Healthcare Vehicle Maintenance Contracting in Developing Economies: The Role of Social Enterprise

Li Chen
Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853, li.chen@cornell.edu

Sang-Hyun Kim
Yale School of Management, Yale University, New Haven, CT 06520, sang.kim@yale.edu

Hau L. Lee
Graduate School of Business, Stanford University, Stanford, CA 94305, haulee@stanford.edu

Problem definition: Difficulties in healthcare delivery in developing economies arise from poor road infrastructure of rural communities, where the bulk of the population reside. Although motorcycles are an effective means for delivering healthcare products, governments in developing economies lack expertise in proper maintenances, resulting in frequent vehicle breakdowns. Riders for Health, a nonprofit social enterprise (SE), has developed specialized capabilities that enable significant enhancements in vehicle maintenance. Riders has engaged with the governments and provided its services using different contract schemes. We analyze a game-theoretic model of outsourced vehicle maintenance, shedding light on the costs and benefits of two different contracting approaches adopted by Riders. **Relevance:** This paper presents one of the first rigorous analyses of how SEs achieve their goals through innovations in operations. Our analysis highlights the relationship between a social mission objective and a service contract choice, contrasting it to the choice by a profit-maximizing organization traditionally assumed in the service contracting literature in Operations Management. **Methodology:** We conduct a game-theoretic analysis of a model that combines the elements of reliability theory and contract theory. **Results:** We find that the "total solution" approach of providing all aspects of vehicle maintenance, including fleet ownership, is a preferred choice for an SE like Riders who prioritizes improving health outcomes; by contrast, an organization that focuses on profit generation would find this approach less attractive. Our analysis also suggests that Riders would further benefit from the total solution approach if it manages a large fleet that consists of vehicles with strong resale values. **Managerial implications:** Our findings provide a theoretical support for Riders' recent move toward the total solution approach as it expands its service into wider rural areas in many countries. The insights obtained from our analysis offer actionable guidelines to Riders, and the model developed in this paper can be used to quantify the costs and benefits of implementing a contract.

Key words: Public Healthcare, Developing Economies, Social Responsibility, Reliability Theory
History: September 17, 2017

1. Introduction

Populations in developing economies face big challenges in their living standards. Besides poor economic conditions, healthcare provisions are often grossly inefficient or simply lacking, leading

to low life expectancy. Such is the case of sub-Saharan Africa, where the average life expectancy for men and women was reported to be about 53 years, compared to 78 years in the United States (World Health Organization 2013). The fraction of the population infected by diseases such as HIV/AIDS, malaria, measles, pneumonia, tuberculosis, and dehydrating diarrhea were much higher than that of the rest of the world (Lee et al. 2013). Maternal mortality rate exceeded 500 per 100,000 live births (Hogan et al. 2010). Despite the enormous needs, health provisions have been dismal, contributed by the fact that the majority of the population reside in rural areas where logistics infrastructures are inadequate or nonexistent.

The United Nations (2014) reports that 60 percent of Africans lived in rural communities, defined to be the ones without adequate infrastructure (e.g., paved roads, electricity, piped water or sewers), education, or health services. Yet, 53 percent of the roads in Africa were unpaved (African Development Bank 2014). As a result, many communities are accessible only by single-lane sand or dirt paths. People requiring medical care are often forced to walk or be carried in a wheelbarrow pushed by a relative in order to get to the nearest clinic, as doctors and nurses are unable to visit such communities on a regular basis.

Well-maintained vehicles and motorcycles have therefore become a crucial missing link in the healthcare delivery supply chain. In fact, given the poor road infrastructure, motorcycles are often the most effective means of transport that allow health workers to visit sick patients and deliver healthcare products such as medicines, test samples and results, education kits, condoms, and malaria nets. Running a cost-effective operation of a vehicle fleet consisting of well-maintained and reliable motorcycles requires specialized expertise and management control. Unfortunately, health ministries of developing economies do not always possess these skills. Preventive maintenance or repair were often not done properly, spare parts for repair were missing, and old vehicles were not replaced at the right time. This is why a *social enterprise* (SE) with such capabilities can make a difference, as it supplements the health ministries' efforts in this endeavor and offers improvements. However, because the goals of an SE and a government agency are not necessarily aligned, the effectiveness of such a joint effort depends on the structure of the contract established between the two parties. This is the subject of our study.

Our work is motivated by an SE named Riders for Health (Riders), a nonprofit organization set up to provide African health ministries with cost-effective operations of consistently reliable vehicle fleets used for healthcare deliveries. Before Riders' involvement, the vehicles were managed by local governments. Vehicles suffered from high operating costs and low availability due to a variety of inefficiencies, such as lack of training and expertise, inconsistent maintenance procedures, and shortages of service parts. With poor maintenance, vehicles needed repairs often, or they simply

stopped working. With poor repair skills and shortages of service parts, it took excessive amount of time to repair or replace vehicles, leading to low vehicle availability.

Riders addressed these issues in multiple ways. The backgrounds of Riders' cofounders were instrumental, as they were a group of professional motorcycle racers with deep knowledge of the operating characteristics of motorcycles. They were able to transform such knowledge into maintenance and repair procedures, which were then passed onto technicians. They trained health workers to perform simple self-maintenance procedures. They also set up an elaborate hub-and-spoke service parts distribution network to ensure that service parts are supplied and replenished quickly when needed. Armed with these innovations, the organization worked with health ministries of the Zimbabwean, Gambian, and Zambian governments and produced very positive results: increased vehicle availability, extended vehicle life, improved health access and interventions, and enhanced cost effectiveness (Tayan 2007, Mehta et al. 2016). As of 2014, Riders operated in seven countries serving population of 21.5 million with 470 staff members and 1,700 vehicles, generating an income of £3.5 million through its programs (Riders for Health 2014).

Despite the success, Riders faced important decisions in its work that raised interesting research questions. When Riders engaged with local governments, it employed two different contracting approaches. An approach called Transport Resource Management, which we refer to as *Repair Only Contract* (ROC) in this paper, was the basis of standard outsourcing arrangements used in most of earlier implementations in Zimbabwe and Gambia. Under ROC, Riders is responsible for maintenance and repair of vehicles (and other maintenance-related tasks such as worker training), while the government controls vehicle replacements and purchases. A newer approach called Transport Asset Management, which we refer to as *Repair and Replacement Contract* (RRC) in this paper, was more prominently used in a later implementation in Gambia and Zambia, and it became Riders' preferred choice. Under RRC, in addition to the same maintenance and repair tasks as under ROC, Riders is also responsible for vehicle replacements and purchases. The running costs of this added responsibility can be spread over time and be reimbursed by health ministries in multiple payment installments. The main benefit of RRC for Riders is that vehicles can be replaced at time intervals deemed optimal by Riders. Another benefit is that RRC allows Riders to standardize vehicle models, enabling streamlined fleet management and reduction of complexity in the overall supply chain. Both ROC and RRC are performance-based contracts, as Riders are compensated based on the actual mileage used; no payment is made for the duration of vehicle downtime. Hence, successful implementations of the two contracting approaches depend on the level of vehicle availability that Riders can provide through its share of fleet operation.

Why does Riders prefer RRC, which actually involves added work and challenges? It is not easy to secure necessary funding for purchasing a fleet and pay for the setup, and doing so often requires

a long-term contractual commitment from the governments, who might be reluctant to agree to such an arrangement. For example, financing for the RRC fleet in Gambia involved a pioneering arrangement among Riders, Africa-based GT Bank, US-based Skoll Foundation, and the Gambian Ministry of Health. (This is in contrast to ROC, under which the government is free to switch on and off the services provided by Riders.) Clearly, adoption of RRC requires a careful assessment of tradeoffs that it presents. In this paper, we explore the research questions that arose from Riders' experience. How do the two contracting approaches compare in improving health outcomes and achieving cost effectiveness? Which features of vehicle maintenance cause the differences in relative performances of the two approaches? Under what conditions does an SE prefer RRC to ROC? Does the contract choice depend on whether the SE is for-profit or nonprofit?

To answer these questions, we conduct a game-theoretic analysis that, for the first time in the literature, combines an SE's contracting decisions with the constructs of optimal machine maintenance models found in the reliability theory literature. Our model features an SE and a government agency that enter into a contractual arrangement under which they divide responsibilities for managing a fleet of vehicles used for health product deliveries. It is the SE who possess the expertise in efficient fleet management and proposes a contract to the government. In making their decisions, the two parties each weigh health outcomes against their costs/profits.

Analyzing this model, we find that RRC is a preferred choice for an SE like Riders who prioritizes improving health outcomes over profitability; by contrast, an SE who focuses on profit generation would find ROC more attractive. Our analysis also suggests that Riders will further benefit from adopting RRC when it manages a large fleet of vehicles that have strong resale values. Together, these findings provide a theoretical support for Riders' recent move toward RRC adoption as it expands its service into wider rural areas in many countries.

Although the role of SEs has been studied from different perspectives in the literatures of law, economics, and organizational studies, few have examined how SEs achieve their goals through innovations in operations. This paper represents one of the first such attempts, as we explicitly link operational details of vehicle maintenance with improved health outcomes in sub-Saharan countries based on Riders' experience. Our analysis reveals managerial insights and actionable guidelines, and furthermore, the model developed in this paper can be used as a means to quantify the costs and benefits of implementing a contract.

2. Related Literature

Because of its unique value proposition, an analysis of Riders' problem requires a combination of knowledge found in different areas of research, including reliability theory, service contracting in Operations Management (OM), and the growing literature on social enterprises.

Motivated by emergence of SEs in recent years, researchers of organizations and entrepreneurship have offered many perspectives on the role of SEs (e.g., Austin et al. 2006, Mair and Marti 2006, Doherty et al. 2014). Among them is the *hybrid organizations* view, which we adopt in our paper: "social enterprises pursue the dual missions of achieving both financial sustainability and social purpose" (Doherty et al. 2014). Under this definition, an SE may be a nonprofit or for-profit entity (Katz and Page 2010). Riders is an example of nonprofit SE with a singular focus on improving public health in sub-Saharan Africa, but many other SEs strive to balance their social objectives with profitability (see Mair and Marti (2006), Wilson and Post (2013), Lee and Tang (2017) for examples). In our paper we consider a spectrum of SE goals, which allows us to better understand the relationship between Riders' social welfare maximization objective and its decisions on service provisions.

The role of SE has been recognized in the OM literature in a broad context of socially-responsible operations (see Besiou and van Wassenhove (2015) and Sodhi and Tang (2017) for overviews). Sodhi and Tang (2011) emphasize SEs' role as "enablers," who facilitate flows of goods, services, and information in supply chains. Lee and Tang (2017) discuss examples of SEs that offer innovative operational solutions. While the importance of capabilities brought by SEs is recognized, few studies have examined their operational challenges using rigorous analytical frameworks. Exceptions include Hu et al. (2016), who examine an agricultural procurement contracting problem for an SE in developing countries, and Calmon et al. (2017), who study an SE's innovative use of distribution channels in India. Also related are Chen et al. (2012) and Uppari et al. (2017), who study operational impacts of SEs' involvements in agriculture and energy markets in developing economies.

Our paper sets itself apart from these studies in that we pay close attention to an SE's operational process that directly impacts social outcomes. Specifically, motivated by Riders' focus on vehicle maintenance operations, we develop a model that customizes the classical machine repair and replacement models found in the reliability theory literature (e.g., Barlow and Proschan 1996, Nakagawa 2005). In particular, we borrow ideas from the *age replacement models* by Barlow and Hunter (1960) and Beichelt (1976, 2006), extending them to capture both fatal and non-fatal vehicle failures that disrupt healthcare deliveries, an important consideration for Riders. Although we maintain key model features found in the reliability theory literature, our model is not a straightforward extension of them because these features are combined with the elements of performance-based contracting in a decentralized supply chain (Kim et al. 2007, Guajardo et al. 2012, Jain et al. 2013). This modeling approach is natural since Riders' contractual payment for vehicle maintenance is based on realized vehicle uptime or usage. To the best of our knowledge, our paper

represents the first attempt at combining the features of age replacement models with a game-theoretic analysis of contracting. Furthermore, we are not aware of other studies that consider the social welfare objective of a decision-maker who provides maintenance services. Therefore, we contribute to the OM literature on service contracting by examining a non-standard setting in which decisions are based on neither profit maximization or cost minimization, and by demonstrating how such a departure from the traditional assumption influences a contract choice.

Healthcare deliveries in developing economies through innovative logistics systems, the subject of our paper, have been discussed in the literature but on the topics that do not overlap with ours; they include centralized vs. decentralized supply chain control (Bossert et al. 2007), operational skill set development (Matome et al. 2008), optimized capacity allocation (Deo et al. 2015), and inventory control (Leung et al. 2016). The papers that come closest to ours are those by Pedraza-Martinez and van Wassenhove (PW; 2012, 2013), who study vehicle fleet management for humanitarian logistics. Based on cases and field studies, PW (2012) provide insights into the persistence of aging fleets, low fleet standardization, and service delays, many of the same issues that Riders has encountered and corrected. PW (2013) show empirically that the standard vehicle replacement policy recommended by the International Committee of Red Cross is suboptimal, and suggest more frequent replacements. This conclusion, however, is based on the assumption of a centralized fleet cost minimization objective, which contrasts with the decentralized social welfare maximization objective that we consider in our paper. Moreover, our focus is on comparative evaluation of two contracting approaches adopted by an SE, a subject PW (2013) do not consider.

Finally, the case of Riders has been studied by other researchers. They include McCoy and Lee (2014) who focus on the equity issue of allocating vehicles to different villages, and Mehta et al. (2016) who empirically evaluate Riders' contribution to health worker productivity. Our paper complement these works by analyzing Riders' contractual vehicle maintenance problem through the lens of reliability theory and contract theory.

3. Model

We consider a contractual relationship between two risk-neutral entities: a social enterprise ("SE," such as Riders) and a government agency in charge of its country's public healthcare policies ("government," such as the Ministry of Health in Zambia). The government funds and operates a healthcare delivery program ("program") aimed at distributing healthcare products and services to the population in remote regions. The government, who used to run the program on its own, considers running the program in collaboration with SE, who brings an expertise in vehicle maintenance and management such as an understanding of industry best-practices or an exclusive access

to a service parts distribution network. A contract that defines the program's division of responsibilities and payment terms is set up between the two parties. For expositional convenience, we use the pronoun "she" to refer to SE and "he" to refer to the government.

Under the program, health workers are deployed to remote regions by means of light transportation vehicles such as motorcycles ("vehicles"). A key driver of the program's success is vehicle uptime, as health workers cannot be deployed if vehicles are inoperative. This is especially important because vehicles are subject to frequent breakdowns due to harsh road conditions in remote regions. Vehicle uptime can be managed through careful maintenance planning, which SE has an expertise on. A fleet of vehicles is owned and maintained by either the government or SE, depending on a contractual arrangement.

3.1 Vehicle Repairs and Replacements

In order to highlight the key tradeoffs, in the main part of the paper we focus on a simplified setting where the size of vehicle fleet is normalized to one, i.e., at any given point in time, the fleet consists of a single vehicle. (Note that this normalization does not mean the same vehicle remains in the fleet indefinitely, because an existing vehicle can be replaced by a new vehicle.) We examine a multi-vehicle extension in §5.1 and discuss the implications of varying fleet sizes.

At a given point in time, a vehicle in the fleet is either available for service ("up") or unavailable after experiencing a random failure and being disabled ("down"). The vehicle can create value only if it is available and utilized. Reliability of the vehicle degrades over time as it ages, becoming more prone to random breakdowns ("failures") as it gains mileage. It is important to distinguish vehicle age from calendar time; since the vehicle ages only when it is used, aging is suspended while it is down. Since failures occur only when the vehicle is used, the time unit on which the failure arrival process is based is vehicle age, not calendar time.

A vehicle experiences two types of failures: *minor failure* and *major failure*. A minor failure disables the vehicle temporarily but is non-fatal, as the vehicle will be fixed and restored to service (e.g., a flat tire that can be replaced by a spare tire). On the other hand, a major failure is fatal and it disables the vehicle permanently (e.g., failure of an engine that cannot be swapped). We assume that the rates at which these two types of failures occur increase with age. The overall failure rate is specified as $\lambda h(t)$, where λ is a scaling factor and $h(t)$ is an increasing function that starts from $h(0)=0$ and diverges to $h(\infty)=\infty$. Let $H(t) \equiv \int_0^t h(z)\,dz$. We adopt the convention that failures occur following a nonhomogeneous Poisson process that gets split into major and minor types with probabilities $p/(p+q)$ and $q/(p+q)$, where p and q satisfy $p+q=\lambda$ (Beichelt 2006, pp. 138-139). Hence, major and minor failures occur at rates $ph(t)$ and $qh(t)$, respectively.

Once a failure occurs, it is addressed by either a *repair* or a *replacement*. A repair is performed on a vehicle disabled by a minor failure, and it restores the vehicle back to working condition. If,

on the other hand, a repair cannot be done because the vehicle suffers a major failure, the disabled vehicle is scrapped and replaced by a new unit. To be precise, a vehicle replacement triggered by a major failure (vehicle's "premature death") is categorized as an *unscheduled replacement* because such a replacement cannot be planned in advance due to the random nature of failure events. By contrast, a *scheduled replacement* refers to a planned replacement of a vehicle that reaches its end-of-life (vehicle's "retirement").

