Impact of Performance-Based Contracting on Product Reliability: An Empirical Analysis

Jose A. Guajardo ● Morris A. Cohen ● Serguei Netessine

The Wharton School, University of Pennsylvania, Philadelphia, PA 19104
josegu@wharton.upenn.edu ● cohen@wharton.upenn.edu ● netessine@wharton.upenn.edu

Sang-Hyun Kim

Yale School of Management, Yale University, New Haven, CT 06511
sang.kim@yale.edu

July 2009, revised February 2010

Abstract

Using a proprietary dataset provided by a major manufacturer of aircraft engines, we empirically investigate the impact of incentives present in after-sales repair and maintenance support contracts on product reliability. In particular, we compare the reliability of products under time and material contracts (T&MC), which have been used traditionally in the aerospace industry, with the reliability under performance-based contracts (PBC), which are gaining wide acceptance. Adverse selection of customers into contract types, and consequently the endogeneity of contract choice, is explicitly modeled by means of a two-stage econometric model. After controlling for this selection process, we find evidence for the positive and significant effect of performance incentives created by PBC on product reliability. We test this hypothesis using a number of reliability proxies, alternative model specifications, and we perform a number of robustness checks. Our estimates indicate an improvement of product reliability in the 20-40% range under PBC, compared to the reliability observed under T&MCs. Our research provides a valuable input into the ongoing policy debate about the effectiveness of performance-based maintenance contracts which are currently being introduced extensively in both the government and the private sectors.

1. Introduction

The ongoing move to performance-based contracting (PBC) in the aerospace and defense industry represents a fundamental shift in customer-supplier relationships for after-sales product support. A PBC contract will pay the service provider on the basis of product use or "up-time"[1], whereas under traditional Time and Material (T&MC) contracts the customer pays the supplier for resources consumed due to the occurrence of product failures and maintenance events. The former approach, which in some industries is referred to as "servicization" (see, for example, Toffel 2002), essentially converts the sale of service products (such as spare parts and repair labor) to the customer into the sale of a service that enables the customer to generate value through the use of the product.

While the provision of after-sales support is a major driver of revenue and profit in many industries, the movement to PBC is especially relevant for after-sales product support in the aerospace and defense industry where products are "mission critical". The movement toward performance-based servicization is motivated by the premise that PBC aligns incentives between the customer and the supplier, so that both benefit when the product's use generates value to the customer. As a consequence, most observers expect that the adoption of PBC will lead to more effective value creation, i.e., with a higher level of performance and a lower cost to the customer, as well as a higher level of profit to the supplier of support services. A growing body of literature (e.g., Kim et al. 2007, 2009a, 2009b) analyzes this issue from an economic and operations modeling perspective. The results of that research indicate that it is possible to design coordinating contracts based on performance and that under such contracts, suppliers and customers have a strong incentive to increase product availability through various means that include improving support capabilities, investing in an appropriate level of support resources, and improving the underlying reliability of the products. While the managerial and analytical arguments for PBC are pervasive, empirical research to support these conclusions is currently non-existent.

The focus of this paper is to evaluate whether PBCs have led to the enhancement of product reliability. Reliability is an important driver of overall system performance (i.e., product availability) which in turn drives customer value creation and the cost of ownership for complex, mission critical products. Further, reliability of complex systems in the aerospace and defense industry is especially critical since product failures lead to disastrous consequences. The analysis we propose is based on observations drawn from an aircraft engine maintenance process and makes a distinction between planned (or scheduled) maintenance and unplanned (or random) maintenance events. In this paper we argue that the concept of reliability is especially relevant for unplanned maintenance events since they lead to costly disruptions in value creation due to unavailability of the product. Planned events, on the other hand, are

[1] A typical performance-based contract will include terms that lead to payment to suppliers based on the number of hours the product has been used and/or the level of availability of key resources.

typically associated with the scheduled use of repair capacity and backup products, and hence their impact is less costly.

The expectation that PBC for aftermarket support and system sustainment has a positive impact on performance, especially through the mechanism of enhanced product reliability, primarily originates from industry reports that are not based on rigorous analysis. These reports and conventional wisdom suggest that PBC improves service outcomes, i.e., it can lead to higher product availability, lower cost of ownership, and reduced customer wait times. Kirk and DePalma (2006), for example, analyze a PBL[2] program in the Navy and, based on a review of historical repair frequency data for several programs, conclude that "there is some evidence that the PBL contract may have helped to improve availability and reliability". Not surprisingly, questions have been raised about the quality of the data and the associated analysis presented in such reports. For example, in a recent report the Government Accountability Office (GAO) states that, "Many DOD program offices that implemented PBL arrangements have limited cost data, and various other factors – such as the lack of business case analyses – further limit an evaluation of the costs of this support strategy. Available data from the programs GAO reviewed indicated mixed results" (GAO 2008).

The findings of this paper are especially timely as PBCs for after-sales support have become increasingly popular in industries such as aerospace and defense, automobile, semiconductors, information technology and software development (e.g., Software as a Service). The adoption of contractual relationships based on performance for after-sales support and other services also spans the public sector. In the U.S., the federal government, and the Defense Department in particular, has mandated this form of contracting for services on a wide-spread basis. Nonetheless, there is an ongoing debate between suppliers of support services and the various federal agencies which are engaged in implementation of performance based programs, regarding the value of contractual relationships based on performance. In addition, the GAO, as noted above, has questioned the accuracy of predictions of the positive impact of PBC and the House of Representatives recently held a hearing on the benefits and costs of PBC.[3]

In this paper, we focus on repair and maintenance services for commercial aircraft engines in the aerospace and defense industry where PBCs for after-sales support has been in place for many years. Our proprietary dataset comes from Rolls-Royce, which, as a major supplier of aircraft engines, provides repair and maintenance services to its customers under two different types of contracts: T&MC and PBC. The main question we analyze is the following: Does the use of PBC have a positive effect on product

[2] Performance-based contracting is also referred to as Performance-based Logistics (PBL) and Power by the Hour® in the defense and commercial aerospace industries, respectively. The latter is a registered Rolls-Royce trademark.

[3] "Performance-based Acquisitions: Creating Solutions or Causing Problems?" Full Committee Thursday, May 08, 2008. http://homeland.house.gov/hearings/index.asp?ID=136.

reliability over the use of T&MC? Reliability, driven by unplanned removals, is a key performance metric in the aerospace industry in which failures of equipment have a significant and direct effect on financial performance.[4]

Measuring the effect of PBC on reliability is a challenging task since adverse selection of customers into contract types generates endogeneity, which leads to complexities from an econometric point of view. In short, and as will be developed in detail throughout the paper, while a positive effect of PBC on reliability can be expected, to capture its true effect we need to consider that certain customers are inherently more likely to choose PBC, depending on their expected usage of the repair facilities. Our analysis, which uses a two-stage framework that explicitly deals with the endogeneity inherent in contract choice by a customer, provides evidence that PBC, indeed, improves product reliability at a rate higher than a naïve analysis (without accounting for endogeneity in contract choice) suggests. This finding is robust to a number of specifications and modeling assumptions. Our results quantify the observed benefits – about a 20-40% reliability improvement – that these contracts have generated in our sample relative to T&MCs. While the focus of this paper is on PBC for commercial aircraft after-sales services, our findings are also relevant to other industries that provide after-sales support for mission critical products.

The paper is organized as follows. In Section 2, we review relevant literature on supply chain contracting. In Sections 3 and 4 we describe the industrial context and the data, and continue with presentation of the econometric analysis and the model specification in Section 5. In Section 6 we discuss estimation, results and robustness checks for our two-stage econometric model. In section 7 we discuss the limitations of the study along with some extensions, and section 8 concludes the paper.

2. Literature Review

There has been substantial interest in the operations management literature concerning the role of contracts in supply chains. Cachon (2003) provides an extensive and relatively recent review of more than 200 papers in this area. As noted there, "the literature contains a considerable amount of theory, but an embarrassingly paltry amount of empiricism." Although papers such as Novak and Eppinger (2001) and Novak and Stern (2008, 2009), for example, empirically examined the impact of product characteristics on vertical integration decisions in the automobile industry, the numerous theoretical predictions in the OM literature related to contractual incentives, for the most part, have not received comparable empirical scrutiny. The findings of this paper contribute to filling this gap in the literature by

[4] Beyond relevance, the focus on reliability is determined by data availability; while we could pursue a more ambitious goal of evaluating the overall impact of PBC on profitability or other financial metrics, those data are not available due to their proprietary nature. Further, such data is typically owned by numerous individual customers and obtaining either the data or the customer's permission to use it for our data set was impractical.

providing an empirical validation of the hypothesis that suppliers under performance contracts will act to improve product reliability, which was generated by an earlier theoretical study (Kim et al. 2009a).

In contrast to the current state of empirical research on supply chain contracting, researchers in other areas have been increasingly interested in studying contracting. In economics, several papers have studied revenue sharing contracts in the video rental industry. Mortimer (2008), for example, analyzes fixed-price vs. revenue sharing contracts between distributors and retailers using a structural econometric approach. Her results indicate that both downstream and upstream firms' profits for popular titles increase by 10% under revenue sharing contracts and consumers are also generally better off. Ho et al. (2008) and Ioannou et al. (2009) investigate other aspects of revenue sharing contracts in this industry. In addition, several empirical papers examine franchising contracts (e.g., Lafontaine and Shaw 1999, Lafontaine and Slade 2001). Finally, there is a stream of literature that studies incentive alignment between firms and their employees by means of different variants of pay-for-performance contracts. Examples include Prendergast (1999, 2002a, 2002b), Lazear (2000) and Banker et al. (2001). Although this research on performance-based contracts is distinct from ours because of its focus on a labor setting, we note two results from this literature that are relevant to our study: (i) employees respond to incentives usually improving firm performance, and (ii) there is self-selection of better employees through pay for performance schemes. The findings of our paper offer related insights in the firm-to-firm aftermarket support setting.

