistics: supply chain management in high gear. *Journal of the Operational Research Society* 57 475–489.

Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Sci.* 38(5) 708–723.

Yi, W., L. Ozdamar. 2007. A dynamic logistics coordination model for evacuation and support in disaster response activities. *European Journal of Operational Research.* 179(3) 1177–1193.