A repair and a replacement have different impacts on the age of a vehicle in the fleet. Whereas completion of a replacement resets the age of a vehicle in the fleet to zero as a new vehicle is brought in, completion of a repair does not change the age; the age at repair completion remains the same as that at the repair start because the vehicle does not gain mileage during repairs. (Such repairs are called *minimal repairs*; see Nakagawa (2005).) We assume that a fixed cost k is incurred each time a repair is performed, while a fixed cost K is incurred whenever a vehicle is replaced, regardless of whether it is unscheduled or scheduled. With the latter assumption, we implicitly assume that the majority of what constitutes K is the cost of acquiring a new vehicle. In §5.2 we consider a modification of this assumption that takes into account the residual value of a vehicle.

A scheduled replacement is done whenever the age of a vehicle currently in the fleet reaches τ time units, provided that no major failure has occurred by then. We refer to τ as vehicle *retirement age*. (Recall that the age τ does not include vehicle downtimes due to repairs.) If a major failure occurs before τ is reached, then an unscheduled replacement is triggered and the vehicle age is reset to zero. We assume that a scheduled replacement is completed instantaneously because τ is known in advance and therefore a replacement vehicle can be procured early. By contrast, it takes positive lead times to complete a repair and an unscheduled replacement because they are prompted by unpredictable failures. A repair lead time depends on such factors as labor capacity or service parts inventory. A replacement lead time depends on accessibility to a vehicle provider, complexity of customs clearance process if the vehicle is to be imported, and vehicle registration and documentation, etc. A vehicle is down during these lead times. We use the notations l and L to denote expected lead times for a repair and for an unscheduled replacement, respectively.

The retirement age τ is the decision variable that we focus on in this paper (the same focus is commonly found in the reliability theory literature). It is a key determinant of long-run success of the program, and its choice reflects the tradeoffs among key performance outcomes including vehicle availability, repair cost, and replacement cost. All other variables are assumed to be exogenously given, except in §5.3 where we consider varying the values of l and λ (hence p and q) to study the impacts of failure prevention and repair lead time reduction efforts.

3.2 Performance Measures

All vehicle performance measures are defined in long-run time averages, evaluated by applying the renewal-reward theorem. A natural renewal cycle is the interval between two successive vehicle replacements, since the age of a vehicle in the fleet is reset to zero whenever a replacement is complete. The following performance measures are used in our analysis: 1) long-run average vehicle *availability*, denoted by $A(\tau)$; 2) long-run average vehicle *repair cost*, denoted by $C(\tau)$; 3) long-run average vehicle *replacement cost* due to both scheduled and unscheduled replacements, denoted by $R(\tau)$. Availability $A(\tau)$ measures the vehicle uptime within each cycle. (Since one-to-one correspondence exists between availability and uptime, we use the two terms interchangeably in most of our discussions.) Repair cost $C(\tau)$ is equal to fixed repair cost k times the expected number of repairs performed in a cycle. Replacement cost $R(\tau)$ is equal to fixed replacement cost K times the number of replacements in a cycle, which is equal to one since each cycle ends with a replacement.

To evaluate these performance measures, we first define the following notations used throughout the analysis. Let Y be the random variable of vehicle age at the time of first major failure conditional on no vehicle retirement ($\tau = \infty$). Let $F(t) \equiv \Pr(Y < t)$ and define $\overline{F}(t) \equiv 1 - F(t)$ and $M(\tau) \equiv \int_0^\tau \overline{F}(t)\,dt$. Because major failures occur at rate $ph(t)$, the following relationships hold: $\overline{F}'(t)/\overline{F}(t) = -ph(t)$ and $\overline{F}(t) = \exp(-pH(t))$ where $H(t) = \int_0^t h(z)\,dz$. The mean of Y is denoted by $\mu \equiv E[Y] = M(\infty)$. In addition, define $S(t) \equiv h(t)M(t) - (1/p)F(t)$, which appears frequently in our analysis (its properties are found in Lemma A1 in the Online Appendix). Note that $F(t)$, $M(t)$, and $S(t)$ all depend on the failure rate scaling factor p, unlike $h(t)$ and $H(t)$.

Based on these definitions, the performance measures are evaluated as

$$A(\tau) = M(\tau)/T(\tau), \quad C(\tau) = (kq/p)F(\tau)/T(\tau), \quad R(\tau) = K/T(\tau). \tag{1}$$

where

$$T(\tau) \equiv M(\tau) + (L + lq/p)F(\tau) \tag{2}$$

is the expected length of each renewal cycle (see Lemma A2 in the Online Appendix).

The long-run average health benefit $W(\tau)$, which we simply call "welfare" throughout the paper, is determined through the choice of τ via vehicle availability $A(\tau)$. It is availability that directly impacts welfare $W(\tau)$ because higher vehicle uptime allows for greater vehicle usage, which in turn enables more healthcare deliveries. (Although in general, the welfare of health benefit is a function of many aspects of service, in this paper we focus on patient access.) The positive relationship between availability and welfare exists because the healthcare needs in vast rural areas far exceed what can be supplied by deliveries; since the needs will not be fully satisfied even if all vehicles in the fleet are 100% utilized, higher vehicle uptime will result in more patient access and better

health outcomes. Such a relationship between $A(\tau)$ and $W(\tau)$ can be captured by a mapping $W(\tau) = V(A(\tau))$, where $V(\cdot)$ is an increasing function.

The exact shape of $V(\cdot)$ is difficult to specify explicitly, since it depends on many factors that are beyond the control of the management and maintenance of vehicles. For example, in Riders' implementation in Gambia, the health outcomes—including life expectancy, infant mortality, and reduction of disease incidences—were influenced by such local factors as efficient handling of test kits, weather condition, staffing, and availability of drugs (Leung et al. 2016). Recognizing this difficulty, the Gates Foundation, which commissioned the Stanford research project on Riders' program in Zambia (Mehta et al. 2016), suggested focusing on vehicle uptime as the primary metric for evaluating the program's effectiveness. While the Stanford research team did collect data on some measures of health outcomes affected by vehicle uptime and effectiveness of health workers (see Table 1 in §3.4), no attempt was made to infer the functional form of $V(\cdot)$.

In this paper we circumvent this modeling difficulty by assuming that the function $V(\cdot)$ is exogenously given, and furthermore, in the main part of our analysis we assume $V(\cdot)$ to be a linear function $V(A(\tau)) = vA(\tau)$ with a constant coefficient v. (Since the mapping between availability and welfare is uncertain in reality, $V(\cdot)$, and hence v, is to be interpreted as the expectation shared by SE and the government in a rational expectation sense.) Although the linearity assumption is restrictive, as we demonstrate, the main tradeoffs are captured with this simplification, which allows for analytical tractability. A more general specification would be a concave increasing function, which we consider as an extension; see §5.1 for more discussions.

3.3 Contracting

Reflecting Riders' experiences, we assume that the government already has his version of the program running, performing all fleet management tasks by himself including vehicle repairs and replacements. Since the government has an option to continue with this "do-it-alone" approach at the time of SE's contract offer, we call this the government's *status quo option*. The contract offered by SE specifies the division of tasks and payment terms, and the government accepts the contract if it is preferable to maintaining the status quo. We assume that SE's program operation is funded by the government only. Although the government funds the program operations, it is SE who provides an expertise to improve the program. Hence, we assume that a contract proposal is made by SE who offers the terms that the government finds acceptable.

We consider two contracting scenarios that differ by contracting parties' responsibilities.

1. Repair Only Contract (**ROC**): The government owns the fleet and replaces vehicles by determining τ. Vehicle repairs are outsourced to SE. The government compensates SE based on realized vehicle uptime.

2. Repair and Replacement Contract (**RRC**): Fleet ownership is transferred to SE. SE replaces vehicles by determining τ and performs vehicle repairs. The government compensates SE based on realized vehicle uptime.

As we noted in §1, ROC and RRC capture the essence of Transport Resource Management and Transport Asset Management contracts employed by Riders. Under ROC, SE is responsible for vehicle maintenance including repairs, but not vehicle replacements. Under RRC, by contrast, SE is responsible for entire fleet operations including vehicle repairs and replacements. The latter arrangement requires a transfer of fleet ownership, because only the rightful owner of assets can make decisions on replacing them periodically and incur associated costs; decisions on asset acquisitions cannot be outsourced without ceding control of them. Due mainly to financial hurdles associated with asset transfers under RRC, in practice ROC has been more widely used by Riders, especially in the earlier implementations in Zimbabwe and Gambia (Tayan 2007).

Under both ROC and RRC, SE is compensated based on realized vehicle uptime. (In practice, the compensation can also be based on actual usage of the vehicle. Because usage and uptime are directly correlated i.e., higher uptime leads to more usage, for simplicity, we assume in the main part of our analysis that the compensation is based on uptime.) Therefore, all contracts are performance-based: SE's compensation depends on realized performance outcome rather than the amount of resources consumed. Such performance-based contracts are gaining popularity in a number of industries for managing outsourced maintenance operations (Kim et al. 2007, Guajardo et al. 2012, Jain et al. 2013), and Riders has adopted the same practice. To be consistent with observed contracts, we only consider a linear contract that stipulates the government to transfer a constant dollar amount r for each unit of realized vehicle uptime.

We assume that SE incurs constant *overhead costs* per unit time, equal to ψ_1 under ROC and ψ_2 under RRC. These represent the continuing expenditures needed to maintain SE's operational capabilities, such as managing a spare parts distribution network. We assume $\psi_2 > \psi_1$ to reflect additional costs associated with vehicle ownership under RRC, such as amortized loan payments to financial institutions that aided vehicle acquisitions for Riders. For example, Riders' RRC implementation in Gambia was made possible by loan guarantees from the Africa-based GT Bank and the US-based Skoll Foundation (Lee et al. 2013). Similar arrangements were required in other countries, such as financial support from the Gates Foundation for RRC implementation in Zambia (Mehta et al. 2015).

3.4 Objectives

Both the government and SE make their decisions to balance the costs and benefits of improving the welfare through the program. To capture this tradeoff in a general framework, we define the

objectives of the government and SE as an weighted average of welfare and their costs/profit. A similar approach is found in the economics literature on SEs, including Szymańska and Jegers (2016) and Besley and Ghatak (2017). We assume that the government and SE each assign relative weights α and γ, respectively, to welfare W (with $0<\alpha<1$ and $0<\gamma<1$). Hence, the government's objective is to maximize $\alpha W-(1-\alpha)B$, where B denotes the government's expected total costs of running his share of the program. On the other hand, SE's objective is to maximize $\gamma W+(1-\gamma)\Pi$, where Π denotes SE's expected surplus, i.e., the difference between the payments she receives from the government and the costs of running her share of the program. (The exact forms of B and Π depend on a contracting scenario, and they will be specified later in §4.) By varying the values of α and γ, we cover a wide range of preferences by the government and SE. For example, $\gamma\to 0$ represents the case where SE runs the program purely for the profit purpose. As γ increases, SE puts more emphasis on improving welfare, with $\gamma\to 1$ mimicking a nonprofit entity. For convenience, we use the terms "profit-oriented SE" and "welfare-oriented SE" to refer to the cases with small γ and large γ, respectively. Similarly, "cost-oriented government" and "welfare-oriented government" correspond to small α and large α.

The government and SE face different constraints in their decisions. We assume that the government makes his decisions subject to a *budget constraint* $B\le b_0$ and a *welfare constraint* $W\ge w_0$. The constants b_0 and w_0, respectively, represent the government's total costs and the level of welfare achieved with his "do-it-alone" status quo. These constraints capture common concerns raised by government agencies, namely that increasing a budget allocated to a program beyond the existing level is often difficult and that outsourcing a government-sanctioned program can be approved only if such an arrangement promises to bring added values. If SE's contract offer violates any of the two constraints, then the government does not accept the offer. As such, satisfying the two constraints ensures the government's participation. (Participation is guaranteed since the two constraints $B\le b_0$ and $W\ge w_0$ together imply that the government increases the value of his objective by switching from the status quo to contracting with SE; $\alpha W-(1-\alpha)B\ge \alpha w_0-(1-\alpha)b_0$.)

Since the program requires the government's participation, SE sets the terms of ROC and RRC to ensure that the government's budget and welfare constraints are satisfied. In addition, the contract terms should satisfy SE's *surplus constraint* $\Pi\ge 0$, i.e., SE receives sufficient funding from the government that leaves her with a nonnegative surplus by running her share of fleet operations.

Finally, throughout the paper we impose the following technical conditions on parameter values that rule out a number of exceptional cases and allow us to focus only on nontrivial cases.

Assumption 1 *(i)* $\frac{kq}{Lp+lq}>\frac{K}{\mu}$*; (ii)* $b_0>\psi_2+\frac{Kp+kq}{(\mu+L)p+lq}$*; (iii)* $w_0<\frac{v\mu p}{(\mu+L)p+lq}$.

Table 1 Performance Comparisons (Source: Mehta et al. 2016)

	Riders RRC Test Regions	Government Control Regions
Number of motorcycles	70	47
Uptime days per week	5.5	1.3
% of Vehicles died at end of study period	0%	19%
No. of outreach trips per week	1.0	0.1
No. of patient visits per week	34.3	8.6
No. of immunizations per week	10.3	0.08
No. of child growth monitoring per week	13.7	10.3
No. of health education sessions per week	23.0	2.4

None of the conditions in Assumption 1 are very restrictive. The first two conditions are satisfied if the expected vehicle life μ is sufficiently long and the RRC overhead cost ψ_2 is in reasonable range below the government budget b_0. The last condition implies that the welfare constraint $W \geq w_0$ does not bind in equilibrium, which is consistent with Riders' experience. This is illustrated in Table 1, which summarizes performance outcomes of Riders' operations in Zambia reported by Mehta et al. (2016). As the table shows, Riders' involvement in the healthcare delivery program resulted in significant jumps in vehicle availability and various health outcome measures. (Although the table only shows the data for RRC, similar improvements have been reported under other ROC programs.) Finally, in addition to Assumption 1, we restrict attention only to a moderate range of values for b_0, which ensures that the problems are well-behaved.

4. Equilibria Under Two Contracting Approaches

4.1 Repair Only Contract (ROC)

Under ROC, the government has ownership of the fleet and is responsible for periodically replacing an aging vehicle in the fleet by choosing vehicle retirement age τ. While the government incurs replacement costs $R(\tau)$, it is SE who incurs repair costs $C(\tau)$ since vehicle repairs are outsourced to SE. In addition, the government pays SE by the rate r per unit of vehicle uptime. Thus, the government's total cost is $B(r,\tau) = rA(\tau) + R(\tau)$ and SE's surplus is $\Pi(r,\tau) = rA(\tau) - C(\tau) - \psi_1$, where ψ_1 is SE's overhead cost of running her share of ROC.

At time zero, SE sets the uptime fee r and offers it to the government, anticipating the government's optimal response $\tau = \tau_G(r)$ that satisfies the latter's budget and welfare constraints as well as SE's surplus constraint. The government then accepts the offer in equilibrium. In sum, SE solves the following optimization problem under ROC:

$$
\begin{aligned}
(\mathcal{P}_1) \quad \max_{r \geq 0} \quad & \gamma W(\tau_G(r)) + (1-\gamma)\Pi(r, \tau_G(r)) \\
\text{s.t.} \quad & \Pi(r, \tau_G(r)) \geq 0, && (3) \\
& \tau_G(r) \in \arg\max \{\alpha W(\tau) - (1-\alpha) B(r,\tau) \text{ s.t. } B(r,\tau) \leq b_0 \text{ and } W(\tau) \geq w_0\}. && (4)
\end{aligned}
$$

The solution to the government's subproblem (4) is found in Lemma A6 in the Online Appendix. We note the following properties of the government's optimal response.

Lemma 1 *Given* $r < \frac{\alpha}{1-\alpha}v - \frac{K}{\mu}$, *the government sets* $\tau = \tau_G(r) < \infty$ *that satisfies* $\tau_G'(r) > 0$; *otherwise, he sets* $\tau = \infty$.

The lemma indicates that, unless the government finds it optimal to never perform a scheduled replacement (set $\tau = \infty$), he responds to a higher fee r by holding onto an existing vehicle longer in order to save vehicle replacement costs. As a result, an attempt by SE to charge a higher fee may backfire, as the government's replacement delay causes SE to spend more time and money on repairing an aging vehicle that is prone to breakdowns. This illustrates the fundamental inefficiency that exists under ROC, a phenomenon that has been reported by Riders (Tayan 2007).