PBC has also received attention in such application areas as health care, public policy, and software development. The empirical results in Lu et al. (2003) show that PBC leads to more referrals and a better match between illness and treatment intensity, suggesting that PBC induces incentive alignment. Shen (2003) finds that PBC provides incentives for nonprofit providers of substance abuse treatment to select less severe patients into treatment. Other examples from the public sector include Heinrich (2002), who analyzes outcome-based performance management using data from federal job-training programs and a case study by Heinrich and Choi (2007), based on social welfare programs in Wisconsin, showing that the service provider responded to the incentives afforded by PBC.

In the context of offshore software development, Gopal et al. (2003) analyze the impact of fixed-price vs. T&MCs on software vendor profits. As we do, they use a two-stage approach that includes modeling of the determinants of contract choice in the vendor-developer relationship. Their results indicate that vendor profits are higher under T&MCs, controlling for variables such as project type and effort. Gopal and Sivaramakrishnan (2008) extend this analysis further by examining the impact that factors such as project size and duration, team size, and risk of employee attrition, have on contract choice.

The papers noted above study the influence of contracts on different performance outcomes. In our research, an equally important issue is determining the factors that influence contract choice. A number of papers examine contract choice in different contexts. Examples include Slade (1996), who studies contracts between oil companies and their service stations and tests various multitask agency hypotheses as drivers of contract choice. Ackerberg and Botticini (2002) examine contract choices using archival data on agricultural contracts between a landlord and tenants, and test hypotheses related to the role of risk sharing and transaction costs on contract selection. Using data from U.S. Air Force engine procurement, Crocker and Reynolds (1993) analyze how different variables affect the degree of contract completeness (i.e., how precisely the contract specifies future duties and contingencies) that parties choose.

Our paper also contributes to the growing stream of empirical research in operations management. Papers that are related to ours include Ramdas and Randall (2008), who find that component sharing in the automotive industry can hurt product reliability. We also study reliability empirically but in a different industry and in the contracting choice context. In an empirical study that focuses on after-sales support for defense systems, Deshpande et al. (2003) analyze how the interaction of attributes such as the criticality and cost of service parts and the nature of inventory policies to manage them affect performance. As in our case, this is an empirical study in an aftermarket context, although it is not focused on reliability or contracting. Terwiesch et al. (2005) empirically analyze demand forecast sharing in the buyer/supplier relationship in an application to the semiconductor industry. Their analysis indicates that non-optimal gaming behavior among all parties occurs as a consequence of conflicting incentive schemes. While this paper is not about contracting, it is related to our research since it empirically studies the role of incentives in an operational context. Finally, in a study related to PBC, Lee and Zenios (2007) develop a performance-based payment system for Medicare by empirically estimating data from patients needing kidney dialysis.

To summarize, we note that while fields such as economics, public policy, information systems, and healthcare have generated several examples of empirical research on the role of performance contracting, this has not been the case in operations management, despite significant attention that has been given to supply chain contracting research. Thus, in addition to providing a scientific input to the ongoing policy debate on performance incentives for government and defense service procurement, our paper contributes to the existing OM literature as it represents one of the few empirical studies of supply chain contracting. As a result, we believe that our paper contributes to closing the gap between theory and practice in this important area of OM research. Finally, to the best of our knowledge, our paper provides the first empirical comparison between performance contracts and non-performance contracts that are used for aftermarket customer support.

3. Industry Background

In this paper we focus on the maintenance, repair and overhaul (MRO) industry for commercial aircraft. According to Standard and Poor's industry report (2009), this sector generated revenues of $117 billion in 2007, of which $60 billion was related to military MRO, $45 billion to air transport (commercial aircraft) MRO, and $12 billion to business and general aviation MRO. More generally, reported statistics (see Cohen et al. 2006 and the references therein) indicate that sales of spare parts and after-sales services in the U.S. represents 8% of annual domestic product, meaning that customers spend approximately $1 trillion every year on assets they already own. The investment in resources required to enable the delivery of services to support products is also huge, e.g. spare parts inventory investment has been observed to be 5% of sales in computer and high technology industries (Cohen et al. 1997).

Aircraft owners (typically airlines) face the problem of properly managing maintenance and repair of aircraft equipment, including managing the risk of infrequent equipment failures and performance of scheduled maintenance checks, so as to preserve aircraft availability, thus avoiding the high opportunity cost of having an aircraft on-the-ground. Customers typically purchase after-sales service support from the OEM and/or other service suppliers on either a transaction basis (i.e., through T&MCs), or purchase a contract for support such that payments are based on the number of flying hours (i.e., PBC).

Aircraft fleet availability therefore is the most important performance metric for airlines and other customers. We note that availability is influenced by several factors that include subsystem and part reliability, spare parts inventory, repair capacity and repair lead time. For example, if a critical part of an aircraft subsystem stops functioning, it must be replaced by a working unit drawn from the spares inventory (if it is available) as soon as possible to minimize disruption to flight schedules. All broken units are ultimately returned to a support depot where they are either repaired or scrapped. A key challenge for both customers and suppliers in this industry is to reduce the cost incurred from failures while maintaining an acceptable level of fleet uptime (availability) through the management of resources (i.e., spare parts inventory and repair capacity) and through interventions that could affect the reliability of the product and/or the performance of the support processes.

Recently, the airlines have been increasingly adopting outsourcing strategies for MRO services in order to focus on their core activities and reduce costs. This trend has led to expansion of the range of MRO services offered by suppliers of various types of aircraft subsystems (e.g., hydraulic power, engine, landing gear, avionic systems). Typically, it is the OEMs themselves that offer support services for their subsystem products because the highly customized and complex nature of their products makes it difficult for a third party to provide similar product care. The provision of such services is also very profitable with margins that often exceed those associated with the sale of the product. Examples of such major (OEM) suppliers in the MRO industry include Pratt & Whitney, General Electric Co., Rolls-Royce, Honeywell Aerospace, Lockheed Martin and Boeing. There are also many smaller MRO companies that provide a wide range of services that include scheduled maintenance checks and parts repair. Such providers include the Triumph Group Inc., AAR Corp and Heico Corp. MRO service providers usually offer different types of contracts under which their customers can receive support services, including time and material (T&MC), fixed-price (FP) and performance-based contracts (PBC). Different versions of PBC have become widely adopted, evolving and transforming the relationship between manufacturers and customers in the industry. For example, recently Boeing aggressively started pursuing contracts for its 787 GoldCare program. As noted in 2006 by a company's vice-president, "before we were just selling parts, now we are selling airlines a power-by-the-hour service and we are guaranteeing availability".[5] Although PBC is an important factor in the industry, there are no published estimates of its actual effect on any of the possible performance metrics.

In this paper, we study the performance implications of the contractual relationship between Rolls-Royce, a major supplier of aircraft engines and services to support them, and its customers. The dataset we analyze was provided by Rolls-Royce. Rolls-Royce delivers after-sales repair and maintenance services for its customers under two different types of contracts: T&MC and PBC. The main difference between the two types of contracts is the mechanism under which the customer pays for the support services. Under a T&MC the customer pays for the materials and resources that are consumed each time a maintenance event occurs. Under PBC, the customer agrees to pay a fee that is proportional to the actual flying hours the customer generates from their fleet of aircraft. (For example, a customer pays $$x$ per flying hour to the supplier with a guaranteed minimum number of hours flown per quarter;[6] note that flying hours can be converted into aircraft fleet availability.) In other words, compensation to Rolls-Royce under PBC is directly tied to the performance outcome that the customer values.

4. Data and main hypothesis

4.1 Data description and preprocessing

The dataset which Rolls-Royce made available to us consists of five years of data (July 2002 - July 2007, hereafter the observation period) of maintenance events (product removals) for different models of aircraft engines produced by Rolls-Royce. A *removal* of the aircraft engine may be necessary due to a part failure or for maintenance purposes, resulting in a *shop visit* to the service provider. Our dataset allow us to distinguish between planned and unplanned removals. Planned removals are associated with

[5] Lou Mancini, in *Airline Business*, "Maintenance special report: service culture" (09/26/2006).
[6] Variations of this payment scheme are observed in other settings, especially in the defense industry. There, many PBC contracts are based on a weighted average of different performance metrics (including "soft" metrics like customer satisfaction). However, flying hours is the only metric used in Rolls-Royce's PBC ("Power by the Hour®").

regular scheduled maintenance tasks; unplanned removals are associated with unexpected maintenance events, and are thus related to the notion of random or unexpected failure of an engine, and consequently are the focus of interest when it comes to estimating engine reliability.

For a better understanding of the data, it is useful to describe more precisely how the data for this research was provided to us (see also Figure 1). We obtained two different data files: a spreadsheet containing all of the removal events between July 2002 and July 20007 for the engines in our sample (hereafter called the 'removals file'), and a spreadsheet containing a list with all engines registered with Rolls-Royce for each customer (hereafter called the 'engines file'). For each removal of an engine unit, a list of the relevant information contained in the removals file is as follows:

- engine unit ID,
- engine model,
- date at which the engine entered the repair shop,
- cumulative flying hours ("time since new", TSN) and cycles ("cycles since new", CSN; a cycle is defined as the interval between a takeoff and a landing) when the engine entered the repair shop,
- removal type: planned/unplanned,
- aircraft tail number in which the engine is installed,
- aircraft model id,
- ID of the customer that owns the product,
- the contract type (T&MC or PBC) under which the product receives service.

The engines file, on the other hand, contains a list with all engines registered with Rolls-Royce at the end of the observation period, i.e., July 2007. For each engine in this file we know:

- engine unit ID,
- engine model,
- cumulative aircraft flying hours and cycles at the end of the observation period,
- ID of the customer that owns the product.

After cleaning data by removing inconsistent observations (e.g., for a given unit, reported flying hours at a shop visit in 2006 are less than the flying hours reported in a shop visit in 2005), our sample consisted of 763 engine units for which at least one engine removal is observed in the 5 years observation period. There are essentially two engine models in our sample: for one of these product models there are 3 different versions, and for the other there are 2 different versions. The engine models are installed in three different types of aircraft. For all types of product, we observe engine units covered by either PBC or T&MC. In the sample of 763 engine units with removals, 21.4% are covered by T&MC, 78.6% are covered by PBC. Among the pool of 763 engines, 305 of them (40%) had at least one unplanned removal during the observation period. These 305 units are associated with 48 different customers.

Our approach for capturing the impact of PBC on product reliability focuses on unplanned removal events. Unplanned removals are highly undesirable events for aircraft owners since an aircraft on-the-ground generally results in high opportunity costs, with estimates being as high as hundreds of thousands of dollars per day for an unplanned removal for a fully loaded wide body commercial aircraft. In fact, serious unplanned removals for the types of engines in our sample usually require several weeks at the repair shop. Furthermore, there is always a possibility that unplanned failure may lead to a catastrophic event. In contrast, in the case of planned removals the shop visit is programmed in advance, and the appropriate replacements can be scheduled to be available to avoid having an aircraft on-the-ground. Therefore, the fact that a removal is planned/unplanned largely determines the unavailability of an aircraft. Indeed, we have observed in practice that release of a product on the scheduled completion date for a planned maintenance event has a high priority. In order to achieve on-time job completion, service providers will employ tactics like cannibalization of the incoming product and use expediting to source any missing parts.

Hence, for the purpose of studying product reliability, our analysis focuses on the sample of 305 engines with at least one unplanned removal. We note that, in the sample of 305 engines with unplanned removals, 21.3% of the engines are covered by T&MC, and 78.7% are covered by PBC, i.e., exactly the same proportions as observed in the full sample: 39.9% of the T&MC engines had unplanned removals, and 40% of the PBC engines had unplanned removals. In what follows, we focus on the sample of 305 engine units with at least one unplanned removal, and the observed proportions suggest that focusing the analysis on unplanned removals does not generate, a priori, a sample bias. We note also that reliability is, by definition, a product unit variable, i.e., what fails is a product. Consequently, the unit of analysis of interest in this research is an engine unit. As will be noted throughout, focusing on an engine unit allows us to capture several details associated with the problem (e.g., initial condition of the product, product type, etc.) and make full use of the dataset we have available for this research, together with allowing for appropriate sample size. From the two data files described earlier, we are able to obtain and calculate several variables of interest which describe the product and the customer. Table 1 provides definitions and descriptive statistics for the variables used in our analysis.

Note that, from the description of our dataset, we do not observe the initial age of the engine (ini_age). As described in the next section, we estimated this variable from the observed data. Note also the sparseness of unplanned removals (nremovals_up), with an average of 1.2 in our sample. Finally, it is important to emphasize that Rolls-Royce did not provide any financial data (prices charged for each

Unit	No. obs.	Variable Name	Variable definition	Mean	Median	Std. Dev.	Min	Max
Engine	305	ini_age	Time since new in July 2002 ($TSN(T_B)$)	3,230.8	2,961.6	2,760.4	0	12,915.5
Engine	305	final_age	Time since new in July 2007 ($TSN(T_E)$)	10,423.8	11,094.3	4,819.2	363.0	21,599.6
Engine	305	eng_avgflighttime	Average flight time $TSN(T_E) / CSN(T_E)$	1.216	1.103	0.265	0.821	2.503
Engine	305	nremovals_up	Number of observed unplanned removals	1.207	1	0.459	1	3
Customer	48	fleetsize	Number of eng. units registered in Rolls-Royce	40.0	8	106.2	2	593
Customer	48	fleetmix	Number of dif. eng. models registered in Rolls-Royce	1.646	1	1.000	1	5
Customer	48	ow_avgflighttime	Average flight time (average across engine units)	1.409	1.388	0.375	0.826	2.497

Table 1: Definition of variables and descriptive statistics

contract, repair costs, etc.), nor do we know the contract terms (e.g. duration of the contract, price charged per hour of utilization of the product) other than whether a product is covered by PBC or T&MC. Also, all our data is for the period 2002-2007 and we do not have access to the prior history of each engine or any information about previous relationships between a customers and Rolls-Royce. We discuss potential limitations due to data availability in the concluding section.

4.2 The dependent variable: product reliability

While there are a number of performance metrics that are relevant to value creation derived from the after-sales use of products (e.g. availability, reliability, cost of ownership), our analysis focuses on product reliability. We do so because reliability is clearly a key metric that is of interest in practice and its evaluation was supported by the dataset that was provided to us. Data on other metrics were either unavailable or deemed too sensitive since our dataset is based on the maintenance experience of Rolls-Royce's commercial customers. The focus on reliability none the less provides valuable insights since, as mentioned in the Introduction, theory guides us to believe that different incentive structures under either T&MC or PBC contracts should be reflected in variations of product reliability. While there are several possible ways to approach the measurement of product reliability, we have chosen the mean time between unplanned removals (MTBUR) to be the measure employed in our main analysis since (1) it can be computed from available data and (2) it is in fact a key reliability metric that practitioners in the aerospace, defense and other industries constantly monitor. MTBUR represents the average time (flying hours) that a product is used without the need for an unplanned removal for repair and maintenance purposes. Unlike some other metrics (such as mean time between removals, which includes both planned and unplanned removals), MTBUR is a good representation of the physical reliability *inherent* in the product since an unplanned removal event occurs only when an engine fails randomly, free of managerial intervention. Planned removals are typically managed through specific scheduling of resources to carry out the maintenance and to provide backup capacity while the product is under repair. In the rest of our discussion we use the terms *product reliability* and *mean time between unplanned removals* interchangeably.

Although MTBUR is an appropriate metric of product reliability, it still poses nontrivial issues for our analysis because unplanned removals – and, in fact, removals of any kind – are quite rare events. In our observation period of 5 years, the majority of the products in the dataset (81.6%) exhibit only one unplanned removal; the remaining units had either two or three unplanned removals (16.1% and 2.3% of the sample, respectively)[7] which somewhat limits our ability to compute the true MTBUR of a product. Additionally, the data suffers from censoring since information on any unplanned removals that occurred before July 2002 or after July 2007 is excluded. Defining a rule to calculate MTBUR, therefore, is a challenging task. Our approach is to define a few alternative proxies that deal with the censoring issue in different ways, and evaluate our models with each of them. In total, we consider four proxies. As we will show subsequently, the main results of our paper are fairly robust under alternative proxy choices.

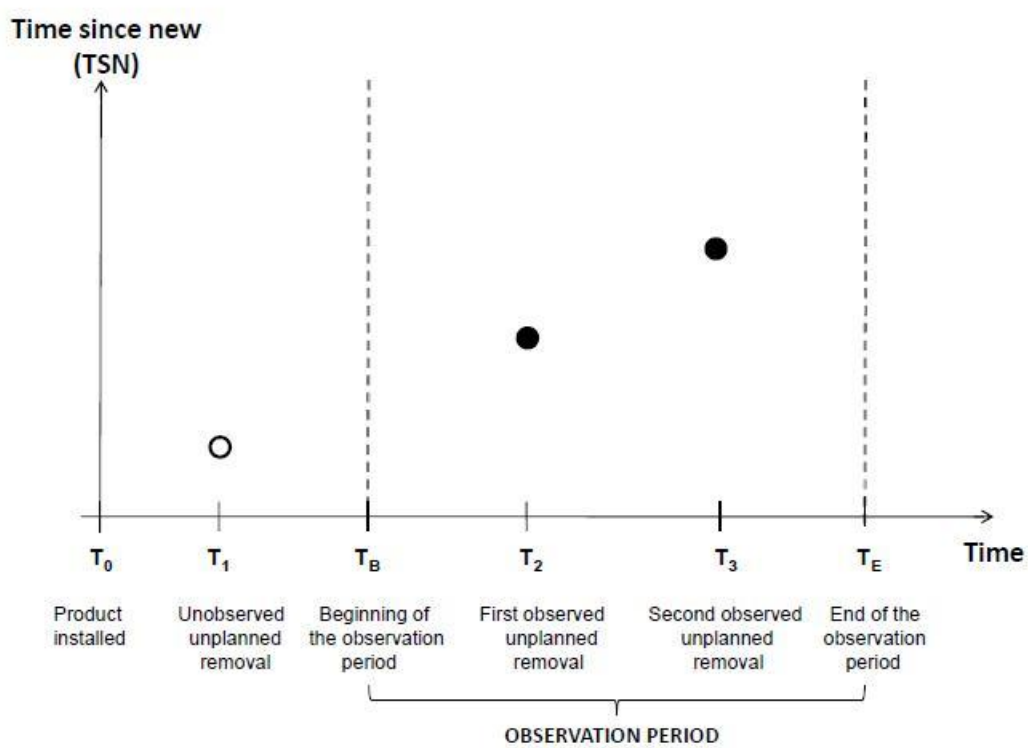

Figure 1: Example of an unplanned removal sequence for a product.

[7] It should be noted that the products that did not have unplanned removals during the observation period are not included in the sample. Therefore, we are in fact considering a *conditional* MTBUR in our analysis: i.e., MTBUR of a product given that an unplanned removal has occurred at least once in the sample period.