We now specify the ROC equilibrium, denoted by the superscript †. For ease of analysis, it is convenient to convert the unconstrained version of SE's problem $(\mathcal{P}_1)$ of choosing the fee r into an equivalent problem of optimizing over τ, by inverting the government's optimal response $\tau_G(r)$ using its monotonicity proved in Lemma 1. Define

$$\rho(\tau) \equiv \frac{\alpha}{1-\alpha}v - \frac{K}{S(\tau)}\left(h(\tau) + \frac{1}{Lp+lq}\right), \tag{5}$$

which represents the value of r that induces the government's optimal choice of τ. With this inversion, $\Pi(r, \tau_G(r))$ and $B(r, \tau_G(r))$ in $(\mathcal{P}_1)$ are rewritten as

$$\widetilde{\Pi}(\tau) \equiv \Pi(\rho(\tau), \tau) = \rho(\tau) A(\tau) - C(\tau) - \psi_1 \quad \text{and} \quad \widetilde{B}(\tau) \equiv B(\rho(\tau), \tau) = \rho(\tau) A(\tau) + R(\tau). \tag{6}$$

In general, the functions $\widetilde{\Pi}(\tau)$ and $\widetilde{B}(\tau)$ are not well-behaved; there are instances where they are neither monotone or unimodal. We observe numerically, however, that such instances rarely occur under parameter restrictions given in Assumption 1 and other assumptions stated in §3. To focus on the most salient features of the equilibrium, we assume a mild regularity condition that $\widetilde{\Pi}(\tau)$ and $\widetilde{B}(\tau) - b_0$ each cross zero at most once, a property readily satisfied under the aforementioned assumptions. The ROC equilibrium is specified as follows.

Proposition 1 (Equilibrium under ROC) *Let* $\overline{\alpha} \equiv \left(1 + \frac{v}{b_0}\frac{\mu p}{(\mu+L)p+lq}\right)^{-1}$ *and assume that* $\widetilde{\Pi}(\tau)$ *and* $\widetilde{B}(\tau) - b_0$ *each cross zero at most once at* τ_1^s *and* τ_1^b, *respectively. In addition, let (i)* $\underline{\tau}_1 \equiv \min\{\tau : C(\tau) + R(\tau) = b_0 - \psi_1\}$, *(ii)* $\widetilde{\tau}_1 \equiv \arg\max_{\tau \geq \max\{\tau_1^s, \underline{\tau}_1\}}\{\gamma W(\tau) + (1-\gamma)\widetilde{\Pi}(\tau)\}$, *and (iii)* $\widehat{\tau}(\gamma)$ *be the unique solution to* $Q_S(\tau) = 0$ *for given* γ *where* $Q_S(\tau) \equiv \left(\frac{\gamma}{1-\gamma}v(Lp+lq) + kq\right)S(\tau) - K(Lp+lq)h(\tau) - K$. *The ROC equilibrium occurs at* $\tau = \tau^\dagger$ *as follows.*

(a) If $\alpha \leq \overline{\alpha}$ *then* $\tau^\dagger = \infty$, *at which the budget constraint binds but the surplus constraint does not.*

(b) If $\alpha > \overline{\alpha}$ *and* $\widetilde{B}(\widetilde{\tau}_1) < b_0$ *then* $\tau^\dagger = \widetilde{\tau}_1$, *at which the budget constraint does not bind.*

(c) If $\alpha > \overline{\alpha}$ and $\widetilde{B}(\widetilde{\tau}_1) \geq b_0$ then $\tau^\dagger = \max\{\tau_1^b, \underline{\tau}_1, \widehat{\tau}(\gamma)\}$, at which the budget constraint binds.

Proposition 1 shows that distinct equilibrium outcomes are possible under ROC, depending especially on the values of α and b_0. (For fixed α, parts (a)-(c) of the proposition are sequenced in decreasing values of b_0.) Part (a) illustrates a potential pitfall of employing ROC; if the government is cost-conscious despite a large budget (small α and large b_0), he never conducts scheduled vehicle replacements. As a result the fleet consists of an old, failure-prone vehicle on average, negatively impacting the welfare and SE's repair costs. Moreover, because the government's decision is made independent of r in this case, SE is able to charge a large fee and saturate the government's budget.

Part (b) is a case where the government is more welfare-oriented and has a sufficient budget (medium α and medium b_0). In this case, the government ends up spending less than his budget. The government enjoys such a cost saving, but it represents an efficiency loss for SE because a non-biding budget constraint means SE is unable to bring in the maximum funding from the government via contracting. This inefficiency arises due to the aforementioned tradeoff, namely that charging a higher fee r prompts the government to delay vehicle replacements, which in turn increases SE's repair costs. Finally, part (c) is a case of a welfare-oriented government facing a tight budget (large α and small b_0). In this case, the government's willingness to improve welfare through frequent vehicle replacements leads him to exhaust his budget, depriving him of the freedom to choose τ. This allows SE to set the fee r such that the her desired level of τ is chosen in equilibrium, thus potentially eliminating the tradeoff that existed in the previous case.

Next, we examine the equilibrium under RRC.

4.2 Repair and Replacement Contract (RRC)

RRC empowers SE with a complete control over all aspects fleet management, including the vehicle replacement decision. The downside is that SE has to assume ownership of the fleet, and as a result she bears both replacement costs and repair costs, as well as extra overhead costs. Under RRC, SE determines both vehicle retirement age τ and the uptime fee r that maximize her objective $\gamma W(\tau) + (1-\gamma)\Pi(r,\tau)$, a weighted average of welfare and her surplus, making sure that she receives enough fees from the government to cover the costs of her operations. Since the government accepts SE's contract offer only if it presents improvements over the status quo option, SE has to determine the values of τ and r that satisfy the government's budget and welfare constraints. Hence, SE solves the problem

$$(\mathcal{P}_2) \quad \max_{r\geq 0, \tau \geq 0} \ \gamma W(\tau) + (1-\gamma)\Pi(r,\tau) \quad \text{s.t.} \ \ \Pi(r,\tau) \geq 0 \ \ \text{and} \ \ B(r,\tau) \leq b_0 \ \ \text{and} \ \ W(\tau) \geq w_0.$$

Since the only cost that the government incurs under RRC is the uptime-based payment to SE, the government's expected spending is equal to $B(r,\tau) = rA(\tau)$. In addition, SE's expected

surplus is equal to $\Pi(r,\tau) = rA(\tau) - C(\tau) - R(\tau) - \psi_2$ because she receives the payment $rA(\tau)$ and incurs the sum of repair and replacement costs $C(\tau) + R(\tau)$ as well as the overhead cost ψ_2. After substituting these in $(\mathcal{P}_2)$ and solving it, we characterize the RRC equilibrium, denoted by the superscript ‡, as follows.

Proposition 2 (Equilibrium under RRC) *Let* $\underline{\tau}_2 \equiv \min\{\tau : C(\tau) + R(\tau) = b_0 - \psi_2\}$ *and* $\widehat{\tau}(\gamma)$ *be the unique solution to* $Q_S(\tau) = 0$ *as defined in Proposition 1. The RRC equilibrium occurs at* $\tau^{\ddagger} = \max\{\underline{\tau}_2, \widehat{\tau}(\gamma)\}$*, where the budget constraint binds while the surplus constraint binds if and only if* $\underline{\tau}_2 \geq \widehat{\tau}(\gamma)$*.*

The proposition reveals a key difference between RRC and ROC: unlike under ROC (Proposition 1), the government's budget constraint always binds at the RRC equilibrium. Therefore, RRC guarantees that SE receives the maximum amount of payments allowed within the government's budget. Although a similar equilibrium outcome is possible under ROC as well, it is not guaranteed; the difference arises from the fact that the government's optimal response under ROC does not always aligns with what SE desires. We now compare RRC with ROC in more detail.

4.3 Comparison of Equilibria

We now discuss the structural differences in the equilibria under ROC and RRC. As we observed above, RRC allows SE to extract surplus away from the government in all situations, leaving the government's budget constraint to be always binding. This is possible because SE directly controls vehicle replacements and determines the vehicle retirement age τ; unlike ROC, which requires SE to use the contract term r as an inducement for a desired value of τ, RRC frees SE from having to use r for such an indirect purpose. As a result, an efficiency gain is achieved by employing RRC. The downside of employing RRC is that it requires an additional overhead cost $\psi_2 - \psi_1$, which offsets the efficiency gain. Hence, SE may or may not prefer RRC. Under what circumstances does SE prefer RRC to ROC, and vice versa?

As a precursor to answering this question, we first examine how SE's contract preference is related to the vehicle retirement age τ and the system-wide total costs of fleet operations $\Gamma(\tau) = C(\tau) + R(\tau)$, sum of repair and replacement costs, focusing on two extreme cases $\gamma \to 0$ and $\gamma \to 1$.

Lemma 2 *Let* $\Gamma(\tau) \equiv C(\tau) + R(\tau)$*. At the ROC equilibrium* $\tau = \tau^{\dagger}$ *and the RRC equilibrium* $\tau = \tau^{\ddagger}$ *described in Proposition 1 and Proposition 2, the following relationships hold for a fixed* α*.*

(a) In the limit $\gamma \to 0$*,* $\Gamma(\tau^{\dagger}) \geq \Gamma(\tau^{\ddagger})$*. If* $\Gamma(\tau^{\dagger}) - \Gamma(\tau^{\ddagger}) > \psi_2 - \psi_1$*, SE prefers RRC to ROC. If* $\Gamma(\tau^{\dagger}) - \Gamma(\tau^{\ddagger}) < \psi_2 - \psi_1$ *and the budget constraint binds under ROC, SE prefers ROC to RRC.*

(b) In the limit $\gamma \to 1$, $\Gamma(\tau^\dagger) - \Gamma(\tau^\ddagger) \leq \psi_2 - \psi_1$ *for* $\alpha > \overline{\alpha}$. *SE prefers RRC to ROC if (i)* $\alpha \leq \overline{\alpha}$ *or (ii)* $\alpha > \overline{\alpha}$ *and* $\tau^\dagger > \tau^\ddagger$, *which arises if* $\Gamma(\tau^\dagger) < \Gamma(\tau^\ddagger)$. *If* $\alpha > \overline{\alpha}$ *and* $\tau^\dagger < \tau^\ddagger$, *on the other hand, SE prefers ROC to RRC.*

According to the lemma, if SE is profit-oriented ($\gamma \to 0$; Lemma 2(a)), then SE prefers RRC only if employing RRC leads to a significant saving in total cost that more than compensates for the additional overhead cost ($\Gamma(\tau^\dagger) - \Gamma(\tau^\ddagger) > \psi_2 - \psi_1$). By contrast, if SE is welfare-oriented ($\gamma \to 1$; Lemma 2(b)), her contract preference is determined by whether scheduled vehicle replacements occur more frequently under one contract than the other ($\tau^\dagger < \tau^\ddagger$ or $\tau^\dagger > \tau^\ddagger$), rather than how much costs are saved. In fact, it is possible that a welfare-oriented SE prefers RRC even if employing RRC results in a higher total cost ($\Gamma(\tau^\dagger) < \Gamma(\tau^\ddagger)$). This happens because a welfare-oriented SE reinvests the efficiency gain from RRC into performing frequent vehicle replacements ($\tau^\dagger > \tau^\ddagger$), since doing so lowers the average age of a vehicle in the fleet and increases availability and welfare.

Lemma 2 also indicates that a welfare-oriented SE's contract preference depends on α, i.e., whether or not the government is also welfare-oriented, whereas dependency on α is not prominent if SE is profit-oriented. In the next result, we expand on this observation and relate SE's contract preference with different combinations of γ and α.

Proposition 3 *At the ROC equilibrium* $\tau = \tau^\dagger$ *and the RRC equilibrium* $\tau = \tau^\ddagger$ *described in Proposition 1 and Proposition 2, given sufficiently large* $\psi_2 - \psi_1$, *the following relationships hold.*

(a) For sufficiently small γ *and small* α, *SE prefers ROC to RRC with* $\tau^\dagger > \tau^\ddagger$.

(b) For sufficiently small γ *and large* α, *SE prefers ROC to RRC with* $\tau^\dagger \geq \tau^\ddagger$.

(c) For sufficiently large γ *and small* α, *SE prefers RRC to ROC with* $\tau^\dagger > \tau^\ddagger$.

(d) For sufficiently large γ *and large* α, *SE prefers ROC to RRC with* $\tau^\dagger < \tau^\ddagger$.

Note that Proposition 3 assumes $\psi_2 - \psi_1$, the additional overhead cost under RRC, is not insignificant such that a tradeoff between ROC and RRC exists. Proposition 3 reveals that, among the four contrasting combinations of γ and α, only the case with a large γ and a small α results in an outcome where SE prefers RRC to ROC (part (c)). That is, RRC is preferred in an environment in which a welfare-oriented SE contracts with a cost-oriented government. By contrast, if SE is profit-oriented (parts (a) and (b)) or if both SE and the government are welfare-oriented (part (d)), ROC is preferred instead. See Figure 1 for illustrations. (The following parameter values are chosen to generate the examples in Figure 1: $H(t) = t^2$, $v = 30$, $K = 5$, $k = 1$, $L = 0.1$, $l = 0.02$, $p = 0.05$, $q = 0.25$, $b_0 = 6$, $\psi_1 = 2$.) Comparing the two preference maps in Figure 1, we see that, as expected, the RRC-preference region (lower right corner) shrinks as $\psi_2 - \psi_1$, the additional overhead cost of employing RRC, increases from 0.2 to 0.8. (It is numerically observed that the regions shrinks gradually as $\psi_2 - \psi_1$ increases.)

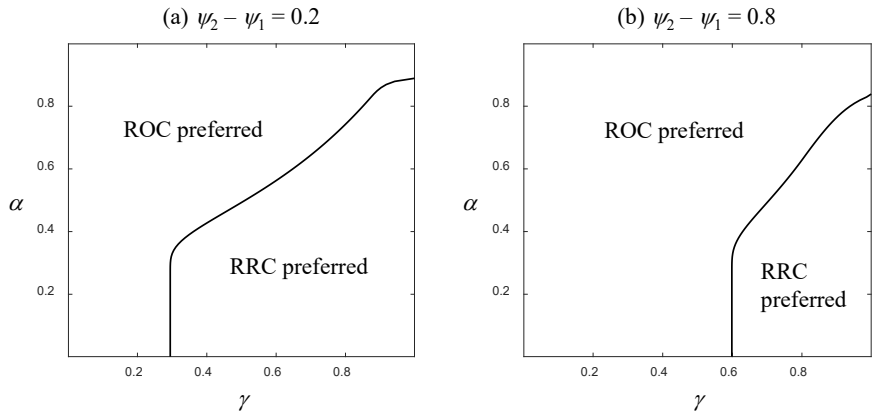

Figure 1 **SE's contract preference for two different values of overhead cost difference** $\psi_2 - \psi_1$

Interestingly, Proposition 3(d) indicates that it is ROC, not RRC, that SE prefers when both SE and the government are welfare-oriented (upper right corners in Figure 1(a) and (b)). This is despite the efficiency-gain advantage of RRC, which a welfare-oriented SE can utilize to achieve her goal of increasing vehicle availability and welfare. This happens because, under ROC, a welfare-oriented government voluntarily replaces vehicles frequently, lessening the adverse effects of higher fee vs. delayed vehicle retirement tradeoff that exists under ROC (see the discussions below Proposition 1). This, combined with a lower overhead cost under ROC, makes ROC a preferred choice for a welfare-oriented SE. Hence, in this case, an incentive alignment favors ROC.

In Rider's case, recently there has been a push for more RRC adoption. Given our observations above, this indicates that Riders often find themselves in a situation described in Proposition 3(c), namely a welfare-oriented SE facing a cost-oriented government. To gain intuition about the welfare and cost implications that arise in this situation, in the next result we examine a special case with $\gamma \to 1$ and $\alpha \to 0$.

Corollary 1 *SE prefers RRC to ROC for* $\gamma \to 1$ *and* $\alpha \to 0$*, resulting in* $\tau^\dagger > \tau^\ddagger$*,* $W(\tau^\dagger) < W(\tau^\ddagger)$*,* $C(\tau^\dagger) > C(\tau^\ddagger)$*, and* $R(\tau^\dagger) < R(\tau^\ddagger)$*. Moreover,* $C(\tau^\dagger) + R(\tau^\dagger) < C(\tau^\ddagger) + R(\tau^\ddagger)$ *for* $\psi_2 - \psi_1 \to 0$ *while* $C(\tau^\dagger) + R(\tau^\dagger) > C(\tau^\ddagger) + R(\tau^\ddagger)$ *for sufficiently large* $\psi_2 - \psi_1$*.*

Corollary 1 reveals how a welfare-maximizing SE takes advantage of efficiency gain under RRC when she contracts with a cost-minimizing government: she invests it in vehicle replacements, shortening vehicle retirement age ($\tau^\dagger \geq \tau^\ddagger$) and bringing new vehicles to the fleet more often. In fact, a welfare-maximizing SE operates on a nonprofit basis, reinvesting all of her surplus into improving welfare (see Corollary A3 in the Online Appendix). As a result, replacement cost is higher under RRC ($R(\tau^\dagger) \leq R(\tau^\ddagger)$) but vehicle breakdowns occur less often than under ROC, lowering repair cost ($C(\tau^\dagger) \geq C(\tau^\ddagger)$) and increasing welfare ($W(\tau^\dagger) \leq W(\tau^\ddagger)$). The welfare increase occurs even if

RRC requires a higher total cost ($C(\tau^\dagger) + R(\tau^\dagger) \leq C(\tau^\ddagger) + R(\tau^\ddagger)$), which the welfare-maximizing SE tolerates. Thus, a high-cost, high-welfare equilibrium may arise in this situation.

In closing, we note that our model can be used to quantify relative performances of ROC and RRC under various environmental conditions. For instance, by estimating the model parameters such as p, q, L, and l and computing the average vehicle uptime and costs, one may evaluate the costs and benefits of switching from ROC to RRC. Clearly, these calculations depend critically on parameter values and some of the simplifying assumptions used in our analysis. To calibrate model predictions, one can add more features to the model that reflect actual practice, which we consider in the next section.