We illustrate the procedure used to calculate the reliability proxies with an example (see Figure 1). Consider a product that was installed in an aircraft at time T_0, before the beginning of the observation period T_B (July 2002 in our case). Assume that a first unplanned removal occurred at time $T_1 < T_B$, i.e., this event was unobservable to us. Suppose we observe the 2 unplanned removals at times T_2 and T_3, which occurred before the end of the observation period T_E. Let TSN(T) denote the time since new (flying hours) of a product at time T.[8] In the example, the (true) MTBUR is given by TSN(T_3)/3, but we do not observe the first unplanned removal and we do not even know if it took place or not. In other words, we only know the values of $T_{B,}$ T_2, T_3, T_E, and the respective measures TSN(T_2), TSN(T_3), and TSN(T_E), but not the values of T_0 (the time at which the product was installed), T_1 (the time the first unplanned removal occurred), the corresponding flying hours TSN(T_1), and the *initial age* of the product at the beginning of the observation period TSN(T_B). We build our first proxy for MTBUR as defined by the following equation:

$$\text{MTBUR} = \frac{\text{TSN(latest observed unplanned removal)} - \text{TSN}(T_B)}{\text{\# observed unplanned removals}}$$

In the example, our proxy for the MTBUR is given by [TSN(T_3)- TSN(T_B)]/2. However, as we pointed out, the data do not include the value TSN(T_B). We compute an estimate for TSN(T_B), say $\text{TSN}^*(T_B)$, by assuming that there was a constant rate of usage for the product throughout the observation period. Specifically, we estimate this value as a linear projection of the line defined by the first observed removal and the age of the product measured at the end of the observation period, i.e. we estimate the slope of the line using the first removal and the end of the observation period as the two data points. We then project the line back to T_B in order to obtain an estimate of the initial age of the product defined as $\max\{0, \text{TSN}^*(T_B)\}$. We believe that this approximation provides a reasonable estimate for MTBUR. Note that our adjusted measure gives the correct value for MTBUR if the product was not installed before July 2002. The bias introduced by this metric will vary since it depends on the number of unobserved removals. We attempt to reduce the potential bias by subtracting the initial product age, as indicated in the MTBUR formula. Note that if we do not apply such correction and omit subtracting the initial age of the engine, the proxy would overestimate the true MTBUR for all engines that had unplanned removals before the beginning of the observation period.

While MTBUR is a reasonable and perhaps the most straightforward proxy of reliability, it is not complete as its definition above does not take into account the fact that there were no unplanned removals between T_3 and T_E. We therefore explore three alternative proxies that involve a different treatment of the *right tail* of the data, which we define using the example in Figure 1 (we use different names for clarity):

[8] Note that TSN is measured in flying hours, i.e., hours of actual activity in the air, which is different from calendar time. In Figure 1, the former is shown on the y-axis and the latter is on the x-axis.

- MTBUE = average{TSN(T_2) – TSN(T_B), TSN(T_3) – TSN(T_2), TSN(T_E) – TSN(T_3)},
- MAXMTBURMTBUE = max{MTBUR, MTBUE},
- $\text{InvFRate} = \dfrac{\text{TSN}(T_E) - \text{TSN}(T_B)}{\text{\# observed unplanned removals}}$.

Thus, the MTBUE (mean time between unplanned events) includes the most recent portion of the time in which an engine was not subject to an unplanned removal, i.e., it considers that as an "event". The MAXMTBURMTBUE also includes that information by contrasting the MTBUE with the MTBUR. Finally, the InvFRate represents the inverse of the failure rate, the latter being calculated as the number of observed unplanned removals over the flying hours of an engine during the observation period. Note that, although we repeat our analysis using all these proxies, for ease of exposition we focus mainly on MTBUR when we discuss our econometric model below.

The descriptive statistics for each reliability proxy are displayed in Table 2; all four proxies are measured in flying hours. Note that for all four alternative dependent variables, the mean is consistently higher for PBC than for T&MC engines (ranging from a 10% to 20% increase under PBC). Instead of computing the mean time between unplanned removals for each individual product and using it as a dependent variable in a regression model, an alternative approach would be to infer the mean time between unplanned removals (as an output of the analysis) by estimating the underlying distribution of the time between removals using techniques drawn from duration modeling (see, for example, Cameron and Trivedi 2005 Ch. 17). This alternative approach, for which we provide results later in Section 6.2, has some econometric challenges on its own. In what follows, we focus our discussion on the results which are based on the well-established two-stage econometric framework that explicitly deals with endogeneity of contract choice.

4.3 Main hypothesis: the effect of PBC on engine reliability

Our goal in this paper is to establish the effect of contract type on engine reliability, as defined by the mean time between unplanned removals. We postulate that contract type can affect a product's MTBUR.

	Overall sample		**T&MC only**		**PBC only**	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
MTBUR	3,328	2,458	2,872	2,632	3,451	2,400
MTBUE	3,507	1,837	3,147	2,029	3,605	1,773
MAXMTBURMTBUE	4,079	2,227	3,609	2,536	4,206	2,123
InvFRate	6,456	3,278	6,016	3,844	6,575	3,106

Table 2: Mean and standard deviation for alternative reliability proxies

As we have discussed above, under PBC the supplier of the service is paid in proportion to the flying hours generated by the customer, creating an incentive to maintain or improve product reliability. In this section we formulate the main hypothesis predicting this relationship. Much of this development relies on the insights derived from the contract theory (Bolton and Dewatripont 2005; hereafter, B&D).

Our first consideration comes from the incentive effects of the two contracts due to hidden actions and the resulting moral hazard (B&D, pp. 129-136). Namely, how much care of an engine is taken by the supplier is not directly observable to the customer under either contract, and therefore, we need to carefully consider what kind of incentives each contract creates for the supplier with regard to engine care. The basic intuition is quite simple. Under the T&MC, the supplier is paid each time the engine fails and the customer requests repairs. Thus, the supplier has less motivation to increase reliability: lower reliability means more repairs and hence more business. On the other hand, under the PBC, the supplier is paid for the uptime of the engine. While there are multiple ways to increase the uptime (e.g., through higher inventory of spares, faster repair turnaround times etc.), one important way is reliability improvement, which can be achieved by using better (higher performance) parts, preemptively replacing parts during scheduled/unscheduled maintenance visits, investing in parts redesign and supersession of old parts with improved versions, and inspecting parts more carefully during maintenance visits. Therefore, PBC creates a greater incentive to increase reliability, than that which occurs under T&MC. This reasoning is formalized in Kim et al. (2009a), who employ the multi-task principal-agent model (B&D, pp. 216-223) to demonstrate that the PBC contract provides the supplier with better incentive to invest into reliability improvement than T&MC does. By contrast, a T&MC leads to distorted incentives and consequently to lower reliability.

Our second consideration comes from an analysis of decision rights under the two contracts. Under a T&MC the customer has the decision rights (or approval rights) with respect to what exact procedures are to be performed during a maintenance event, because he is billed for every part and labor expense. Naturally, this arrangement leads the customer to be hesitant to approve those repairs that are deemed "unnecessary" from his point of view or even to opt for a bare minimum of service. Further, T&MC customers might not even bring in engines regularly for scheduled maintenance because of the cost associated with doing so. These decisions by the customers are based on their assessment of the tradeoff between the immediate cost of a repair and the long term revenue/cost consequences of delaying or minimizing maintenance effort. PBC customers, on the other hand, do not pay for scheduled maintenance and hence do not have any reason to not closely follow all maintenance procedures by the supplier and the product OEM. Further, PBC customers who do bring the engine in for scheduled maintenance may not have full decision rights regarding what is replaced/repaired, and hence such customers may not be able to interfere with suggestions of the supplier who has an incentive to improve reliability (as described above). At the same time, the supplier is much more likely to have better information regarding necessary repairs based on experience with the entire installed engine base, and hence he has a better chance to utilize this knowledge under a PBC. Thus, the PBC structure is such that information asymmetry is resolved by allocating decision rights to the most informed party, i.e., the supplier. Summarizing, under a PBC the supplier has not only a greater incentive to provide the best quality of service and perform the necessary actions to improve reliability, but also has the decision rights to do so. This is less likely to occur under a T&MC, where the decision rights ultimately reside with the customer.

Finally, consistent with our reasoning above, there is anecdotal evidence from practice that PBC contracts, and in particular, those used for Rolls-Royce engine maintenances, lead to higher reliability. For example, Thomas (2005) reports that: "… Rolls-Royce had officially won praise from the US Navy for its innovative 'PBtH' support for the F405 engine". According to Captain Win Everett, Program Manager for the US Navy's Undergraduate Flight Training Systems at NATC 'Patuxent River' (Maryland), 'under Rolls-Royce, engine availability has exceeded the current target of 85%, the average time between engine removals has increased from 700 hours to over 900 hours, and expected engine removals have fallen by 15 per cent.'" Likewise, Business Wire (2008) cites reliability improvements after introduction of PBC contracts by Rolls-Royce. In PBC environments not directly related to engines or Rolls-Royce, Geary and Vitasek (2008) observe that, "... there were also 90 improvements made to the APU (auxiliary power unit), with 20 of those being reliability improvements", and "the contract with Raytheon... the system design and support concept used in this program have resulted in a 200% improvement in MTBOF (mean time between operations failures) and a 400% improvement in mean time to repair." Hence, based on the above arguments and evidence, we formulate the main hypothesis of interest for this research: *PBC increases engine reliability relative to T&MC.*

While the arguments presented above describe mechanisms under which PBC could improve reliability to greater extent than T&MC, arguments counter to this hypothesis could also be made. We have discussed moral hazard on the supplier side, and have argued that PBC resolves that problem by providing greater incentives for the supplier to improve reliability under PBC. We have also argued that under PBC the customer has the incentive to closely follow appropriate scheduled maintenance protocols, as the costs of doing so are incurred by the supplier. Beyond that, it could be argued that PBC creates a moral hazard problem on the customer side: the actions of the customer beyond following regular maintenance at the supplier – i.e., the way customers regularly operate their equipment – are unobserved by the supplier. Noting that PBC provides full protection to the customer, it can be hypothesized that some PBC customers may operate their equipment less carefully, as the costs of potential failures will be covered by the supplier and as a result reliability under PBC would be reduced. Moreover, as discussed in the agency literature, e.g., Holstrom (1979), monitoring is of value in such cases. In fact, under PBC, in our application, the supplier typically keeps track of relevant engine metrics such as actual flying hours, which qualifies the relevance of the potential moral hazard problem described above.

Finally, we note that suppliers under T&MC also have an incentive to improve reliability due to the cost of not having engines available and their concern about the long term potential impact on their brand reputation regardless of the contract in place. The empirical analysis of this paper explores the relative impact of PBC on reliability. While we believe that the arguments in favor of our hypothesis are more solid and should prevail, determining the relative effect of PBC on engine reliability is ultimately an empirical question.

5. Econometric Model

5.1 Econometric Framework

Our goal is to build a model that properly captures the effect of contract type on product reliability, measured by our alternative proxies MTBUR, MTBUE, MAXMTBURMTBUE, InvFRate. A major challenge associated with isolating the true marginal effect of different contract types on product reliability is the inherent endogeneity associated with contract type choice by customers. Indeed, customers do not sign on for either a T&MC or a PBC randomly, but rather respond to several factors that influence this decision, i.e., self-selection is expected in our setting. A number of empirical studies have considered the issue of contract choice decisions by firms in different contexts (see, for example, Slade (1996) for an application in the oil industry and Crocker and Reynolds (1993) for an application to the Air Force engine procurement process). In these and other studies, endogeneity of contract choice has been regarded as a key econometric issue in testing contract design hypotheses (Masten and Saussier 2002). Endogeneity of contract choice by a customer is an important issue in our problem as well. As a consequence, usual ordinary least squares (OLS) estimation would lead to biased estimates of the marginal effect of contract choice on product reliability. General econometric discussions on the importance of accounting for self-selection and related methods can be found elsewhere (e.g., Heckman 1979, Maddala 1983, Greene 2008).

To account for endogeneity in contract selection, we utilize a two-stage treatment effects model (see Maddala 1983, p. 120). This approach allows us to estimate the effect of a binary treatment (PBC) on a numeric outcome (product reliability), given that the treatment assignment is not random but rather is determined by an endogenous decision process, which the customer carries out. The approach utilizes a two-stage structure that involves a first stage to explain contract choice (Equation 2) and a second stage to explain product reliability (Equation 1).

$$y_i = x_i\beta + \delta z_i + \varepsilon_i \qquad \text{(Eq. 1)}$$

$$z_i = \mathbf{1}(w_i\gamma + v_i > 0) \qquad \text{(Eq. 2)}$$

The observed reliability of product i – denoted by y_i – is explained by the exogenous covariates x_i and the binary endogenous variable z_i (that in our case is equal to 1 for products covered by PBC contracts and 0 otherwise). As in standard discrete choice models with latent variable representation, e.g., probit, the binary variable for contract choice (z_i) is modeled as an indicator function, dependent on a set of exogenous covariates w_i, which drive the choice process. The error terms (ε_i, v_i) of the outcome and choice equations, respectively, account for unobservable characteristics which are allowed to be correlated, and are modeled as a bivariate normal random variable with distribution $N_2(0, 0, \sigma^2, 1, \rho)$; where the variance of v_i is normalized to one for identification purposes. If the correlation between both error terms is equal to zero then the outcome and choice equations can be estimated independently (Equation 1 could be estimated by OLS), i.e., endogeneity is not relevant for the problem. For additional information on this model, including a discussion of identification, the reader is referred to Maddala (1983, p.117-125). A good illustration of the biases that can be generated by not accounting for self-selection in an application to firm entry and performance can be found in Shaver (1998).

5.2 Model Specification

In order to properly capture the main effect of our interest, i.e., significance and the magnitude of the coefficient δ in Equation 1, we need to include x_i, the factors influencing engine reliability other than the contract type, as well as the variables that may influence the contract choice that are captured by the covariates w_i. There are characteristics of both the product and the product owner that can play a role. Of those, we need to identify the most salient ones in order to avoid collinearity issues.

An obvious factor that can influence the observed MTBUR is the initial condition of the product. On one hand, reliability theory (see Rausand and Høyland 2004) often argues that very young and very old engines are more likely to have low reliability: younger engine units typically require adjustments at the beginning of their life, while older engines may fail more often due to part wear and tear caused by usage over time. To check these conjectures, we plot the distribution of MTBUR for different ranges of initial product age in Figure 2. The graph suggests that MTBUR is lower for both *new* and *old* products, and is higher for *medium age* products, which is in line with the reasoning proposed above. In fact, there appears to be a concave relationship between MTBUR and initial age. In order to account for such effects, we include both linear and quadratic terms for the initial age of the product in our model specification. The linear term should take care of the initial increasing trend in product reliability, while the quadratic term should reflect the decreased MTBUR for *old* products; polynomial function has been used by Hubbard (1988), among others, to capture the effect of initial age.

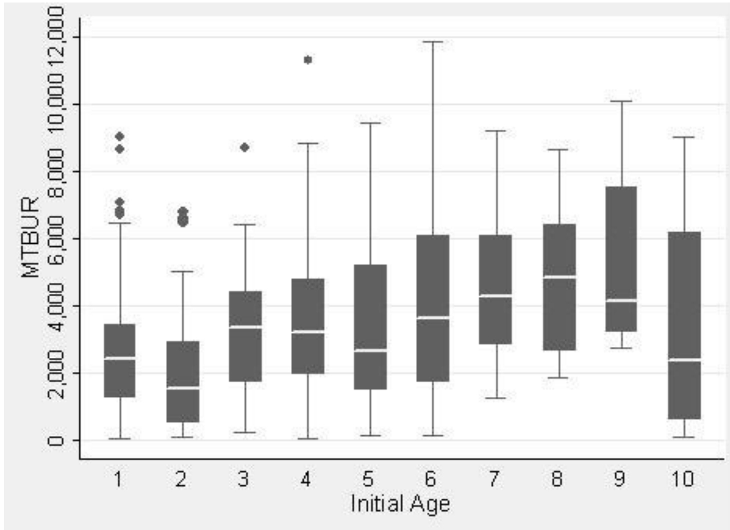

Figure 2: MTBUR vs. Initial age (flying hours). Variable initial age is categorized as follows: '1'→ 0<initial age<=1000, '2' → 1000<initial age<=2000, ..., '10'→ 9000<initial age.

A second product characteristic that can be related to reliability is the average flight time of an engine. This is a relevant variable since it reflects how the engine is used by the customer, directly impacting reliability. For example, more take-off/landing cycles per flight hour may decrease the reliability of the engine. Hence, it is important that we control for this variable. In the dataset there are records of the time since new (TSN) and the cycles since new (CSN) for each engine, measured at the end of the observation period. The average flight time for each engine is then the ratio TSN/CSN. We include this control variable in the outcome equation.

While the initial age of the engine and the average flight time capture a relevant part of both the initial conditions and usage patterns of the product, there are product characteristics related to, for example, the capacity of the aircraft in which the engine is installed, the stage of the engine model in its product cycle and other product characteristics. As mentioned in the data description, there are 2 engine models in our sample: one of them comes in 3 different versions (1a, 1b, 1c), and the other in 2 different versions (2a, 2b). These engines are installed in 3 different aircraft models. Not every engine model can be installed in

Aircraft	Engine Model				
Model	1a	1b	1c	2a	2b
1	13	51	11	0	0
2	0	156	5	0	0
3	0	0	0	49	20

Table 3. Engine models and aircraft models.

every aircraft model. Table 3 shows the number of units found in each combination in our sample.[9] Figure 3 shows the distribution of the MTBUR in each case. The aircraft model captures the essence of the differences across products more parsimoniously. We include aircraft model dummies as control variables in the outcome equation of our model specification.[10]

Finally, the characteristics of the aircraft owner may influence reliability. For example, geographic location of the owner can be a proxy for proximity and availability of repair shops and for other local market conditions such as price. We include dummy variables for the geographical region of the owner: these variables are grouped into three categories, "usa", "europe", and "other", where the latter includes the countries that are neither in Europe nor in the U.S.[11] Naturally, we would like greater granularity to capture local effects at the level of countries or even regions within a country. Recall, however, that we only have 48 customers in our sample, and so defining the geographical variable too narrowly may not capture relevant differences. At the same time if we define geography too broadly, it will not be meaningful from a statistical estimation point of view.

This completes the specification of the outcome equation that we henceforth refer to as the basic specification. Below, we also report results for more complete model specifications that include additional customer characteristics. More precisely, these models include additional variables *fleetsize* and *fleetmix*. The size of a customer's fleet may be seen as an indicator of his/her capability to develop maintenance in house, or alternatively, as a sign of the quality of the service provided by Rolls-Royce (for example, Rolls-Royce may give priority/better service to larger customers). On the other hand, whether a

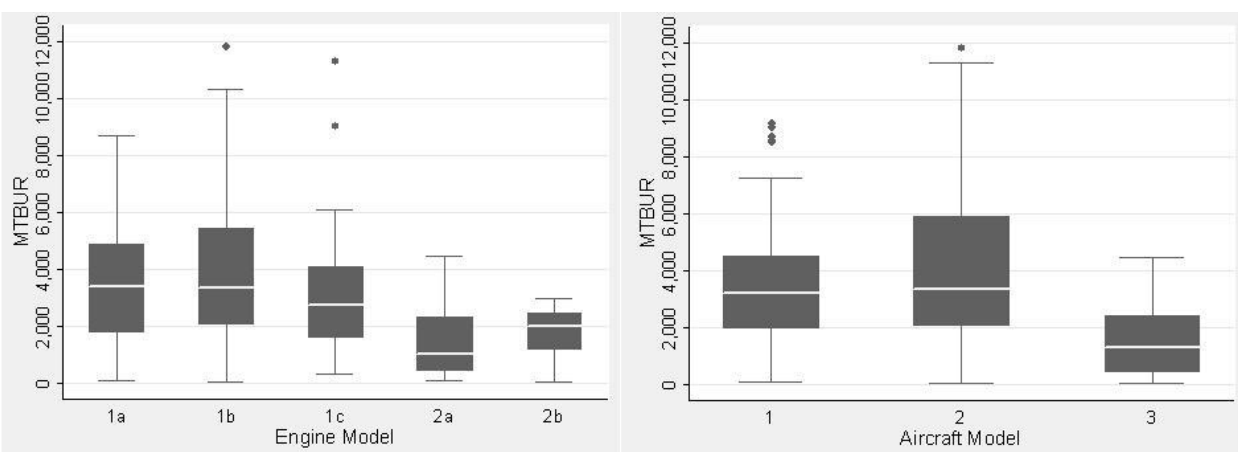

Figure 3: MTBUR vs. Engine Model (left). MTBUR vs. Aircraft Model (right).