5. Extended Analysis and Numerical Studies

5.1 General Fleet Size and Concave Valuation of Vehicle Uptime

Thus far, our analysis has relied on a simplifying assumption that the uptime-to-welfare mapping $V(\cdot)$ scales linearly in vehicle uptime. In addition, the fleet size was normalized to one in order to facilitate analysis. We now generalize the model by relaxing both restrictions, accommodating a general fleet size $n \geq 1$ and examining the effects of concavity in $V(\cdot)$.

Concavity assumption is a natural way to capture the realities of healthcare deliveries. Even if a vehicle is available, it will not be utilized if an immediate delivery is prevented by unpredictable local factors such as unstocked medicine and health kits, flooding of roads, and insufficient staffing levels. In one of Riders' episodes, a severe fuel shortage led Riders to temporarily cease their operations in Mozambique, Kenya, and Zimbabwe, as it was unnecessary to keep motorcycles available when there is no fuel. In such instances, increasing total vehicle uptime does not lead to a tantamount increase in welfare; the relationship between the two is likely to feature diminishing marginal returns. Hence, we expect that the effects of concavity will become more pronounced as the total vehicle uptime increases.

For simplicity, we assume that individual vehicles in the fleet are identical and are managed independently in parallel, and that their failures occur according to independent and identically distributed processes that satisfy the properties discussed in §3. This implies that the long-run average total repair and replacement costs are proportional to the fleet size n, i.e., the total cost is equal to $nC(\tau) + nR(\tau)$, and that the long-run average total uptime provided by the fleet operation is equal to $nA(\tau)$; identicalness and independence guarantee that a single choice of τ applies to all identical vehicles. Concurrently, to enable fair comparisons, we adjust the exogenous model parameters as follows: $\psi_1 \to n\psi_1$, $\psi_2 \to n\psi_2$, $b_0 \to nb_0$, and $w_0 \to nw_0$.

To capture the effects of concavity succinctly, we specify $V(\cdot)$ as $V(z) = vz^a$ with $0 < a \leq 1$. The case $a = 1$ corresponds to the assumption of our main model, i.e., linear mapping between uptime

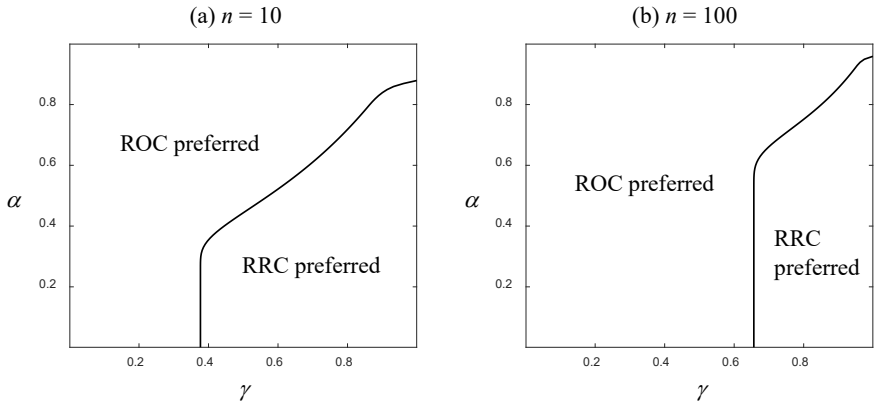

Figure 2 SE's contract preference for two different fleet sizes $n = 10$ and $n = 100$, in the presence of a concave increasing mapping between vehicle uptime and welfare

and welfare. With the new specification, uptime and welfare are related as $W(\tau) = V(nA(\tau)) = vn^a A(\tau)^a$. Since vehicle availability $A(\tau)$ is close to one in most real-world situations, we can further simplify the expression by applying the approximation $A(\tau)^a \approx A(\tau)$, which results in $W(\tau) = vn^a A(\tau)$. Therefore, the effect of concavity is captured by the term n^a.

It can be shown that the analysis of this generalized model is analogous to that presented in §4. In fact, for identifying ROC and RRC equilibria, the n-vehicle problem described here becomes equivalent to the single-vehicle problem, achieved by replacing v from §4 with v/n^{1-a} and factoring out n. Hence, all of the results in §4 are retained with this adjustment in v. Clearly, the fleet size n will impact the comparison between ROC and RRC equilibria, which we examine next.

Figure 2 presents two maps depicting SE's preference for ROC vs. RRC, similar to Figure 1 except that two different fleet sizes $n = 10$ and $n = 100$ are considered for a fixed value of $\psi_2 - \psi_1$. (In these examples, $a = 0.5$, $v = 100$, $\psi_1 = 2$, and $\psi_2 = 2.3$ are chosen along with other parameter values identical to those from Figure 1; note vehicle availability close to one arises from these parameter values, as assumed above.) From the two maps in Figure 2, we observe that an enlarged fleet size n leads to two changes. First, the RRC-preference region shrinks for a small α. Second, the RRC-preference region expands for a large γ. These two changes are related. As the fleet size n grows, the diminishing returns to scale on welfare exhibited in $V(\cdot)$ push SE and the government to become less welfare-oriented and focus more on their profit/costs, effectively reducing γ and α. If the government was already cost-oriented (small α), then a more profit-oriented SE (effectively reduced γ) finds the welfare-enhancing advantage of RRC less attractive, which explains the first change. If SE was already welfare-oriented (large γ), on the other hand, RRC becomes more attractive to her if she faces a more cost-oriented government (effectively reduced α), since the latter party erodes the advantage of ROC by delaying vehicle retirements and hence driving up SE's repair costs; this explains the second change. Additionally, we make the following observation.

Proposition 4 *For $\gamma \to 1$ and $\alpha \to 0$, SE's preference for RRC over ROC widens as n increases.*

In other words, for a welfare-maximizing SE contracting with a cost-minimizing government, the value of RRC relative to that of ROC increases with the fleet size; the efficiency gap between RRC and ROC widens as more vehicles are used. From these observations, we conclude that the value of RRC is especially high for a welfare-oriented SE like Riders if she operates a large fleet of vehicles. These findings provide a theoretical support for Riders' recent move toward RRC adoption as it expands its service into wider rural areas in many countries.

We note that the effects described in Figure 2 and Proposition 4 will be moderated if the cost functions do not scale linearly as assumed above, e.g., if the total cost increases in n in a concave manner instead of $nC(\tau) + nR(\tau)$. Such a situation will arise if management utilizes resource pooling, for example through grouped repairs, parts sharing, and discounted bulk purchasing of new vehicles. Given the challenging local factors that we mentioned above, however, it is likely that such pooling effects are dominated by diminishing returns on welfare.

5.2 Residual Value of Vehicle

Another simplifying assumption in the main analysis was that a constant fixed cost K is incurred whenever a vehicle is replaced, regardless of whether a replacement is performed on a scheduled or an unscheduled basis. This assumption, which enables analytical tractability, is consistent with an interpretation that the replacement cost is equal to the cost of acquiring a new vehicle. In reality, however, a vehicle has a residual value at the time of a scheduled replacement, which can be recouped; for example, an aged but functioning vehicle may be sold in an open market at a price that reflects the vehicle's cumulative mileage, thus offsetting the cost of acquiring a new vehicle. Absent this consideration, we overestimate the long-run average cost of replacements, $R(\tau)$. (Other performance measures, such as $A(\tau)$ and $C(\tau)$, are not impacted.)

To accommodate this residual value consideration, we modify $R(\tau) = K/T(\tau)$ derived in (1) as follows. (Note $T(\tau)$ is the length of each replacement cycle and $R(\tau)$ satisfies $R'(\tau) < 0$; a delayed scheduled replacement lowers the average replacement cost.) An ability to recoup the residual value implies that $R(\tau)$ is lowered for any given length of cycle $T(\tau)$, which suggests that the numerator K of $R(\tau) = K/T(\tau)$ should be substituted by a smaller value. In addition, this substitution should be a τ-dependent function—which we denote as $\kappa(\tau)$—since the residual value is determined by a replaced vehicle's age τ at the time of scheduled replacement.

Intuitively, the cost-reduction effect of residual value should be more significant when scheduled replacements occur more often; a vehicle sold in an open market with a lower mileage brings in more revenue than a vehicle with a higher mileage does. On the other hand, if scheduled replacements are rarely performed (very large τ), then the cost-reduction effect of residual value is negligible. These

suggest that, with K denoting the cost of acquiring a new vehicle, $\kappa(\tau)$ should be an increasing function that converges to K as $\tau \to \infty$. A functional form for $\kappa(\tau)$ that naturally satisfies these criteria is

$$\kappa(\tau) = K(1 - e^{-\tau/\delta}),$$

where $1/\delta$ captures the rate of vehicle value depreciation. Higher δ for fixed τ corresponds to a greater residual value, and in the limit $\delta \to 0$, our earlier assumption of zero residual value is restored.

Based on the modified average replacement cost $R(\tau) = \kappa(\tau)/T(\tau)$, we numerically examine the effects of varying δ. It is observed that, for large γ (welfare-oriented SE), higher δ results in a lower value of $\tau^{\ddagger}$ under RRC; a lower rate of value depreciation prompts SE to perform scheduled replacements more frequently. This makes intuitive sense, since a higher residual value of a vehicle makes it less costly for SE to replace a vehicle, whose high reliability improves welfare through increased vehicle availability. Under ROC, it is observed that a similar effect is either substantially more muted or nonexistent. This is because the impact of residual value is small if vehicles are replaced infrequently; since the government tends to delay replacements under ROC, he does not take full advantage of recouping a vehicle's residual value.

Since more frequent scheduled replacements under RRC increase vehicle availability and welfare, the above observations imply that a strong residual value of a vehicle enhances a welfare-oriented SE's preference for RRC. Indeed, this prediction is supported by numerical examples that show RRC-preference region in Figure 1 expanding to a greater range of α as the residual value increases (higher δ). Therefore, we conclude that RRC is more attractive to a welfare-oriented SE like Riders if she utilizes vehicle models that enjoy strong resale values.

5.3 Failure Prevention and Repair Lead Time Reduction Efforts

To enable a fair comparison between ROC and RRC, it has been assumed that all environmental parameters are fixed except for the decision variable τ. In practice, SEs invest in measures aimed at improving the operating environment. For Riders, failure prevention and repair lead time reduction are two important components of its fleet operations. Failure prevention is achieved by investments in measures such as healthcare worker training, through which the workers learn how to self-check vehicles to ensure that the filters are clean, tires have adequate pressure, etc. In addition, Riders achieved reduced repair lead time by building a hub-and-spoke distribution network for service parts, which enabled faster responsiveness to repair requests. As illustrated in Table 1 in §3.4, these efforts by Riders in its Zambia operations (Mehta et al. 2016) resulted in increased vehicle uptime (from 1.3 days per week to 5.5 days per week) and prolonged vehicle life (major failures reduced

from 19% to 0%), as well as higher health worker productivity measured in the number of outreach trips, patient visits, immunizations, child growth monitoring, and health education sessions.

Motivated by this practice, we extend our model to capture SE's decisions on failure prevention and repair lead time reduction, to understand how SE would optimally invest in the two measures to improve its operations. We focus exclusively on RRC, reflecting a recent trend towards more RRC adoption by Riders. We introduce new decision variables $x \geq 0$ and $y \geq 0$, which represent "failure prevention effort" and "repair lead time reduction effort," respectively. The optimal effort levels are assumed to be chosen at time zero and remain unchanged afterwards. For simplicity, we assume that SE incurs linear effort costs $c_x x$ and $c_y y$ per unit time to maintain the chosen levels of x and y. (Hence, $\psi_2 = c_x x + c_y y$ under RRC at the optimally chosen values of x and y.) Exerting the effort y reduces the repair lead time from a constant l_0 to $l = l_0/(1+y)$. Similarly, exerting the effort x reduces the overall failure rate from $\lambda_0 h(t)$ to $\lambda_0 h(t)/(1+x)$ where λ_0 is a constant (see §3.1). As a result, the scaling factors p and q for major and minor failure rates are related to x as

$$p = \frac{\phi \lambda_0}{1+x} \quad \text{and} \quad q = \frac{(1-\phi)\lambda_0}{1+x}, \tag{7}$$

where the parameter ϕ in (7) denotes the fraction of prevention effort directed at reducing the rate of major failures as opposed to minor failures. Note that the factors $1/(1+x)$ and $1/(1+y)$ capture diminishing marginal returns in effort.

Exerting the efforts x and y has both direct and indirect impacts on performance outcomes, as the changes in (7) and $l = l_0/(1+y)$ not only shift performance measure curves $A(\tau)$, $C(\tau)$, and $R(\tau)$ for each τ but also lead to adjustments in the optimal choice of τ. It is the net of the direct and indirect impacts that determine the optimal effort level choices, denoted by the superscript $*$. Characterizing the optimal effort choices is, in general, analytically intractable. As such, we numerically investigate the optimal choices and discuss the observations; see the examples in Figure 3, which plot x^* and y^* as functions of γ for two different values of ϕ. (In these examples, $c_x = c_y = 0.2$, $\lambda_0 = 0.3$, and $l_0 = 0.1$ are chosen along with other parameter values identical to those from Figure 1.) The values $\phi = 0.01$ and $\phi = 0.25$ in Figure 3(a) and Figure 3(b) correspond to 1% and 25% of the total failures being identified as major failures.

Three patterns are observed from the plots in Figure 3 (the same patterns are found in other numerical examples). First, both x^* and y^* increase in γ. In other words, a welfare-oriented SE (large γ) exerts more efforts than a profit-oriented SE does. This is in line with intuition, and it confirms Riders' aforementioned commitment in these efforts. Second, a welfare-oriented SE (large γ) facing a high chance of major failure (large ϕ) rely more on failure prevention and less on lead time reduction, compared with an SE who faces a low chance of major failure (for a large γ, the

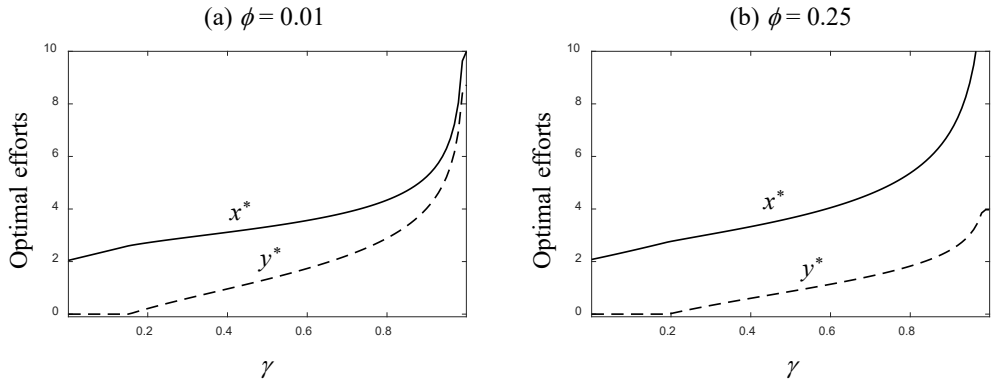

Figure 3 **Optimal prevention effort x^* and repair lead time reduction effort y^* for two different values of ϕ.**

gap $x^* - y^*$ is larger for $\phi = 0.25$ than for $\phi = 0.01$). This is consistent with Riders' experience, as Riders focused on failure prevention when it first took over the maintenance operations run by local governments under which a large fraction of vehicles in the fleets had been discarded after experiencing premature deaths (Lee et al. 2013).

Finally, a profit-maximizing SE ($\gamma \to 0$) prioritizes failure prevention over lead time reduction ($x^* > y^*$; in fact, $y^* = 0$ in the examples in Figure 3). This pattern is observed even in cases where the prevention effort is much more costly than the lead time reduction effort ($c_x \gg c_y$). This reflects a downside of lead time reduction, namely that it increases SE's costs (see Lemma A7 in the Online Appendix for a proof); quicker repairs allow a vehicle to spend more time being utilized, which exposes the vehicle to more wear-and-tear that increases the chance of failures. As a result, vehicle repairs and replacements are needed more frequently, leading to higher costs that a profit-maximizing SE avoids. By contrast, a welfare-oriented SE does not necessarily focus on failure prevention, opting instead to balance its failure prevention effort with lead time reduction effort. This is evidenced by other numerical examples that show reversed priority ($x^* < y^*$).

In summary, our analysis in this subsection corroborates some of Riders' existing practices and brings new insights that highlight the difference between a welfare-oriented SE and a profit-oriented SE. These insights shed light on how an SE would optimally allocate resources between failure prevention and repair lead time reduction.

6. Conclusion

In general, one reason why an organization would outsource its operations to a social enterprise (SE) is that the SE can offer superior operational performance than what the organization can achieve by itself. Such is the case of Riders for Health, an SE that, on behalf of governments in developing countries, manages maintenance operations of vehicles used for delivering life-saving health products to rural communities in those countries. In this outsourcing arrangement, a right

choice of contracting arrangement is of paramount importance, as it influences the incentives of contracting parties and affects the eventual performance outcomes.