[9] Different versions of a given engine model have essentially the same technical specifications. The two engine models in our sample are designated as small aircraft engines for civil aerospace.
[10] In the robustness section we discuss modifications using engine model dummies or year of introduction of the engine model to the market; the results remain robust to those modifications.
[11] We collect this information based on the customer ID, which we know from our database.

customer owns fewer engine models (which can be viewed as a contrast between a more focused customer and a more diversified customer) can also have an impact on product reliability since smaller number of models allows focusing maintenance on fewer and more specific tasks. We choose to report results for models with and without these two variables in the outcome equation because both *fleetsize* and *fleetmix* are important variables to explain contract choice, which can be seen as a source of exogenity to estimate the effect of PBC on reliability, i.e., they can be used as instruments if they are not included in the outcome equation.

In order to specify the choice equation, we need to include covariates that influence the type of contract selected by a customer. Results from the adverse selection models (B&D, p. 14-20) suggest that customers who intend to overuse engines and take poor care of them will be incentivized to opt for PBC contracts to begin with (i.e., a self-selection bias). The challenge is then to define variables that correspond to the causes of this self-selection problem.

Contract theory suggests that the allocation and sharing of the risk (B&D p. 7-10) induced by different contract types is a reason for self-selection by customers. Clearly, two contract types have different implications for risk allocation: under PBC, the risk is shifted entirely towards the supplier, while under T&MC the risk is shifted towards the customer. Naturally, we would expect that risk imposed on the supplier by the PBC to be factored into the price: i.e., over the lifetime of two identical engines, expected payment to Rolls-Royce is probably higher under PBC than under T&MC. Thus, PBC can be thought of as an insurance policy that creates predictable cash flows for the customer at a cost. Individual customers would then analyze this trade-off using their internal knowledge about their risk-aversion: a customer more (less) tolerant to risk might opt for T&MC (PBC). Unfortunately, and as has been recognized in previous empirical research, it is virtually impossible to measure this factor (Lafontaine and Slade 2001). While it is notoriously difficult to find good proxies for risk-aversion, one of the commonly used proxies is the size of the company. We therefore include the *fleetsize* of the customer in w_i. As noted earlier, *fleetsize* is measured as the number of engine units that a customer has registered with Rolls-Royce at the end of the observation period. Measures of customer's size have been widely used to explain contract choice in different contexts; see Lafontaine and Slade (2001), and the references therein. According to supplier managers we interviewed, *fleetsize* is probably the most relevant variable to explain contract selection, as it is observed that customers with a larger fleet are more likely to choose a PBC contract for after-sales support. This conjecture is in line with the data: the median fleet size of T&MC and PBC customers are 2 and 12, respectively. A larger fleet size is expected to be associated with greater total fleet-flying-hours (at the customer level), e.g., a customer with a fleet of 50 products is likely to have more flying hours per year than a customer with 10 products. This may cause larger firms to expect to use the MRO service more frequently, which may influence the likelihood to sign on for PBC.[12] .

In addition to *fleetsize*, a second variable that reveals some information about the customer is the complexity of its portfolio of engine types. For example, a customer might have internal expertise to service some engine types, but ownership of multiple engine types would require a complex mix of internal capabilities and therefore could lead the customer to opt for a comprehensive support arrangement through PBC. Hence, the mix of the customer product portfolio needs to be controlled for in the contract selection process. That is what we call the *fleetmix* variable, which is defined as the number of different engine models in the portfolio of a customer.

Fleetsize and *fleetmix* give a partial characterization of hidden information/adverse selection that determine the contract choice. In terms of usage, we would like to capture a pattern that describes the behavior of the owner. Overuse may manifest itself, for instance, in higher number of take-off/landing cycles (or equivalently, shorter flight time per cycle) because take-off and landing cause maximum wear and tear on the engine. This may not only decrease the reliability of the engine, as discussed earlier, but may also be a key to explaining contract choice as it reflects the usage pattern of the owner. In order to reflect the customer's overall usage pattern, we aggregate observations across the engines belonging to each owner, and calculate the average flight time for a given owner. This variable is thus included in w_i as well. Similarly, the location of the owner reflects local characteristics, e.g. prices, availability of repair shops, competition, marketing efforts by Rolls-Royce that may influence the decision to choose PBC. Therefore, we have included geographical dummies in w_i, as we did for the outcome equation. Finally, another factor that can influence contract choice for a given product is the value of the aircraft equipment. As discussed earlier, the aircraft model seems to summarize product characteristics that can be related to the type of customers that prefer either type of contract, for example, though the size of the aircraft and the type of engine that it uses. We include aircraft model dummies in the choice equation.

Note that our modeling approach allows for correlation between the unobservable effects of the outcome and contract choice equations. This feature is especially useful since we do not observe all variables related to the risk profile of the customer, which can influence both product reliability and contract choice. As we argued previously, in our case we expect that the correlation between the error terms in the two equations to be different from zero due to self-selection of the customers. In particular, we expect this correlation to be negative due to adverse selection, i.e., customers that are more likely to use engines more intensely (increasing failure risk) have more propensity to sign for PBC, which is hypothesized to increase reliability. Similar arguments have been discussed in the case of extended

[12] Of course, another major variable that is likely to affect contract choice is pricing. As noted earlier, we do not possess pricing data and it is not clear how to include such information even if it was available since PBC prices are per hour of engine or aircraft use and T&MCC prices are per part/unit of labor.

warranties for new car buyers (Padmanabhan 1995), where it is argued that heavy users have stronger incentives than light users to sign on for extended warranties, since their products are more likely to experience failures.

6. Analysis and Results

6.1 Two-stage model for the MTBUR proxy

We now turn to the estimation and results obtained for the two-stage model defined by Equations 1 and 2. The model can be estimated using the usual two-step procedure, defined e.g., in Maddala (1983, p.120-122). First, a probit model is estimated for the choice equation, where the probability that observation *i* receives the treatment (engine unit *i* covered by PBC) is given by: $\Pr(z_i = 1 | w_i) = \Phi(w_i\gamma)$. From the results of this first stage, a selectivity term can be derived as follows:

$$\text{Selectivity term}\_i = \begin{cases} \dfrac{\phi(w_i\gamma)}{\Phi(w_i\gamma)} & \text{if } z_i = 1, \\ \dfrac{-\phi(w_i\gamma)}{1-\Phi(w_i\gamma)} & \text{if } z_i = 0, \end{cases} \qquad \text{(Eq. 3)}$$

where Φ and ϕ are the cdf and the pdf of the standard normal distribution. The selectivity term calculated from the first stage is used as regressor in the outcome equation, which allows consistent estimation of the effect of the endogenous treatment, PBC, on product reliability, by accounting for the endogeneity in contract choice.

We estimate the model in STATA using the aforementioned two-step procedure. For clarity of the exposition, we focus initially on the MTBUR proxy. Table 4 displays the results obtained from the two-stage model, for the outcome equation and the MTBUR proxy. Column (1) represent the basic model specification; columns (2), (3) and (4) represent variations of the same model, including *fleetsize*, *fleetmix*, and both *fleetsize* and *fleetmix* in the outcome equation, respectively. Table 5 displays the results for the choice equation. We report clustered standard errors at the customer level in all cases. Clustered standard errors at the customer level allow for correlation between products of the same customer, while maintaining the independence assumption for products of different customers. This is particularly important in our case, as we have argued that the error terms of the outcome and choice equations involve customer unobservables. In the case of clustered standard errors at the customer level, the variance-covariance matrix of the estimates involves the computation of the sum of the interactions between the residuals and the covariates for each of the products of a given customer, which is repeated

VARIABLES	(1)	(2)	(3)	(4)
PBC	**1304****	**1151***	**1331*****	**1299***
	(498)	**(679)**	**(478)**	**(707)**
ini_age	.253	.262*	.246	.249
	(.153)	(.15)	(.157)	(.164)
ini_age_square	-0.00008	-0.000009	-0.00008	-0.00008
	(.000021)	(.0000209)	(.0000211)	(.0000222)
eng_avgflighttime	1257*	1219	1148*	1150*
	(714)	(733)	(684)	(675)
region_europe	-9.04	38.9	64.8	68
	(452)	(481)	(508)	(505)
region_other	-1402***	-1265**	-1336***	-1315**
	(378)	(568)	(419)	(559)
aircraft_model2	1141*	1076*	1232*	1211
	(583)	(603)	(684)	(790)
aircraft_model3	-1868**	-1832**	-1678*	-1687*
	(712)	(726)	(845)	(844)
selectivity_term	-721**	-637	-737**	-719
	(333)	(409)	(331)	(429)
fleetsize		.466		.0913
		(1.24)		(1.36)
fleetmix			71.9	65.9
			(163)	(188)
Constant	85.7	175	-59	-29.2
	(985)	(1068)	(1205)	(1433)
Observations	305	305	305	305

Table 4: Two-stage model for MTBUR, outcome equation. Clustered standard. errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

for each customer (see e.g. Greene 2008 pp.188-190, Wooldridge 2002 pp.328-331, for further details).[13]

Our estimate for the effect of PBC on MTBUR is positive and indicates that, on average and all else being equal, PBC increases the MTBUR of a product in the range of 1,151 to 1,331 hours, depending on the model specification under consideration. This effect is significant or highly significant for model specifications (1) and (3), and significant at the 10% level when the exclusion restriction for *fleetsize* is relaxed (model specifications (2) and (4)). For the four specifications, a 90% confidence interval for the effect of PBC on MTBUR is given by [469, 2140], [12, 2291], [529, 2133], [112, 2485], respectively.

[13] In our problem, clustered standard errors at the customer level give estimates of comparable magnitude to the ones obtained with the usual two-stage correction, and in many cases are more conservative than the latter. In particular, our conclusions on how PBC affects reliability remain unaffected by the use of the common two-stage std. error correction. We have alternatively considered clustered standard errors at the aircraft level, as it is possible to argue that products of the same aircraft type may exhibit some correlation, e.g., given that an aircraft is on-the-ground due to a product failure, a customer may be more likely to include other products of the same aircraft type in a shop visit. Results under clustered standard errors at the aircraft level are more significant than clustered standard errors at the customer level, i.e. we are reporting the most conservative case. In particular, under clustered standard errors at the aircraft level, the p-value for the PBC coefficient is 0.01, 0.085, 0.008, 0.043, for models 1 to 4, respectively.

VARIABLES	Estimates
Fleetsize	.0328**
	(.0145)
Fleetmix	-.75**
	(.378)
ow_avgflighttime	-.07
	(.91)
region_europe	.436
	(.974)
region_other	2.88***
	(1.11)
aircraft_model2	-1.91**
	(.776)
aircraft_model3	-.37
	(1.29)
Constant	.991
	(1.6)
Observations	305

Table 5: Two-stage model, choice equation. Clustered standard Errors are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

The estimate of the correlation between the error terms of the choice and outcome equations ranges from -0.34 to -0.30, depending on the model specification, in accordance with our expectation that they are negatively correlated. This is in line with our hypothesis concerning the endogeneity for the variable PBC, suggesting that the model cannot be estimated by OLS due to the self-selection in the contract choice decision.

Results in Table 4 also show that aircraft models and the geographical region of the customer are significant control variables to explain the MTBUR of an engine; the predicted effects (direction) are in line with the data. Variables that have impact at the 10% confidence level in at least one specification on the MTBUR include the initial age[14] of the product (p-value ranging from 0.087 to 0.137) and the average flight time for the engine (p-value ranging from 0.085 to 0.103). The selectivity term is significant in specifications 1 and 3, and is barely non-significant in specification 2 and 4 (p-value of 0.126 and 0.100, respectively). We also note that when *fleetsize* and/or *fleetmix* are included in the outcome equation, none of them are significant, and the coefficient of PBC remains relatively unaffected by these variations.

With respect to the choice equation, the results indicate that *fleetsize*, *fleetmix*, and the geographical region have explanatory power for the contract type of an engine. In particular, our results show that engines owned by customers with greater fleet sizes are more likely to be covered by PBC than T&MCs, in line with our hypothesis, data, and managerial expectations. Importantly, the first stage probit

[14] As will be noted in the robustness section, when other proxies are used to calculate the MTBUR, the significance of the initial age of the engine increases.

Proxy for reliability	PBC coefficient			
	(1)	(2)	(3)	(4)
MTBUE	683**	1003**	688**	1278*
	(283)	(438)	(299)	(665)
MAXMTBURMTBUE	955**	1076*	989***	1501**
	(392)	(567)	(362)	(658)
InvFRate	518	1451**	545	2394**
	(558)	(689)	(564)	(973)

Table 6: Two-stage model for alternative reliability proxies - PBC coefficients. Clustered standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

predictions are fairly reasonable. In particular, the prediction for the number of engines covered by each contract type is very close to the actual values observed in the data (predicted ratio of PBC to T&MC engines 231/74, actual 240/65). The choice model correctly predicts 90% of the observations in the sample, which is, e.g., much higher than accuracy levels reported in related applications in the literature (e.g. Terwiesch et al. 2005). It should be noted that our two-stage model does not requires instrumental variables; in fact, the δ coefficient can be estimated consistently without any instruments (Maddala 1983 p.121). However, it is recommended to have at least one variable in w_i that is excluded from the outcome equation, i.e., to act as an instrument. In our case, the strongest instrument (when excluded from the outcome regression) is *fleetsize*. The pseudo R^2 for the first stage probit is 0.44. Univariate Kolmogorov-Smirnov and t-tests shows strong significant association between *fleetsize* and PBC, and less strong association – but still significant under the Kolmogorov-Smirnov test – between PBC with both *fleetmix* and owner average flight time. We conclude that, overall, the results obtained for the first stage are fairly reasonable.

6.2 Robustness tests

Here we report results of our two-stage model for the alternative reliability proxies defined in previous sections, i.e. MTBUE, MAXMTBURMTBUE, InvFRtate. The first stage results remain unchanged. For clarity we only report results for the PBC coefficient in this section. Table 6 lists the value of the coefficient obtained for PBC and its clustered standard error, under the usual 4 model specifications, i.e., column (1) represent the basic model specification; columns (2), (3) and (4) represent variations of the same model, including *fleetsize*, *fleetmix,* and both *fleetsize* and *fleetmix* in the outcome equation, respectively.

Our main findings that PBC increases product reliability remain fairly robust to these three variations. In most cases the estimates of PBC remain significant (most of them at the 5% confidence level, and some at the 10% level). The only exceptions occur when the InvFRate proxy is used in model specifications that exclude *fleetsize* from the outcome equation. Thus, when found significant, our estimates indicate an effect for PBC on product reliability roughly in the 20%-40% range of improvement, consistently across alternative proxies, including also the MTBUR proxy, where we found estimates in the 35%-40% range. The effects found for the rest of the explanatory variables are largely the same as the ones reported for the MTBUR; the only important difference is that the effect of the initial age of the engine – and sometimes its quadratic term - is found to be significant in most of the cases under the alternative proxies, and was found to be only significant at the 10% confidence level or barely non-significant in the case of the MTBUR. Overall, results obtained with a two-stage model seem very consistent across different proxies for product reliability.

Finally, here we briefly report some additional variations to further validate our results. The first issue we address is related to the variable used to control for product type. Throughout the paper, we included dummy variables for the aircraft model to capture product fixed effects. As we mentioned in earlier sections, there are different ways to control for product characteristics, e.g., engine model, engine type, or some relevant characteristic of the engine model. They cannot be included simultaneously in the model, as different proxies are almost perfectly collinear. As a robustness check, we estimated our model using alternately the 5 engine models, the 2 engine types, and the number of years since the engine model was introduced into the market. This last variable is a reasonable way to summarize the effect of engine characteristics on product reliability, e.g., it can be hypothesized that less is known about engine models that were introduced more recently to the market, which could influence reliability and how likely are these engines to be covered by PBC. We evaluate these models for the four alternative proxies of product reliability, under the four usual model specifications. Our results remain qualitatively the same.

The second issue is the use of the initial age of the engine in the outcome regression. As we noted, this variable is not available to us, and we used extrapolation to obtain an estimate of it. A concern that can be raised is that we are using this variable as an input to calculate our dependent variable, as well as an independent variable. While we do not believe that this is critical in our case – our proxies for product reliability would be less meaningful if we did not take into account the initial age of the engine, to start with – we try a different proxy for capturing the initial conditions of the engine. For that, we collect the year of production of each aircraft from the Federal Aviation Administration (FAA) and related websites using the aircraft serial number.[15] Using it, we construct a variable reflecting the number of years since

[15] We obtained data on this variable from aeronautic institutions from different countries. Websites consulted include: USA (FAA database, http://www.faa.gov), FRANCE (Direction Générale de l'Aviation Civile, http://www.immat.aviation-civile.gouv.fr/immat/servlet/aeronef_liste.html#), UK (Civil Aviation Authority http://www.caa.co.uk/application.aspx?catid=60&pagetype=65&appid=1), SWEEDEN (Transport Styrelsen, http://www.luftfartsstyrelsen.se/templates/LS_LuftFartyg_Sok____39453.aspx), . For other countries like BRAZIL, ITALY, BELGIUM, SOUTHAFRICA, CANADA, THAILAND, etc., we used different industry websites, e.g. http://www.airport-data.com, www.airframes.org, http://www.aerotransport.org/php/querybuilder.php?tab=regn, http://www.planespotters.net/Production_List/. Overall, we were able to collect data for this variable for 303 out of 305 engines in our sample.

the aircraft was manufactured with respect to 2002 (the beginning of the observation period). While we believe that the initial age of the engine (measured in flying hours) is a more precise proxy for the initial conditions of the engine, the age of the aircraft can capture some relevant initial conditions of the engine. We use age of the aircraft instead of initial age (Equation 2) to estimate all four model specifications. We do not observe any notable changes in the results: in particular, the effect of PBC on product reliability remains significant in a majority of the cases, with estimated coefficients in the same range as before.

Thus far, our analysis has relied on constructing proxies for the mean time between unplanned removals of each engine unit and performing two-stage estimation. As mentioned earlier, a duration model offers an alternative way to analyze our data. Duration modeling is a nonlinear modeling approach that allows estimating the rate at which events – unplanned removals in our case – occur. In fact, the sample in our dataset represents *multiple spells* since some products have more than one unplanned removal. The main advantage of this estimation technique is that it has a built-in mechanism to deal with some of the censoring issues, which are common in duration data such as ours. Unfortunately, research on how to incorporate the endogeneity issue in the duration models is still ongoing, and currently there is no agreed upon method or statistical package that we can adopt for our purposes (see Bijwaard 2007 for a recent contribution to this research stream). However, the following informal approach has been used by some researchers: (1) run the probit model (Equation 2) to predict contract choices, (2) calculate the selectivity term from this analysis (Equation 3), and (3) perform duration analysis using the selectivity term as one of the regressors. A similar approach in the context of sample selection for duration models was used by Rao et al. (2001), based on the generalization of the Heckman selection model proposed in Lee (1983). Although consistency of this approach is not, to our knowledge, yet fully established, we use this procedure in this subsection as a robustness check, which can be taken as complementary evidence with the caveat on endogeneity as described above.

In order to proceed with this approach, we must analyze the data at the removal level (instead of at the product level), and we must examine the influence of contract type on the respective unplanned removal rates. As is standard in duration analysis, we conduct experiments using both semi-parametric (Cox) and parametric (exponential, log-logistic, and log-normal) transition rate models. The explanatory variables are the same that were used in the two-stage regression models; the only modification in explanatory variables we incorporated is to replace the initial age of the product with the age of the product at the time of the previous unplanned removal (if any) in Equation 1.[16] We estimate the models using clustered standard errors at the customer level. For simplicity and readability, we focus on our main hypothesis that PBC increases product mean time between unplanned removals, and report results

[16] For example, for the 3rd removal of a product unit, instead of the initial age (age of the product at the beginning of the sample period) we include the age of the product unit at the moment of the 2nd removal (for both the linear and quadratic terms).