When Riders first started providing its service to the governments of sub-Saharan African countries such as Zimbabwe and Gambia, it only provided vehicle maintenance and repair services, referred to in our paper as Repair Only Contract or ROC (Tayan 2007). It then introduced adding vehicle replacement and purchases to its responsibility, referred to in our paper as Repair and Replace Contract or RRC, in Gambia (Lee et al. 2013). In practice, RRC is more costly to implement than ROC because of added requirements of owning a fleet of vehicles and managing all aspects of fleet operations. For example, one of the challenges that Riders faced was securing a loan arrangement with financial institutions in order to acquire vehicles and enter into a long-term contractual agreement with local governments. Despite these challenges, Riders has moved towards more RRC adoption as RRC became the dominant contracting approach used in Zambia in recent years (Mehta et al. 2015 and 2016).

Such preference for RRC reflects Riders' commitment to improving health outcomes as a welfare-maximizing nonprofit SE, i.e., an SE whose sole focus is on improving social benefits. By contrast, our analysis shows that an SE with a profit motive would prefer ROC instead. It is when an SE reinvests its financial gains into further improvements in fleet operations—as Riders does—that RRC becomes a preferred choice. Furthermore, preference for RRC becomes even stronger if a welfare-oriented SE contracts with a government that focuses on cost reduction. We show that the SE enjoys a higher level of vehicle availability under RRC than under ROC. Increased vehicle availability in turn enables more trips to rural communities in need, thus improving health outcomes. Interestingly, we find that RRC may result in higher operating costs than ROC does in some situations; hence, a welfare-oriented SE adopting RRC may operate vehicles in a high-cost, high-benefit environment.

Our analysis also reveals a number of other useful insights that serve as guidelines for SEs such as Riders. For example, under reasonable assumptions, we find that RRC is more attractive to a welfare-oriented SE when it operates a large fleet that consists of vehicles that have strong resale values. We also find that, under RRC, a profit-seeking SE prioritizes failure prevention over repair lead time reduction, whereas a welfare-oriented SE does not necessarily have the same priority. Finally, the model developed in this paper can be used to quantify the costs and benefits of adopting ROC or RRC; through parameter estimations and model calibrations, one can apply our model to predict the differences in performance outcomes under the two contracts.

References

African Development Bank (2014) Infrastructure development. *Tracking African Progress in Figures*, Tunis, Tunisia.

Austin J, Stevenson H, Wei-Skillern J (2006) Social and commercial entrepreneurship: Same, different, or both? *Entrepreneurship Theory and Practice*. 30, 1–22.

Barlow RE, Hunter L (1960) Optimum preventive maintenance policies. *Oper. Research*. 8(1), 90–100.

Barlow RE, Proschan F (1996) *Mathematical Theory of Reliability* (SIAM).

Beichelt F (1976) A general preventive maintenance policy. *Mathematische Operationsforschung und Statistik*, 7(6), 927–932.

Beichelt F (2006) *Stochastic Processes in Science, Engineering and Finance*, Chapman and Hall/CRC.

Besiou M, van Wassenhove LNV (2015) Addressing the Challenge of Modeling for Decision-Making in Socially Responsible Operations. *Production and Oper. Management*. 24(9), 1390–1401.

Besley T, Ghatak M (2017) Profit with purpose? A theory of social enterprise. *American Economic Journal: Economic Policy*. 9(3), 19–58.

Bossert TJ, Bowser DM, Amenyah JK (2007) Is decentralization good for logistics systems? Evidence on essential medicine logistics in Ghana and Guatemala. *Health Policy and Planning* 22, 73–82.

Calmon AP, Jue-Rajasingh D, Romero G, Stenson J (2017) Consumer education and regret returns in a social enterprise. Working paper, University of Toronto.

Chen Y, Shanthikumar JG, Shen ZM (2013) Training, production, and channel separation in ITC's e-Choupal network. *Prod. and Oper. Management.* 22(2), 348–364.

Deo S, Gallien J, Jonasson JO (2015) Improving HIV early infant diagnosis supply chains in sub-Saharan Africa: Models and application to Mozambique. Working paper, Indian School of Business.

Doherty B, Haugh H, Lyon F (2014) Social enterprises as hybrid organizations: A review and research agenda. *International Journal of Management Reviews*, 16(4), 417-436.

Guajardo J, Cohen MA, Kim S-H, Netessine S (2012) Impact of performance-based contracting on product reliability: An empirical analysis. *Management Sci.* 58(5), 961–979.

Hogan MC, Foreman KJ, Naghavi M, Ahn SY, Wang M, Makela SM, Lopez AD, Lozano R, Murray CJ (2010) Maternal mortality for 181 countries, 1980–2008: A systematic analysis of progress towards millennium development goal 5. *Lancet*. 375, 9726, 1609-1623.

Hu M, Liu Y, Wang W (2016) Altruistic rationality: The value of strategic farmers, social entrepreneurs and for-profit firms in crop planting decisions. Working paper, University of Toronto.

Jain N, Hasija S, Popescu DG (2013) Optimal contracts for outsourcing of repair and restoration services. *Oper. Research.* 61(6), 1295–1311.

Katz RA, Page A (2010) The role of social enterprise. *Vermont Law Review.* 35, 59–103.

Kim S-H, Cohen MA, Netessine S (2007) Performance contracting in after-sales service supply chains. *Management Sci.* 53(12), 1843–1858.

Lee HL, Rammohan SV, Sept L (2013) Innovative logistics in extreme conditions: The case of health care delivery in gambia. *Handbook of Global Logistics*, edited by Bookbinder JH, Springer, New York, 297–322.

Lee HL, Tang CS (2017) Socially and environmentally responsible value chain innovations: New operations management research opportunities. Forthcoming in *Management Science.*

Leung NHZ, Chen A, Yadav P, Gallien J (2016) The impact of inventory management on stock-outs of essential drugs in sub-Saharan Africa: Secondary analysis of a field experiment in Zambia. *PLoS ONE* 11(5): e0156026. doi:10.1371/journal.pone.0156026.

Mair J, Marti I (2006) Social entrepreneurship research: A source of explanation, prediction, and delight. *J. World Business.* 41, 36–44.

McCoy JH, Lee HL (2014) Using fairness models to improve equity in health delivery fleet management. *Production and Oper. Management*, 23(6), 965–977.

Mehta KM, Rammohan SV, Albohm DC, Muwowo G, Sept L, Lee HL, Bendavid E (2015) Evaluating the effectiveness and efficiency of outsourcing vehicle fleet management to strengthen health delivery: An evaluation of riders for health. Final Presentation to the Gates Foundation.

Mehta KM, Rerolle F, Rammohan SV, Albohm DC, Muwowo G, Moseson H, Sept L, Lee HL, Bendavid E (2016) Systematic motorcycle management and health care delivery: A field trial. *American Journal of Public Health*, 106(1), 87–94.

Nakagawa T (2005) *Maintenance Theory of Reliability.* Springer, London.

Pedraza-Martinez AJ, van Wassenhove LN (2012) Transportation and vehicle fleet management in humanitarian logistics: Challenges for future research. *EURO J. on Transportation and Logistics.* 1(1), 185–196.

Pedraza-Martinez AJ, van Wassenhove LN (2013) Vehicle replacement in the international committee of the Red Cross. *Production Oper. Management.* 22(2), 365–376.

Riders for Health (2014) 2014 Annual Report & Accounts.

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

Sodhi MS, Tang CS (2017) Social responsibility in supply chains. *Sustainable Supply Chains: A Research-Based Textbook on Operations and Strategy.* Eds. Bouchery Y. et al. Springer, New York.

Szymańska A, Jegers B (2016) *Modelling social enterprises. Annals of Public and Cooperative Economics*. 87(4), 501–527.

Tayan B (2007) Riders for Health: Health care distribution solutions in sub-Saharan Africa. Stanford Graduate School of Business, Case GS-58.

United Nations (2014) *World Urbanization Prospects: The 2014 Revision Highlights*. New York.

Uppari BS, Popescu I, Netessine S (2017) Selling off-grid light to liquidity constrained consumers. Working paper, INSEAD.

Wilson F, Post JE (2013) Business models for people, planet (&profits): Exploring the phenomena of social business, a market-based approach to social value creation. *Small Business Economics.* 40, 715–737.

World Health Organization (2010) World health statistics 2010 (part II). `http://www.who.int/whosis/whostat/EN_WHS10_Part2.pdf`.

Appendix: Proofs of Main Results

(Lemmas and equations labeled "A" are found in the accompanying Online Appendix.)

Proof (**Lemma 1**) Since $A'(\tau) < 0$ with $A(0) = \infty$ and $A(\infty) = \frac{\mu p}{(\mu+L)p+lq} > \frac{w_0}{v}$ (Lemma A3), where the inequality follows from Assumption 1, the welfare constraint $W(\tau) = vA(\tau) \geq w_0$ never binds. Thus, the government's problem in (4) can be written as $\max \alpha W(\tau) - (1-\alpha)B(r,\tau)$ s.t. $B(r,\tau) \leq b_0$. Since the left hand-side of the budget constraint $B(r,\tau) = rA(\tau) + R(\tau) \leq b_0$ decreases in τ for fixed r starting from $B(r,0) = \infty$ while it increases in r for fixed τ (Lemma A3), the optimal response determined at the binding budget constraint satisfies $\tau_G(r) < \infty$ and $\tau_G'(r) > 0$. Suppose that the budget constraint does not bind. According to Lemma A6, in this case the optimal response is $\tau = \widehat{\tau}_g(r) < \infty$ if $r < \frac{\alpha}{1-\alpha}v - \frac{K}{\mu}$ and $\tau = \infty$ otherwise, where $\widehat{\tau}_g(r) < \infty$ is the unique root of $Q_g(\tau) \equiv (Lp+lq)\left[\left(\frac{\alpha}{1-\alpha}v - r\right)S(\tau) - Kh(\tau)\right] - K$ (Lemma A6). Implicitly differentiating $Q_g(\widehat{\tau}_g(r)) = 0$ and using the result $Q_g'(\widehat{\tau}_g(r)) > 0$ proved in Lemma A5, we find $\tau_G'(r) = \widehat{\tau}_g'(r) > 0$. □

Proof (**Proposition 1**) Recall $W(\tau) = vA(\tau)$, $\Pi(r,\tau) = rA(\tau) - C(\tau) - \psi_1$, and $B(r,\tau) = rA(\tau) + R(\tau)$ under ROC. Since $A'(\tau) < 0$ with $A(0) = \infty$ and $A(\infty) = \frac{\mu p}{(\mu+L)p+lq} > \frac{w_0}{v}$ according to Lemma A3 and Assumption 1, the welfare constraint $W(\tau) = vA(\tau) \geq w_0$ never binds and it is dropped

from the problem formulation. In addition, $\widetilde{\Pi}(\tau)=\rho(\tau)A(\tau)-C(\tau)-\psi_1$ and $\widetilde{B}(\tau)=\rho(\tau)A(\tau)+R(\tau)$ defined in (6) satisfy $\widetilde{\Pi}(0)=\widetilde{B}(0)=-\infty$, $\widetilde{\Pi}(\infty)=\frac{\alpha}{1-\alpha}vA(\infty)-C(\infty)-R(\infty)-\psi_1$, and $\widetilde{B}(\infty)=\frac{\alpha}{1-\alpha}vA(\infty)$ where $A(\infty)=\frac{\mu p}{(\mu+L)p+lq}$, $C(\infty)=\frac{kq}{(\mu+L)p+lq}$, and $R(\infty)=\frac{Kp}{(\mu+L)p+lq}$. Suppose that $\widetilde{\Pi}(\tau)$ and $\widetilde{B}(\tau)-b_0$ each cross zero exactly once at τ_1^s and τ_1^b, respectively. Since $\widetilde{\Pi}(0)=\widetilde{B}(0)=-\infty$, the crossings occur from negative to positive, from which it follows that $\widetilde{\Pi}'(\tau_1^s)>0$ and $\widetilde{B}'(\tau_1^b)>0$. This implies $\widetilde{\Pi}(\tau)\geq 0$ iff $\tau\geq\tau_1^s$ and $\widetilde{B}(\tau)\leq b_0$ iff $\tau\leq\tau_1^b$. Moreover, $\widetilde{\Pi}(\tau)\geq 0$ and $\widetilde{B}(\tau)\leq b_0$ together imply $C(\tau)+R(\tau)\leq b_0-\psi_1$. As proved in Corollary A1, the function $C(\tau)+R(\tau)$ that appears on the left-hand side of this condition is quasiconvex with a unique interior minimum with the limits $C(0)+R(0)=\infty$ and $C(\infty)+R(\infty)=\frac{Kp+kq}{(\mu+L)p+lq}<b_0-\psi_2<b_0-\psi_1$, where the inequality follows from Assumption 1. Hence, the condition $C(\tau)+R(\tau)\leq b_0-\psi_1$ can be rewritten as $\tau\geq\underline{\tau}_1$. These observations imply that $\widetilde{\Pi}(\tau)\geq 0$ and $\widetilde{B}(\tau)\leq b_0$ can be together rewritten as $\max\{\tau_1^s,\underline{\tau}_1\}\leq\tau\leq\tau_1^b$.

Suppose $\tau=\infty$ in equilibrium. Given $r\geq 0$, according to Lemma A6, this outcome arises iff (i) $\frac{\alpha}{1-\alpha}v\leq\frac{K}{\mu}$ or (ii) $r\geq\frac{\alpha}{1-\alpha}v-\frac{K}{\mu}>0$. With $\tau=\infty$ substituted, $(\mathcal{P}_1)$ becomes $\max_{r\geq 0}\gamma vA(\infty)+(1-\gamma)[rA(\infty)-C(\infty)-\psi_1]$ s.t. $rA(\infty)-C(\infty)-\psi_1\geq 0$ and $rA(\infty)+R(\infty)\leq b_0$, from which it is clear that the optimal solution is found by increasing r until the last constraint (budget constraint) binds, i.e., set $r=\frac{b_0-R(\infty)}{A(\infty)}=b_0\left(1+\frac{Lp+lq}{\mu p}\right)-\frac{K}{\mu}$. The condition $b_0>\psi_2+\frac{Kp+kq}{(\mu+L)p+lq}$ in Assumption 1 ensures that the first constraint (surplus constraint) is satisfied at this value of r and does not bind in equilibrium. At this value of r, the condition (ii) $r\geq\frac{\alpha}{1-\alpha}v-\frac{K}{\mu}>0$ becomes $\frac{K}{\mu}<\frac{\alpha}{1-\alpha}v\leq b_0\left(1+\frac{Lp+lq}{\mu p}\right)$. In addition, the condition (i) above can be written as $\frac{\alpha}{1-\alpha}v\leq\frac{K}{\mu}<b_0\left(1+\frac{Lp+lq}{\mu p}\right)$ where the last inequality follows from $b_0>\psi_2+\frac{Kp+kq}{(\mu+L)p+lq}\geq\frac{Kp}{(\mu+L)p+lq}$ in Assumption 1. The two newly derived conditions together are combined as $\frac{\alpha}{1-\alpha}v\leq b_0\left(1+\frac{Lp+lq}{\mu p}\right)$, or equivalently $\alpha\leq\overline{\alpha}=\left(1+\frac{v}{b_0}\frac{\mu p}{(\mu+L)p+lq}\right)^{-1}$; this is the necessary and sufficient condition for having $\tau=\infty$ in equilibrium, described in part (a) of the proposition.

Next, suppose $\alpha>\overline{\alpha}$, which can be rewritten as $\frac{\alpha}{1-\alpha}v>b_0\left(1+\frac{Lp+lq}{\mu p}\right)$. Then the condition $b_0>\psi_2+\frac{Kp+kq}{(\mu+L)p+lq}\geq\frac{Kp}{(\mu+L)p+lq}$ from Assumption 1 implies $\frac{\alpha}{1-\alpha}v>\frac{K}{\mu}$, which ensures that for $\rho(\tau)$ defined in (5), there exists a sufficiently large τ over which $\rho(\tau)>0$ since $\rho'(\tau)=\frac{Kh'(\tau)T(\tau)}{(Lp+lq)S(\tau)^2}>0$ with $\rho(0)=-\infty$ and $\rho(\infty)=\frac{\alpha}{1-\alpha}v-\frac{K}{\mu}>0$; hence, $r=\rho(\tau)>0$ is assured for some τ. In this case $\tau<\infty$ in equilibrium. According to Lemma A6, the budget constraint $B(r,\tau)\leq b_0$ in (4) binds in equilibrium iff $Q_g(\underline{\tau}_g(r))\geq 0$, where $Q_g(\tau)=(Lp+lq)\left[\left(\frac{\alpha}{1-\alpha}v-r\right)S(\tau)-Kh(\tau)\right]-K$ and $\underline{\tau}_g(r)$ is the unique solution to the binding budget constraint $B(r,\tau)=rA(\tau)+R(\tau)=b_0$ for given r. It can be shown that the condition $Q_g(\underline{\tau}_g(r))\geq 0$ is equivalent to $\widetilde{B}(\tau)\geq b_0$. Therefore, in the case of $\alpha>\overline{\alpha}$, the budget constraint binds at the equilibrium $\tau=\tau^\dagger$ iff $\widetilde{B}(\tau^\dagger)\geq b_0$, which can be rewritten as $\tau^\dagger\geq\tau_1^b$ given the earlier observation that $\widetilde{B}(\tau)-b_0$ crosses zero exactly once at τ_1^b from below.

Given $\alpha > \overline{\alpha}$, consider a relaxed problem with the budget constraint (and the welfare constraint) dropped. Then SE's problem from (4) can be rewritten as $\max_{\tau \geq \max\{\tau_1^s, \underline{\tau}_1\}} \gamma W(\tau) + (1-\gamma)\widetilde{\Pi}(\tau)$, from our earlier observation that the surplus and budget constraints $\widetilde{\Pi}(\tau) \geq 0$ and $\widetilde{B}(\tau) \leq b_0$ can be together rewritten as $\max\{\tau_1^s, \underline{\tau}_1\} \leq \tau \leq \tau_1^b$. If the solution of this problem $\widetilde{\tau}_1$ satisfies $\widetilde{B}(\widetilde{\tau}_1) < b_0$ (or equivalently $\widetilde{\tau}_1 < \tau_1^b$), then the budget constraint indeed does not bind, as stated in part (b) of the proposition. If $\widetilde{B}(\widetilde{\tau}_1) \geq b_0$ (or equivalently $\widetilde{\tau}_1 \geq \tau_1^b$), on the other hand, the budget constraint binds. The solution of this new problem is found by substituting the condition $B(r,\tau) = b_0$ in $(\mathcal{P}_1)$, which becomes

$$\max_{\tau \geq \tau_1^b} \frac{\gamma}{1-\gamma} vA(\tau) - C(\tau) - R(\tau) \text{ s.t. } C(\tau) + R(\tau) \leq b_0 - \psi_1, \tag{8}$$

where the lower bound $\tau \geq \tau_1^b$ ensures that the equilibrium $\tau = \tau^\dagger$ indeed occurs at the binding budget constraint, as we showed above. The constraint $C(\tau) + R(\tau) \leq b_0 - \psi_1$ in (8) represents SE's surplus constraint from (4) given that the government's budget constraint binds, and it was established above that this can be rewritten as $\tau \geq \underline{\tau}_1$. The properties of the quasiconcave objective function in (8) are found in Lemma A5 with a substitution $\theta = \frac{\gamma}{1-\gamma} v$; given the condition $\frac{kq}{Lp+lq} > \frac{K}{\mu}$ stated in Assumption 1, there exists a unique solution $\tau = \widehat{\tau}(\gamma)$ to $Q_S(\tau) = 0$ that maximizes this objective, where $Q_S(\tau) \equiv \left(\frac{\gamma}{1-\gamma} v(Lp+lq) + kq\right) S(\tau) - K(Lp+lq) h(\tau) - K$. Therefore, the solution to (8) without the condition $\tau \geq \tau_1^b$ is equal to $\max\{\underline{\tau}_1, \widehat{\tau}(\gamma)\}$. Incorporating $\tau \geq \tau_1^b$, we see that the solution is $\tau^\dagger = \max\{\tau_1^b, \underline{\tau}_1, \widehat{\tau}(\gamma)\}$, as stated in part (c) of the proposition. □

Proof (**Proposition 2**) Recall $W = vA(\tau)$, $\Pi = rA(\tau) - C(\tau) - R(\tau) - \psi_2$, and $B = rA(\tau)$ under RRC. Since $A'(\tau) < 0$ with $A(0) = \infty$ and $A(\infty) = \frac{\mu p}{(\mu+L)p+lq} > \frac{w_0}{v}$ from Assumption 1, the welfare constraint $W(\tau) = vA(\tau) \geq w_0$ never binds and it is dropped from the problem formulation. SE's problem $(\mathcal{P}_2)$ then becomes $\max_{r \geq 0, \tau \geq 0} \gamma vA(\tau) + (1-\gamma)(rA(\tau) - C(\tau) - R(\tau) - \psi_2)$ s.t. $rA(\tau) - C(\tau) - R(\tau) - \psi_2 \geq 0$ and $rA(\tau) \leq b_0$, from which we find that it is optimal to increase r to its maximum, i.e., when the budget constraint binds: $B = rA(\tau) = b_0$. Substituting this equation, we see that the problem is equivalent to $\max_{\tau \geq 0} \frac{\gamma}{1-\gamma} vA(\tau) - C(\tau) - R(\tau)$ s.t. $C(\tau) + R(\tau) \leq b_0 - \psi_2$. This problem is equivalent to (8) found in the proof of Proposition 1 except that the constraint $\tau \geq \tau_1^b$ is replaced by $\tau \geq 0$. The rest of the proof proceeds similarly as that of Proposition 1. □

Proof (**Lemma 2**) Let $\Gamma(\tau) \equiv C(\tau) + R(\tau)$, which is quasiconvex with $\Gamma(0) = \infty$ and $\Gamma(\infty) = \frac{Kp+kq}{(\mu+L)p+lq}$ with the unique minimum occurring at $\tau = \widehat{\tau}(0)$ that solves the equation $kqS(\tau) - K(Lp+lq) h(\tau) - K = 0$ (Corollary A1). Note $\underline{\tau}_1 \leq \underline{\tau}_2 < \widehat{\tau}(0)$, where $\underline{\tau}_1$ and $\underline{\tau}_2$ are the the smallest solutions of $\Gamma(\tau) = b_0 - \psi_1$ and $\Gamma(\tau) = b_0 - \psi_2$, respectively (Corollary A1). Recall τ_1^s and τ_1^b solves $\widetilde{\Pi}(\tau) = 0$ and $\widetilde{B}(\tau) = b_0$, respectively.

(a) In the limit $\gamma \to 0$, SE's objective becomes $\lim_{\gamma\to 0} \gamma W + (1-\gamma)\Pi = \Pi$. At the RRC equilibrium, this is equal to $\Pi = rA(\tau) - \Gamma(\tau) - \psi_2 = b_0 - \Gamma(\tau) - \psi_2$ as the budget constraint $B(r,\tau) = rA(\tau) \leq b_0$ binds. The RRC equilibrium occurs at $\tau^\ddagger = \max\{\underline{\tau}_2, \widehat{\tau}(0)\} = \widehat{\tau}(0)$ that minimizes $\Gamma(\tau)$ in the interior, at which Π is equal to $\Pi_2 \equiv b_0 - \Gamma(\tau^\ddagger) - \psi_2$ (Proposition 2 and Corollary A3). Under ROC, on the other hand, the equilibrium is found at either (i) $\tau^\dagger = \infty$ or (ii) $\tau^\dagger = \max\{\tau_1^b, \underline{\tau}_1, \widehat{\tau}(0)\} = \max\{\tau_1^b, \widehat{\tau}(0)\} \geq \tau_1^b$ or (iii) $\tau^\dagger = \widetilde{\tau}_1 < \tau_1^b$ that maximizes $\widetilde{\Pi}(\tau) = \rho(\tau)A(\tau) - C(\tau) - \psi_1$ in the interval $(\max\{\tau_1^s, \underline{\tau}_1\}, \tau_1^b)$, which exists iff $\underline{\tau}_1 < \tau_1^b$. The budget constraint binds in the first two cases only (Proposition 1 and Corollary A2). The first result $\Gamma(\tau^\dagger) \geq \Gamma(\tau^\ddagger)$ follows directly from the fact that $\Gamma(\tau)$ is minimized at $\tau = \widehat{\tau}(0)$, which is equal to $\tau^\ddagger$ but may not coincide with $\tau^\dagger$. Suppose that the budget constraint binds under ROC. Then $\Pi(r,\tau) = rA(\tau) - C(\tau) - \psi_1$ together with $B(r,\tau) = rA(\tau) + R(\tau) = b_0$ under ROC implies $\Pi = b_0 - \Gamma(\tau) - \psi_1$. At the ROC equilibrium $\tau^\dagger$, this is equal to $\Pi_1 \equiv b_0 - \Gamma(\tau^\dagger) - \psi_1$. Comparing Π_1 with Π_2, we see that $\Pi_1 < \Pi_2$ iff $\Gamma(\tau^\dagger) - \Gamma(\tau^\ddagger) > \psi_2 - \psi_1$. Suppose that the budget constraint does not bind under ROC. According to Proposition 2, this happens iff $\tau^\dagger = \widetilde{\tau}_1 < \tau_1^b$ at which $\widetilde{B}(\widetilde{\tau}_1) = \rho(\widetilde{\tau}_1)A(\widetilde{\tau}_1) + R(\widetilde{\tau}_1) < b_0$, which implies $\widetilde{\Pi}_1 \equiv \widetilde{\Pi}(\widetilde{\tau}_1) = \rho(\widetilde{\tau}_1)A(\widetilde{\tau}_1) - C(\widetilde{\tau}_1) - \psi_1 < b_0 - \Gamma(\widetilde{\tau}_1) - \psi_1$. Comparing $\widetilde{\Pi}_1$ with Π_2, we see that $\widetilde{\Pi}_1 < \Pi_2$ if $\Gamma(\tau^\dagger) - \Gamma(\tau^\ddagger) > \psi_2 - \psi_1$. Hence, $\Pi_1 < \Pi_2$ if $\Gamma(\tau^\dagger) - \Gamma(\tau^\ddagger) > \psi_2 - \psi_1$ regardless of whether the budget constraint binds or not.

(b) In the limit $\gamma \to 1$, $\widehat{\tau}(1) = 0$ and SE's objective becomes $\lim_{\gamma\to 1} \gamma W + (1-\gamma)\Pi = W = vA$. At the RRC equilibrium, this is equal to $W_2 \equiv vA(\tau^\ddagger)$ where the RRC equilibrium $\tau^\ddagger$ is found at $\tau^\ddagger = \max\{\underline{\tau}_2, \widehat{\tau}(1)\} = \underline{\tau}_2$, at which the budget and surplus constraints both bind and satisfies $\Gamma(\tau^\ddagger) = b_0 - \psi_2$ and $\Gamma'(\tau^\ddagger) < 0$ (Proposition 2, Corollary A3). Under ROC, on the other hand, the equilibrium is found at either (i) $\tau^\dagger = \infty$, at which the budget constraint binds but the surplus constraint does not, or (ii) $\tau^\dagger = \max\{\tau_1^s, \underline{\tau}_1\}$, at which the surplus constraint binds while the budget constraint binds iff $\tau_1^s < \underline{\tau}_1 = \tau^\dagger$. The first and second cases arise iff $\alpha \leq \overline{\alpha}$ and $\alpha > \overline{\alpha}$, respectively (Proposition 1, Corollary A2). Either case, the objective function at the equilibrium is $W_1 \equiv vA(\tau^\dagger)$. Suppose $\alpha \leq \overline{\alpha}$. Then $W_1 = vA(\tau^\dagger) = vA(\infty) < vA(\tau^\ddagger) = W_2$ since $A'(\tau) < 0$; hence, RRC is preferred to ROC. Suppose $\alpha > \overline{\alpha}$. Since the surplus constraint binds while the budget constraint may not under ROC, $\widetilde{\Pi}(\tau^\dagger) = \rho(\tau^\dagger)A(\tau^\dagger) - C(\tau^\dagger) - \psi_1 = 0$ and $\widetilde{B}(\tau^\dagger) = \rho(\tau^\dagger)A(\tau^\dagger) + R(\tau^\dagger) \leq b_0$, together implying $\Gamma(\tau^\dagger) \leq b_0 - \psi_1$. Combining this with the earlier result $\Gamma(\tau^\ddagger) = b_0 - \psi_2$, we find $\Gamma(\tau^\dagger) - \Gamma(\tau^\ddagger) \leq \psi_2 - \psi_1$. From this and the properties of $\Gamma(\tau)$ noted above with the earlier result $\Gamma'(\tau^\ddagger) < 0$, we find that both $\tau^\dagger \leq \tau^\ddagger$ and $\tau^\dagger > \tau^\ddagger$ are possible depending on the values of ψ_1 and ψ_2. RRC is preferred iff $\tau^\dagger > \tau^\ddagger$ since $W_1 = vA(\tau^\dagger) < vA(\tau^\ddagger) = W_2$. Moreover, $\tau^\dagger > \tau^\ddagger$ if $\Gamma(\tau^\dagger) < \Gamma(\tau^\ddagger)$ since $\Gamma(\tau)$ is quasiconvex and $\Gamma'(\tau^\ddagger) < 0$. □

Proof (**Proposition 3**) For brevity, refer to the budget constraint and the surplus constraint as BC and SC. Let $\Gamma(\tau) \equiv C(\tau) + R(\tau)$.

(a) According to Proposition 1(a), a small α satisfying $\alpha \leq \overline{\alpha}$ results in $\tau^\dagger = \infty$, at which BC binds and SC does not bind under ROC. On the other hand, according to Proposition 2 and Corollary A3, in the limit $\gamma \to 0$ the equilibrium under RRC is found at $\tau^\ddagger = \max\{\underline{\tau}_2, \widehat{\tau}(0)\} = \widehat{\tau}(0) < \infty$ at which BC binds and SC does not bind (Note $\underline{\tau}_2 < \widehat{\tau}(0)$ by the properties of $C(\tau) + R(\tau)$; see Corollary A1.) Hence, $\tau^\dagger > \tau^\ddagger$ with binding BC and non-binding SC under both ROC and RRC. It then follows from Lemma 2(a) that SE prefers ROC to RRC if $\Gamma(\tau^\dagger) - \Gamma(\tau^\ddagger) < \psi_2 - \psi_1$, which is satisfied for sufficiently large $\psi_2 - \psi_1$.

(b) According to Corollary A2(a), in the limit $\gamma \to 0$ under ROC, BC binds for a large α satisfying $\alpha > \overline{\alpha}$ while SC does not bind. It then follows from Proposition 1(c) that $\tau^\dagger = \max\{\tau_1^b, \underline{\tau}_1, \widehat{\tau}(0)\} = \max\{\tau_1^b, \widehat{\tau}(0)\}$ in this case (since $\underline{\tau}_1 < \widehat{\tau}(0)$, proved similarly as in part (a)). Under RRC, $\tau^\ddagger = \widehat{\tau}(0)$ as we found in part (a), with binding BC and non-binding SC. Hence, $\tau^\dagger = \max\{\tau_1^b, \widehat{\tau}(0)\} \geq \widehat{\tau}(0) = \tau^\ddagger$. If $\tau^\dagger = \tau_1^b > \widehat{\tau}(0) = \tau^\ddagger$, then it follows from Lemma 2(a) that SE prefers ROC to RRC if $\Gamma(\tau^\dagger) - \Gamma(\tau^\ddagger) < \psi_2 - \psi_1$, which is satisfied for sufficiently large $\psi_2 - \psi_1$. If $\tau^\dagger = \widehat{\tau}(0) > \tau_1^b$, on the other hand, $\Gamma(\tau^\dagger) - \Gamma(\tau^\ddagger) < \psi_2 - \psi_1$ is always satisfied since $\tau^\dagger = \tau^\ddagger = \widehat{\tau}(0)$ implies that the left-hand side of the condition is equal to zero.

(c) It follows from Lemma 2(b) that SE prefers RRC to ROC for a small α satisfying $\alpha \leq \overline{\alpha}$ and a large γ in the limit $\gamma \to 1$. In these parameter ranges, $\tau^\dagger = \infty$ (Proposition 1(a)) and $\tau^\ddagger = \max\{\underline{\tau}_2, \widehat{\tau}(\gamma)\} = \underline{\tau}_2 < \infty$ (Proposition 2), the latter from $\underline{\tau}_2 > 0 = \lim_{\gamma \to 1} \widehat{\tau}(\gamma)$. Hence, $\tau^\dagger > \tau^\ddagger$.

(d) From the proofs of Lemma 2(b) and Corollary A2(b) we see that, for a large α satisfying $\alpha > \overline{\alpha}$ and $\gamma \to 1$, $\tau^\dagger = \underline{\tau}_1$ under ROC and $\tau^\ddagger = \underline{\tau}_2$ under RRC. Since $\underline{\tau}_1 < \underline{\tau}_2$ (see proof of Lemma 2), $\tau^\dagger < \tau^\ddagger$. According to Lemma 2(b), this implies that SE prefers ROC to RRC. □

Proof (**Corollary 1**) Proof is found in the Online Appendix. □

Proof (**Proposition 4**) Following the discussions in §5.1, apply substitutions $v \to vn^a$, $C(\tau) \to nC(\tau)$, $R(\tau) \to nR(\tau)$, $\psi_1 \to n\psi_1$, $\psi_2 \to n\psi_2$, $b_0 \to nb_0$, and $w_0 \to nw_0$. Recall from the discussions that the ROC and RRC equilibria $\tau^\dagger$ and $\tau^\ddagger$ can be determined analogously to those from the single-vehicle problem with a rescaling $v \to v/n^{1-a}$. The SE's objective in the limit $\gamma \to 1$ is equal to $\lim_{\gamma \to 1} \gamma W + (1-\gamma)\Pi = W = vn^a A$. In this limit, the RRC equilibrium of the single-vehicle problem is found at $\tau = \tau^\ddagger = \underline{\tau}_2 < \infty$ that satisfies $C(\tau^\ddagger) + R(\tau^\ddagger) = b_0 - \psi_2$ (proof of Lemma 2(b)). Since $C(\tau)$ and $R(\tau)$ are independent of v (or the rescaled value v/n^{1-a}; see (1)), the RRC equilibrium is found at $\tau = \tau^\ddagger = \underline{\tau}_2$ for any n. At this value, SE's objective is equal to $W_2 \equiv vn^a A(\tau^\ddagger)$. Consider $\alpha \to 0$. From Proposition 1(a) we see that the

ROC equilibrium of the single-vehicle problem is found at $\tau = \tau^{\dagger} = \infty$, which also applies to any n. At this value, SE's objective is equal to $W_1 \equiv vn^a A(\tau^{\dagger})$. Since $\tau^{\ddagger} < \tau^{\dagger}$ with $A'(\tau) < 0$, we have $W_2 - W_1 = vn^a [A(\tau^{\ddagger}) - A(\tau^{\dagger})] > 0$, and moreover, $\frac{\partial}{\partial n}(W_2 - W_1) > 0$ since neither $\tau^{\dagger}$ nor $\tau^{\ddagger}$ depends on n. □

Online Appendix to: "Improving Social Benefits Through Vehicle Maintenance Contracting: The Role of Social Enterprise"

Lemma A1 *The function*

$$S(t) \equiv h(t) M(t) - (1/p) F(t), \tag{A1}$$

has the following properties: $S'(t) = h'(t) M(t) > 0$ *with* $S(0) = 0$ *and* $S(\infty) = \infty$. *Consequently,* $S(t) \geq 0$. *Moreover,* $\frac{\partial S(t)}{\partial p} < 0$ *for any* $t > 0$.

Proof Recall $H(t) = \int_0^t h(z)\,dz$ and $\overline{F}(t) = \exp(-pH(t))$ and the assumptions $h(0) = 0$, $h(\infty) = \infty$, and $h'(t) > 0$. Clearly, $S(0) = 0$ and $S(\infty) = \infty$. Using the identity $\overline{F}'(t) = -ph(t)\overline{F}(t)$ we obtain $S'(t) = h'(t) M(t) > 0$ for $\tau > 0$. The results $S(0) = 0$ and $S'(t) > 0$ together imply $S(t) > 0$ for all $t > 0$. Differentiating $S(t)$ with respect to p yields $\frac{\partial S(t)}{\partial p} = \frac{\omega(t)}{p^2}$ where $\omega(t) \equiv 1 - \overline{F}(t) - pH(t)\overline{F}(t) - p^2 h(t) \int_0^t H(z)\overline{F}(z)\,dz$. Since $h(0) = H(0) = 0$ and $\overline{F}(0) = 1$, we have $\omega(0) = 0$. Moreover, using the identities $H'(t) = h(t)$ and $\overline{F}'(t) = -ph(t)\overline{F}(t)$ we obtain $\omega'(t) = -p^2 h'(t) \int_0^t H(z)\overline{F}(z)\,dz < 0$. The results $\omega(0) = 0$ and $\omega'(t) < 0$ together imply $\omega(t) < 0$ for all $t > 0$. Therefore, $\frac{\partial S(t)}{\partial p} = \frac{\omega(t)}{p^2} < 0$ for $t > 0$. □

Lemma A2 *Long-run average availability, repair cost, and replacement cost are evaluated as* $A(\tau) = \frac{M(\tau)}{T(\tau)}$, $C(\tau) = \frac{(kq/p)F(\tau)}{T(\tau)}$, *and* $R(\tau) = \frac{K}{T(\tau)}$, *where* $T(\tau) \equiv M(\tau) + (L + lq/p) F(\tau)$.

Proof The expected length of a vehicle replacement cycle consists of three components: (i) expected vehicle age at the time of replacement; (ii) expected cumulative repair downtimes until replacement; (iii) expected downtime after an unscheduled replacement. With finite retirement age $\tau < \infty$, a vehicle is replaced at age $\min\{Y, \tau\}$, where Y denotes vehicle age at the time of first major failure conditional on no vehicle retirement ($\tau = \infty$). Hence, the expected replacement age is equal to $E[\min\{Y, \tau\}] = \int_0^\tau \overline{F}(t)\,dt = M(\tau)$, which also represents the expected vehicle uptime in a single cycle. This is the first component of the expected cycle length. The second component, expected cumulative repair downtimes until replacement, is equal to $l \times N(\tau)$ where l is the expected repair lead time and $N(\tau)$ is the expected number of minor failures until replacement at vehicle age $\min\{Y, \tau\}$. The latter is evaluated as $N(\tau) = \int_0^\tau qh(t) \Pr(Y > t)\,dt = (q/p)F(\tau)$ (see Beichelt 2006, pp. 138-140). Finally the last component, expected downtime after an unscheduled replacement, is equal to $L \times \Pr(Y < \tau) = LF(\tau)$ where L is the expected replacement lead time. Adding up the three components yields the expected cycle length $T(\tau) = M(\tau) + \left(L + \frac{lq}{p}\right) F(\tau)$. It then follows that $A(\tau) = \frac{M(\tau)}{T(\tau)}$, the expected vehicle uptime in a cycle, $C(\tau) = \frac{kN(\tau)}{T(\tau)} = \frac{(kq/p)F(\tau)}{T(\tau)}$, fixed repair cost k times the expected number of repairs in a cycle, and $R(\tau) = \frac{K}{T(\tau)}$, fixed replacement cost K times the number of replacements in a cycle, which is exactly one because

each renewal cycle ends with a replacement. □

Lemma A3 *$A(\tau)$, $C(\tau)$, and $R(\tau)$ defined in (1) satisfy the following: (a) $A'(\tau)<0$ with $A(0)=1$ and $A(\infty)=\frac{\mu p}{(\mu+L)p+lq}$; (b) $C'(\tau)>0$ with $C(0)=0$ and $C(\infty)=\frac{kq}{(\mu+L)p+lq}$; (c) $R'(\tau)<0$ with $R(0)=\infty$ and $R(\infty)=\frac{Kp}{(\mu+L)p+lq}$.*

Proof Differentiating $A(\tau)$, $C(\tau)$, and $R(\tau)$ in (1) using the identity $\overline{F}'(\tau)=-ph(\tau)\overline{F}(\tau)$ yields $A'(\tau)=-\frac{(Lp+lq)\overline{F}(\tau)S(\tau)}{T(\tau)^2}$, $C'(\tau)=\frac{kq\overline{F}(\tau)S(\tau)}{T(\tau)^2}$, and $R'(\tau)=-\frac{K\overline{F}(\tau)[1+(Lp+lq)h(\tau)]}{T(\tau)^2}$, where $S(\tau)$ and $T(\tau)$ are defined in (A1) and (2). Since $S(\tau)>0$ (see Lemma A1), it follows that $A'(\tau)<0$, $C'(\tau)>0$, and $R'(\tau)<0$. It is straightforward to evaluate the limiting values at $\tau=0$ and $\tau\to\infty$. □

Lemma A4 *(i) $\lim_{p\to 0}\overline{F}(\tau)=1$; (ii) $\lim_{p\to 0}M(\tau)=\tau$; (iii) $\lim_{p\to 0}\frac{F(\tau)}{p}=H(\tau)$; (iv) $\lim_{p\to 0}S(\tau)=\tau h(\tau)-H(\tau)$; (v) $\lim_{p\to 0}T(\tau)=\tau+lqH(\tau)$; (vi) $\lim_{p\to 0}A(\tau)=\frac{\tau}{\tau+lqH(\tau)}$; (vii) $\lim_{p\to 0}C(\tau)=\frac{kqH(\tau)}{\tau+lqH(\tau)}$; (viii) $\lim_{p\to 0}R(\tau)=\frac{K}{\tau+lqH(\tau)}$.*

Proof The results follow from $\lim_{p\to 0}\overline{F}(t)=\lim_{p\to 0}\exp(-pH(t))\to 1$. Note $\lim_{p\to 0}\frac{F(\tau)}{p}=\lim_{p\to 0}\frac{\partial F(\tau)}{\partial p}=\lim_{p\to 0}H(\tau)\overline{F}(\tau)=H(\tau)$. □

Lemma A5 *Let $\beta_1\equiv\frac{kq}{Lp+lq}$ and*

$$Q(\tau)\equiv(\theta(Lp+lq)+kq)S(\tau)-K(Lp+lq)h(\tau)-K. \tag{A2}$$

Recall $A(\tau)$, $C(\tau)$, $R(\tau)$ defined in (1). The function $U(\tau)\equiv\theta A(\tau)-C(\tau)-R(\tau)$ is quasiconcave, starting from $U(0)=-\infty$ and converging to $U(\infty)=\frac{(\theta\mu-K)p-kq}{(\mu+L)p+lq}$. Moreover:

(a) If $\theta+\beta_1\le\frac{K}{\mu}$, then $U'(\tau)>0$ for all $\tau\ge 0$.

(b) If $\theta+\beta_1>\frac{K}{\mu}$, then $U(\tau)$ peaks at $\tau=\widehat{\tau}$ where $\widehat{\tau}$ is a unique root of the function $Q(\tau)$ that satisfies $Q'(\widehat{\tau})>0$.

Proof We first prove the following properties of $Q(\tau)$ defined in (A2): *(i) If $\theta+\beta_1\le\frac{K}{\mu}$, then $Q(\tau)<0$ and $Q'(\tau)<0$ for all $\tau\ge 0$; (ii) If $\theta+\beta_1>\frac{K}{\mu}$, then $Q(\tau)$ crosses zero exactly once at $\tau=\widehat{\tau}>0$ that satisfies $Q(\widehat{\tau})=0$ and $Q'(\widehat{\tau})>0$.* Using the definition of $S(\tau)$ in (A1), we can rewrite $Q(\tau)$ as $Q(\tau)=\chi(\tau)h(\tau)-(\theta L+(\theta l+k)q/p)F(\tau)-K$ where $\chi(\tau)\equiv(\theta Lp+(\theta l+k)q)M(\tau)-K(Lp+lq)$. Using the relation $\overline{F}'(t)=-ph(t)\overline{F}(t)$, we obtain $Q'(\tau)=\chi(\tau)h'(\tau)$. Since $h'(\tau)>0$, we see that the sign of $Q'(\tau)$ is equal to the sign of $\chi(\tau)$. Observe $\chi'(\tau)=(\theta Lp+(\theta l+k)q)\overline{F}(\tau)>0$ with $\chi(0)=-K(Lp+lq)<0$ and $\chi(\infty)=(Lp+lq)\mu\left(\theta+\beta_1-\frac{K}{\mu}\right)$, where $\mu=M(\infty)$. From the

latter expression, we see that $\chi(\infty) > 0$ iff $\theta + \beta_1 > \frac{K}{\mu}$. Suppose $\theta + \beta_1 \leq \frac{K}{\mu}$. Then $\chi(\tau) < 0$ for all $\tau \geq 0$ and therefore $Q(\tau) < 0$ and $Q'(\tau) < 0$. Suppose $\theta + \beta_1 > \frac{K}{\mu}$. Then $\chi(\tau)$ crosses zero exactly once from below at $\tau = \tau^o \equiv M^{-1}\left(\frac{K(Lp+lq)}{\theta Lp+(\theta l+k)q}\right) > 0$, which implies $Q'(\tau) < 0$ for $\tau < \tau^o$ and $Q'(\tau) > 0$ for $\tau > \tau^o$, i.e., $Q(\tau)$ is a quasiconvex function with a unique interior minimum occurring at $\tau = \tau^o$. Combining this with $Q(0) = -K < 0$ and $Q(\infty) = \chi(\infty) h(\infty) - (\theta L + (\theta l + k) q/p) - K = \infty > 0$, we conclude that $Q(\tau)$ crosses zero exactly once at $\tau = \widehat{\tau} > \tau^o$ such that $Q(\widehat{\tau}) = 0$ and $Q'(\widehat{\tau}) > 0$.

Rewriting $U(\tau) \equiv \theta A(\tau) - C(\tau) - R(\tau)$ using (1) yields $U(\tau) = \frac{\theta M(\tau) - (kq/p)F(\tau) - K}{M(\tau) + (L + lq/p)F(\tau)}$, where $M(\tau) = \int_0^\tau \overline{F}(t)\, dt$. Taking the limit $\tau \to 0$ that results in $H(0) = F(0) = 0$ and substituting them in $U(\tau)$, we find $U(0) = -\infty$. Taking the limit $\tau \to \infty$ that results in $H(\infty) = \infty$, $F(\infty) = 1$, and $M(\infty) = \mu$, we find $U(\infty) = \frac{(\theta\mu - K)p - kq}{(\mu + L)p + lq}$. Differentiating $U(\tau)$ using the identity $\overline{F}'(\tau) = -ph(\tau)\overline{F}(\tau)$ yields $U'(\tau) = -\frac{\overline{F}(\tau) Q(\tau)}{T(\tau)^2}$. The statements of the lemma follow from the properties of $Q(\tau)$ proved above. □

Corollary A1 (Corollary of Lemma A5) *Let* $\beta_1 \equiv \frac{kq}{Lp+lq}$ *and* $\beta_2 \equiv \frac{Kp+kq}{(\mu+L)p+lq}$. *The function* $\Gamma(\tau) \equiv C(\tau) + R(\tau)$ *is quasiconvex with* $\Gamma(0) = \infty$ *and* $\Gamma(\infty) = \beta_2$. *If* $\beta_1 \leq \frac{K}{\mu}$, *then* $\beta_1 \leq \beta_2 \leq \frac{K}{\mu}$ *and* $\Gamma'(\tau) < 0$ *for all* $\tau \geq 0$. *If* $\beta_1 > \frac{K}{\mu}$, *then* $\frac{K}{\mu} < \beta_2 < \beta_1 < \frac{k}{l}$ *and* $\Gamma(\tau)$ *has a unique minimum at* $\tau = \widehat{\tau}_0 > 0$ *that solves the equation* $kqS(\tau) - K(Lp + lq)h(\tau) - K = 0$, *resulting in* $\Gamma(\widehat{\tau}_0) = \frac{kq}{1/h(\widehat{\tau}_0) + Lp + lq} = \frac{Kh(\widehat{\tau}_0)}{S(\widehat{\tau}_0)}$ *that satisfies* $\frac{K}{\mu} < \Gamma(\widehat{\tau}_0) < \beta_2$. *In the latter case:*

(a) $\widehat{\tau}_\theta \leq \widehat{\tau}_0$, *where* $\widehat{\tau}_\theta$ *is the root of* $Q(\tau) = (\theta(Lp + lq) + kq)S(\tau) - K(Lp + lq)h(\tau) - K$ *with* $\theta \geq 0$.

(b) $\tau_m < \widehat{\tau}_0$, *where* τ_m *uniquely solves the equation* $M(\tau) + LF(\tau) = \frac{Kl}{k}$ *and satisfies* $\Gamma(\tau_m) = \frac{k}{l}$.

Proof The statement in the lemma follows from Lemma A5 after setting $\theta = 0$. The inequalities $\beta_1 \leq \beta_2 \leq \frac{K}{\mu}$ and $\frac{K}{\mu} < \beta_2 < \beta_1 < \frac{k}{l}$ follow from the fact that $\frac{A+B}{C+D} < \frac{B}{D}$ iff $\frac{A}{C} < \frac{B}{D}$. The minimum value $\Gamma(\widehat{\tau}_0)$ at $\tau = \widehat{\tau}_0$ is evaluated by rewriting the optimality condition $kqS(\tau) - K(Lp + lq)h(\tau) - K = 0$ as $K = \frac{kqS(\widehat{\tau}_0)}{1 + (Lp+lq)h(\widehat{\tau}_0)}$ and substituting it in $\Gamma(\tau) = \frac{(kq/p)F(\tau) + K}{M(\tau) + (L + lq/p)F(\tau)}$ (see (1)). The inequalities $\frac{K}{\mu} < \Gamma(\widehat{\tau}_0) = \frac{K}{S(\widehat{\tau}_0)/h(\widehat{\tau}_0)} < \beta_2$ follow from the facts that $\frac{S(t)}{h(t)} = M(t) - \frac{F(t)}{ph(t)}$ increases with slope $\frac{h'(t)F(t)}{ph(t)^2} > 0$ with $\lim_{t\to\infty} \frac{S(t)}{h(t)} = \mu$ and that $\Gamma(\tau)$ is quasiconvex with $\Gamma(\infty) = \beta_2$.

(a) From Lemma A5 we infer that $Q(\tau) \geq 0$ iff $\tau \geq \widehat{\tau}_\theta$. Since $Q(\widehat{\tau}_0) = (\theta(Lp + lq) + kq)S(\widehat{\tau}_0) - K(Lp + lq)h(\widehat{\tau}_0) - K = \theta(Lp + lq)S(\widehat{\tau}_0) \geq 0$, we have $\widehat{\tau}_0 \geq \widehat{\tau}_\theta$.

(b) Substituting $M(\tau_m) + LF(\tau_m) = \frac{Kl}{k}$ in $\Gamma(\tau)$ yields $\Gamma(\tau_m) = \frac{k}{l}$. The inequality $\Gamma(\infty) = \beta_2 < \frac{k}{l} = \Gamma(\tau_m)$ proved above implies $\Gamma'(\tau_m) < 0$ since $\Gamma(\tau)$ is quasiconvex with $\Gamma(0) = \infty$ and $\Gamma(\infty) = \beta_2$. Hence, $\tau_m < \widehat{\tau}_0$. □

Lemma A6 (Government's optimal response under ROC) *Let $\underline{\tau}_g(r)$ be the unique solution to the binding budget constraint $rA(\tau)+R(\tau)=b_0$ from (4) that exists if and only if $r<\frac{b_0(\mu+L+lq/p)-K}{\mu}$. Let $\widehat{\tau}_g(r)$ be the unique root of $Q_g(\tau)\equiv(Lp+lq)\left[\left(\frac{\alpha}{1-\alpha}v-r\right)S(\tau)-Kh(\tau)\right]-K$ that exists if $\frac{\alpha}{1-\alpha}v-r>\frac{K}{\mu}$; otherwise, set $\widehat{\tau}_g(r)=\infty$. A feasible solution to the government's subproblem (4) under ROC exists if and only if $\underline{\tau}_g(r)\leq\overline{\tau}_g$ where $\overline{\tau}_g=A^{-1}\left(\frac{w_0}{v}\right)$ if $\frac{\mu p}{(\mu+L)p+lq}<\frac{w_0}{v}$ and $\overline{\tau}_g=\infty$ if $\frac{w_0}{v}\leq\frac{\mu p}{(\mu+L)p+lq}$. Given r and a nonempty interval $[\underline{\tau}_g(r),\overline{\tau}_g]$, the government sets optimal $\tau=\tau_G(r)$ as follows: (i) $\tau_G(r)=\underline{\tau}_g(r)$ if $\widehat{\tau}_g(r)\leq\underline{\tau}_g(r)<\overline{\tau}_g$; (ii) $\tau_G(r)=\widehat{\tau}_g(r)$ if $\underline{\tau}_g(r)<\widehat{\tau}_g(r)<\overline{\tau}_g$; (iii) $\tau_G(r)=\overline{\tau}_g$ if $\underline{\tau}_g(r)<\overline{\tau}_g\leq\widehat{\tau}_g(r)$.*

Proof With $W(\tau)=vA(\tau)$ and $B(r,\tau)=rA(\tau)+R(\tau)$ under ROC, the government's subproblem (4) from $(\mathcal{P}_1)$ is equivalent to $\max_{\tau\geq0}U_g(\tau)$ s.t. $rA(\tau)+R(\tau)\leq b_0$ and $vA(\tau)\geq w_0$, where $U_g(\tau)\equiv\left(\frac{\alpha}{1-\alpha}v-r\right)A(\tau)-R(\tau)$. Since $rA'(\tau)+R'(\tau)<0$ with $rA(0)+R(0)=\infty$ and $rA(\infty)+R(\infty)=\frac{(r\mu+K)p}{(\mu+L)p+lq}$ (see Lemma A3), the constraint $rA(\tau)+R(\tau)\leq b_0$ can be written as $\tau\geq\underline{\tau}_g(r)$ where $\underline{\tau}_g(r)$ uniquely solves $rA(\tau)+R(\tau)=b_0$ provided that $r<\frac{b_0(\mu+L+lq/p)-K}{\mu}\equiv\overline{r}$; otherwise, a feasible solution does not exist. Note $\underline{\tau}_g'(r)>0$ with $\underline{\tau}_g(0)=R^{-1}(b_0)$ and $\lim_{r\to\overline{r}}\underline{\tau}_g(r)=\infty$. Similarly, since $A'(\tau)<0$ with $A(0)=1$ and $A(\infty)=\frac{\mu p}{(\mu+L)p+lq}$, the constraint $vA(\tau)\geq w_0$ can be written as $\tau\leq A^{-1}\left(\frac{w_0}{v}\right)$ if $\frac{\mu p}{(\mu+L)p+lq}<\frac{w_0}{v}$; otherwise, the constraint is always satisfied. Together, these observations imply that a feasible solution to (4) exists iff $0<\underline{\tau}_g(r)\leq\overline{\tau}_g$, where $\overline{\tau}_g$ is defined in the lemma. Comparing $U_g(\tau)$ with $U(\tau)=\theta A(\tau)-C(\tau)-R(\tau)$ discussed in Lemma A5, we see that $U_g(\tau)$ is a special case of $U(\tau)$ with $\theta=\frac{\alpha}{1-\alpha}v-r$ and $k=0$ (which leads to $C(\tau)=0$; see (1)). Thus the results from Lemma A5 apply with the substitutions: (i) $U_g'(\tau)>0$ for all $\tau\geq0$ if $\frac{\alpha}{1-\alpha}v-r\leq\frac{K}{\mu}$ and (ii) $U_g'(\tau)>0$ for $\tau<\widehat{\tau}_g$ and $U_g'(\tau)<0$ for $\tau>\widehat{\tau}_g$ where $\widehat{\tau}_g>0$ is a unique root of $Q_g(\tau)\equiv(Lp+lq)\left[\left(\frac{\alpha}{1-\alpha}v-r\right)S(\tau)-Kh(\tau)\right]-K$ otherwise. Therefore, the maximum of $U_g(\tau)$ occurs either at the boundaries of the interval $[\underline{\tau}_g(r),\overline{\tau}_g]$ or at the interior point $\widehat{\tau}_g$ as stated in the lemma. □

Corollary A2 (Corollary of Proposition 1) *Suppose $\alpha>\overline{\alpha}$. At the ROC equilibrium described in Proposition 1(b) and (c):*

(a) In the limit $\gamma\to0$, the budget constraint may or may not bind while the surplus constraint does not bind. The budget constraint binds for sufficiently large α.

(b) In the limit $\gamma\to1$, the budget constraint may or may not bind while the surplus constraint binds. The budget constraint binds for sufficiently large α.

Proof (a) Note $\underline{\tau}_1$ and $\widehat{\tau}(0)$, the solution to $C(\tau)+R(\tau)=b_0-\psi_1$ and the unique minimizer of $C(\tau)+R(\tau)$, respectively, satisfy $\underline{\tau}_1<\widehat{\tau}(0)$ (Corollary A1). The budget and surplus constraints

both bind in equilibrium iff $\tau^\dagger = \underline{\tau}_1$ (proof of Proposition 1). Following the proof of Proposition 1, we see that in the limit $\gamma \to 0$, SE's problem without the budget constraint becomes $\max_{\tau \geq \max\{\tau_1^s, \underline{\tau}_1\}} \widetilde{\Pi}(\tau)$, whose solution is $\tau = \widetilde{\tau}_1$. Suppose that the budget constraint does not bind, which requires $\underline{\tau}_1 < \widetilde{\tau}_1 < \tau_1^b$ (Proposition 1(b)). Since $\widetilde{\Pi}(\tau)$ satisfies $\widetilde{\Pi}(\tau_1^s) = 0$ and $\widetilde{\Pi}'(\tau_1^s) > 0$, given the assumption that $\widetilde{\Pi}(\tau)$ crosses zero at most once, we have $\widetilde{\tau}_1 > \tau_1^s$, i.e., the surplus constraint $\widetilde{\Pi}(\tau) \geq 0$ does not bind. Suppose that the budget constraint binds. In this case (Proposition 1(c)) the solution is $\max\{\tau_1^b, \underline{\tau}_1, \widehat{\tau}(0)\} = \max\{\tau_1^b, \widehat{\tau}(0)\}$ since $\underline{\tau}_1 < \widehat{\tau}(0)$ as proved above. This implies that the surplus constraint does not bind in this case, either, since the budget and surplus constraints both bind iff $\tau = \underline{\tau}_1$ in equilibrium, as we noted above. Therefore, the surplus constraint does not bind regardless of whether the budget constraint binds.

Note $\rho(\tau)$ defined in (5) satisfies $\frac{\partial}{\partial \alpha}\rho(\tau) > 0$ with $\lim_{\alpha \to 1} \rho(\tau) = \infty$. This implies that $\frac{\partial}{\partial \alpha}\tau_1^s < 0$ and $\frac{\partial}{\partial \alpha}\tau_1^b < 0$ with $\lim_{\alpha \to 1} \tau_1^s = \lim_{\alpha \to 1} \tau_1^b = 0$ since $\widetilde{\Pi}(\tau) = \rho(\tau) A(\tau) - C(\tau) - \psi_1$ crosses zero at $\tau = \tau_1^s$ from below with $\widetilde{\Pi}(0) = -\infty$ and $\widetilde{\Pi}'(\tau_1^s) > 0$, and similarly, $\widetilde{B}(\tau) = \rho(\tau) A(\tau) + R(\tau)$ crosses b_0 at $\tau = \tau_1^b$ from below with $\widetilde{B}(0) = -\infty$ and $\widetilde{B}'(\tau_1^b) > 0$ (proof of Proposition 1). These results imply that, for sufficiently large α, the maximum of $\widetilde{\Pi}(\tau)$ occurs in the region $\tau \geq \tau_1^b$, i.e., $\widetilde{\tau}_1 \geq \tau_1^b$ or equivalently $\widetilde{B}(\widetilde{\tau}_1) \geq b_0$, which corresponds to Proposition 1(c). Hence, the budget constraint binds in this case.

(b) Following the proof of Proposition 1, we see that in the limit $\gamma \to 1$, SE's problem without the budget constraint becomes $\max_{\tau \geq \max\{\tau_1^s, \underline{\tau}_1\}} vA(\tau)$, whose solution is found at the lower bound $\tau^\dagger = \max\{\tau_1^s, \underline{\tau}_1\}$ since $A'(\tau) < 0$ (Lemma A3). If $\tau^\dagger = \tau_1^s$, then $\underline{\tau}_1 \leq \tau^\dagger = \tau_1^s < \tau_1^b$ and therefore the surplus constraint binds but the budget constraint does not (Proposition 1(b)). If $\tau^\dagger = \underline{\tau}_1 > \tau_1^s$, then the budget and surplus constraints both bind as we noted above. By the same argument as in part (a), $\tau_1^s \to 0$ in the limit $\alpha \to 1$ and therefore $\tau_1^s < \underline{\tau}_1$ for sufficiently large α; in this case, $\tau^\dagger = \max\{\tau_1^s, \underline{\tau}_1\} = \underline{\tau}_1$ at which the budget constraint binds. □

Corollary A3 (Corollary of Proposition 2) *At the RRC equilibrium described in Proposition 2:*

(a) In the limit $\gamma \to 0$, the budget constraint binds while the surplus constraint does not bind.

(b) In the limit $\gamma \to 1$, both the budget and surplus constraints bind.

Proof The proof is similar to as that of Corollary A2 and is omitted. □

Corollary A4 (Corollary of Lemma 2) *If $\psi_2 - \psi_1 \to 0$, SE weakly prefers RRC to ROC for $\gamma \to 0$ and $\gamma \to 1$. Moreover, for $\alpha > \overline{\alpha}$:*

(a) $C(\tau^\dagger)+R(\tau^\dagger)\geq C(\tau^\ddagger)+R(\tau^\ddagger)$ *if* $\gamma\to 0$ *while* $C(\tau^\dagger)+R(\tau^\dagger)\leq C(\tau^\ddagger)+R(\tau^\ddagger)$ *if* $\gamma\to 1$.

(b) Either $W(\tau^\dagger)\leq W(\tau^\ddagger)$ *or* $W(\tau^\dagger)>W(\tau^\ddagger)$ *if* $\gamma\to 0$ *while* $W(\tau^\dagger)\leq W(\tau^\ddagger)$ *if* $\gamma\to 1$.

Proof Omitted. □

Proof (**Corollary 1**) Note that preference for RRC remains true for $\psi_2-\psi_1$ smaller than the values considered in Proposition 3 (see Corollary A4 that proves this for $\psi_2-\psi_1\to 0$). Let $\Gamma(\tau)\equiv C(\tau)+R(\tau)$. It is shown in the proof of Proposition 3(c) that $\tau^\dagger=\infty>\underline{\tau}_2=\tau^\ddagger$ for $\alpha\to 0$ satisfying $\alpha\leq\overline{\alpha}$ and $\gamma\to 1$, where $\underline{\tau}_2$ is the smallest solution to $\Gamma(\tau)=b_0-\psi_2$ as defined in Proposition 2. The result $\tau^\dagger>\tau^\ddagger$ in turn implies $W(\tau^\dagger)<W(\tau^\ddagger)$, $C(\tau^\dagger)>C(\tau^\ddagger)$, and $R(\tau^\dagger)<R(\tau^\ddagger)$ since $W'(\tau)=vA'(\tau)<0$, $C'(\tau)>0$, and $R'(\tau)<0$ (Lemma A3). In the limit $\psi_2-\psi_1\to 0$, $\Gamma(\tau^\dagger)=\Gamma(\infty)<\Gamma(\underline{\tau}_2)=\Gamma(\tau^\ddagger)$ since (i) $\Gamma(\tau)$ is a quasiconvex function with a unique interior minimum that starts from $\Gamma(0)=\infty$ and converges to $\Gamma(\infty)=\frac{Kp+kq}{(\mu+L)p+lq}$ (Corollary A1) and (ii) $b_0-\psi_2=b_0-\psi_1>\frac{Kp+kq}{(\mu+L)p+lq}=\Gamma(\infty)$ by Assumption 1. On the other hand, $\Gamma(\tau^\dagger)>\Gamma(\tau^\ddagger)$ for sufficiently large $\psi_2-\psi_1$ that satisfies $b_0-\psi_2<\Gamma(\infty)$. □

Lemma A7 *For given* τ*, the performance measures defined in (1) change in the failure prevention effort* x *and the repair lead time reduction effort* y *defined in §5.3 as follows.*

(a) $\frac{\partial}{\partial x}A(\tau)>0$; $\frac{\partial}{\partial x}C(\tau)<0$; $\frac{\partial}{\partial x}R(\tau)<0$ *if and only if* $L+l\frac{1-\phi}{\phi}<\frac{\int_0^\tau H(t)\overline{F}(t)dt}{H(\tau)\overline{F}(\tau)}$. *Moreover, in the limit* $\phi\to 0$, $\frac{\partial}{\partial x}[C(\tau)+R(\tau)]<0$ *if and only if* $\frac{\tau}{l}>\frac{K}{k}$.

(b) $\frac{\partial}{\partial y}A(\tau)>0$; $\frac{\partial}{\partial y}C(\tau)>0$; $\frac{\partial}{\partial y}R(\tau)>0$; $\frac{\partial}{\partial y}[C(\tau)+R(\tau)]>0$.

Proof The results in part (b) are straightforward, obtained by inspecting the expressions in (1) with the substitution $l=\frac{l_0}{1+y}$. Consider part (a). Let $\ell\equiv L+l\frac{1-\phi}{\phi}>0$ and $m(\tau)\equiv\frac{M(\tau)}{F(\tau)}$. With $\overline{F}(t)=\exp(-pH(t))$ and $M(\tau)=\int_0^\tau\overline{F}(t)\,dt$, we have $\frac{\partial F(\tau)}{\partial p}=H(\tau)\overline{F}(\tau)$ and $\frac{\partial M(\tau)}{\partial p}=-\int_0^\tau H(t)\overline{F}(t)\,dt$. Then $\frac{\partial m(\tau)}{\partial p}=\frac{\partial}{\partial p}\frac{M(\tau)}{F(\tau)}=-\frac{F(\tau)\int_0^\tau H(t)\overline{F}(t)dt+H(\tau)\overline{F}(\tau)M(\tau)}{F(\tau)^2}<0$. Let $\lambda(x)\equiv\frac{\lambda_0}{1+x}$, which satisfies $\lambda'(x)<0$. Substituting $p=\phi\lambda(x)$ and $q=(1-\phi)\lambda(x)$ from (7) into (1) and (2) yields $T(\tau)=M(\tau)+\ell F(\tau)$, $A(\tau)=\left(1+\frac{\ell}{m(\tau)}\right)^{-1}$, and $C(\tau)=\frac{1-\phi}{\phi}\frac{k}{\ell+m(\tau)}$. From these expressions we find $\frac{\partial A(\tau)}{\partial p}=\frac{\partial A(\tau)}{\partial m(\tau)}\frac{\partial m(\tau)}{\partial p}<0$ and $\frac{\partial C(\tau)}{\partial p}=\frac{\partial C(\tau)}{\partial m(\tau)}\frac{\partial m(\tau)}{\partial p}>0$. Since a one-to-one correspondence is made between p and x via $p=\phi\lambda(x)$ with $\lambda'(x)<0$, we then have $\frac{\partial A(\tau)}{\partial x}=\phi\lambda'(x)\frac{\partial A(\tau)}{\partial p}>0$ and $\frac{\partial C(\tau)}{\partial x}=\phi\lambda'(x)\frac{\partial C(\tau)}{\partial p}<0$. Finally, $\frac{\partial T(\tau)}{\partial p}=-\int_0^\tau H(t)\overline{F}(t)\,dt+\ell H(\tau)\overline{F}(\tau)$ together with $R(\tau)=\frac{K}{T(\tau)}$ imply $\frac{\partial R(\tau)}{\partial x}=\phi\lambda'(x)\frac{\partial R(\tau)}{\partial p}<0$ iff $\ell<\frac{\int_0^\tau H(t)\overline{F}(t)dt}{H(\tau)\overline{F}(\tau)}$. To prove the last result in part (a), note $C(\tau)+R(\tau)\to\frac{kqH(\tau)+K}{\tau+lqH(\tau)}$ in the limit $\phi\to 0$ which leads to $p=0$ and $q=\lambda(x)$ with $\lambda'(x)<0$ (see (7) and Lemma A4). Since $\frac{\partial}{\partial q}[C(\tau)+R(\tau)]=\frac{(k\tau-Kl)H(\tau)}{[\tau+lqH(\tau)]^2}$, we have $\frac{\partial}{\partial x}[C(\tau)+R(\tau)]=\lambda'(x)\frac{\partial}{\partial q}[C(\tau)+R(\tau)]<0$ iff $\frac{\tau}{l}>\frac{K}{k}$. □