Duration model	PBC coefficient			
	(1)	(2)	(3)	(4)
Cox	-0.308	-0.483**	-0.323	-0.710**
	(0.22)	(0.24)	(0.25)	(0.34)
Exponential	-0.284*	-0.521***	-0.276	-0.659***
	(0.17)	(0.18)	(0.19)	(0.23)
Log-Logistic	-0.388**	-0.602***	-0.383*	-0.693***
	(0.19)	(0.21)	(0.20)	(0.23)
Log-Normal	-0.419**	-0.681***	-0.408*	-0.798***
	(0.19)	(0.22)	(0.21)	(0.28)

Table 7: PBC coefficient in duration models. Clustered standard Errors are in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

for the coefficient of the PBC variable only. Table 7 summarizes the results obtained for the usual four model specifications. As before, column (1) represent the basic model specification; columns (2), (3) and (4) represent variations of the same model, including *fleetsize*, *fleetmix*, and both *fleetsize* and *fleetmix* in the outcome equation, respectively.

Note that the negative coefficients for PBC implies that PBC decreases the rate at which unplanned removals occur, or equivalently, PBC increases the time between unplanned removal events and the size of the effect is comparable to our earlier findings. The effect of PBC is significant in most of the cases, with the exception of model specifications (1) and (3) for Cox models, and model specification (3) for the exponential model. In addition, as apparent from the results in Table 6, the PBC coefficient changes somewhat with the inclusion of *fleetsize*, i.e., our models seem to react quite sensitively to the inclusion of this variable in the outcome equation (model specifications 2 and 4), which was not the case in our two stage models presented in previous sections. This is perhaps not surprising, as duration models have an event – unplanned removal – as unit of analysis, while the selectivity term (driven largely by *fleetsize*) is calculated at the engine level (Equation 2).

Instead of imposing a duration model *a priori*, our approach has been to test the main hypothesis under different types of duration models, as reflected in Table 7. Among these models, Cox models are semi-parametric (the shape of the transition rate is not imposed a priori), and therefore they allow more flexibility than the rest of the models which make further assumptions on the distribution of the hazard rate at which event occurs. While more flexible in this sense, Cox models assume proportionality, i.e., that the transition rates for different values of the covariates are proportional. Further analysis of the data reveals that the proportionality assumption for Cox models is less plausible for model specifications that exclude *fleetsize* from the outcome equation (columns 1 and 3 in Table 7), suggesting that the models that include *fleetsize* (columns 2 and 4) may be more appropriate under Cox models. For the other (parametric) models, both log-logistic and log-normal are more flexible than the exponential model, as the latter imposes an assumption of the monotonically decreasing hazard rate, while the former models allow for non-monotonicity. In fact, the log-logistic distribution is also more flexible than other distributions such as Gompertz and Weibull (Blossfeld et al. 2007). Our results and data indicate that a non-monotonic hazard rate function is more appropriate in our case, so that results of the log-logistic and log-normal models are more relevant than the other, and both of these models support largely the same conclusions and estimates for the effect of PBC on product reliability as obtained before.

We conclude that our estimates for the effect of PBC on product reliability are reasonably robust to many variations in model specification, reliability proxies, and econometric methodologies. The main finding that PBC has a positive and significant marginal effect on product reliability remains robust in most of the cases.

7. Discussion of limitations and extensions

Although we have shown that there is relevant evidence to support our conclusions, our analysis is not free of limitations. In particular, there are several issues stemming from the nature of the data. First, while we are concerned with product reliability, our ability to accurately measure it is quite limited as the frequency of failures in our sample is very low. Failures of aircraft engines are infrequent events, as is typical in practice, even though we observe the system for a relatively long period of 5 years. This problem creates imprecision in the definition of the dependent variable, the mean time between unplanned removals: essentially, there is no obvious "correct" way to measure our dependent variable. We have attempted to deal with this issue by performing experiments under several alternative proxies for the mean time between unplanned removals, and by alternative modeling approaches (two-stage models, duration model analysis). Thus, we have dealt with the censored nature of our data by exploring different ways of treating that issue, and found results that are consistent overall. However, if data becomes available over a longer time span, our results would have to be revisited.

Second, in our sample we observe a small proportion of products covered by T&MCs (21.3%), which may suggest the risk of sample selection bias. Recall that we use data from only one supplier in the market, so it is possible that customers that prefer a T&MC scheme choose a different supplier for their repair and maintenance service. Naturally, we do not observe such customers, and we do not have any basis to construct a model to explain a potential sample selection of suppliers by customers. A recent note in Business Wire (2008), however, indicates that our sample is representative of the total population, for which engines covered by PBC represent about 80%. While the uneven proportion of T&MC/PBC engines does not appear to raise a significant concern in terms of the results we obtained (recall that the proportion of engines under each contract regime predicted by our model is very accurate), we conduct further experiments estimating our model with balanced samples (50% PBC, 50% T&MC) by appropriately weighting the observations for the T&MC pool, in order to make sure that the uneven proportion of PBC and T&MC products in our sample is not playing a role in our results. Results obtained in these experiments are similar to the ones obtained with the original sample, and the effect of PBC on reliability is found to be significant and in similar ranges. Certainly, obtaining a bigger dataset, possibly from multiple OEMs, would further help eliminate concerns about sample bias.

Third, while our dataset is rich in terms for characterizing the removal incidents for a given product, we have only limited data to characterize a customer. In particular, unobservables related to customer risk profile and behavior might have an impact on our results. We include in our analysis the *fleetsize*, *fleetmix*, average length of a flight, and the region of the owner as a way to account for some relevant customer characteristics, which somewhat alleviate these concerns. It is also true that our dataset is not rich in describing the specific terms of the contracts in each case; we only distinguish between T&MCs and PBC. This does not allow us to explore the influence of price and other contract conditions on the customer's contract choice, although the supplier suggested to us that there are no major differences in contract parameters: most of them are signed at list prices. We also do not observe data before/after adoption of PBC, which would have made possible to study the dynamics, e.g., by using difference-in-differences estimators. Finally, due to a small number of customers in our sample, we are unable to conduct analysis at the customer level but instead we control for relevant customer characteristics and account for possible correlations across the engines owned the same customer using clustered standard errors.

With respect to our modeling approach, the main assumption we impose is the exogeneity of the independent variables (other than PBC) in our model specification. In essence, we assume that the unobservable characteristics related to customer risk profile and behavior are uncorrelated with the included product and customer characteristics. This assumption rules out the possibility that customer behavior may vary depending on the initial age of the product or that customer risk exposure will change with e.g. *fleetsize*. Thus, while our approach explicitly deals with the endogeneity of contract choice, our results need to be understood in the context of the exogeneity assumption for the rest of the covariates.

Finally, we have analyzed the unplanned removals of the engines, and calculated alternative proxies for product reliability, for products that were removed from an aircraft at least once during the observation period. Thus, our analysis is driven by the nature of our dependent variable, the mean time between unplanned removals. Our analysis is silent, however, with respect to those products that were never removed, or for which we observed planned removals only. With respect to the latter, we have already pointed out that the full sample of engines is very similar to the sample of engines with unplanned removals only: 39.9% of the T&MC engines had unplanned removals, and 40% of the PBC engines had

Variable	Overall sample		T&MC only		PBC only	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
MTBPR	3,999	2,004	4,602	2,309	3,848	1,892
MTBPE	3,599	1,648	4,271	1,899	3,430	1,535
MAXMTBPRMTBPE	4,327	1,921	5,069	2,232	4,140	1,789
InvFRateP	6,675	3,114	8,034	3,709	6,332	2,849

Table 8: Mean time between planned removals: statistics for alternative proxies

unplanned removals. This suggests that by focusing on the sample of engines with unplanned removals, there is no obvious bias. Moreover, based on this fact we have no reason to expect that sample selection with respect to engines that were not removed in the observation period – and hence, unobserved to us – is a relevant issue in our problem. Further, as noted in previous sections, a possible explanation for PBC increasing the mean time between unplanned removals, is related to behavior concerning planned removals, i.e., PBC may increase reliability through better (or more frequent) scheduled maintenance, as indicated by the planned removals. For the sample of engines with planned removals, we calculate the mean time between planned removals, using the same four proxy formulations employed for unplanned removals, but now considering planned removals (i.e., planned instead of unplanned removals are used in the calculations). We use the labels MTBPR, MTBPE, MAXMTBPRMTBPE, InvFrateP, to denote the four proxies applied to planned removals. Table 8 displays summary statistics for these proxies.

Table 8 shows that the mean time between planned removals is greater for T&MC engines than for PBC engines, suggesting that PBC engines are subject to more frequent preemptive maintenance regardless of the proxy. We attempted to build two-stage models to test whether there is an effect of PBC on the mean time between planned removals but found no consistently significant effect of PBC in this case. This is consistent with our earlier explanation that planned removals are largely pre-determined by regulated usage milestones and contract terms. Thus, attempting to fit a model to explain planned removals similar to the one we developed for unplanned removals, does not appear to be appropriate. Thus, we only report the evidence in Table 8 to support the explanation that our result concerning the increase in engine reliability (as indicated by MTBUR) due to PBC could be related to the execution of more frequent pre-emptive maintenance, as measured by the mean time between planned removals (MTBPR). We also attempted studying the time interval between a planned and a subsequent unplanned removal but there are only about 50 such observations in our data, covering only a handful of customers, rendering such analysis infeasible

8. Conclusions

We have examined the impact of performance-based contracts on product reliability in an application to the aerospace and defense maintenance and repair services industry. Using a proprietary dataset from

Rolls-Royce, a major supplier of aircraft engines, we proposed and analyzed a model employing the well accepted two-stage approach that allows us to explicitly account for the endogeneity inherent in contract choice by a customer. The first stage of the econometric model describes the customer decision with respect to selecting a contract and the second stage analyzes the impact of contract type on product reliability. Our analysis shows that there is a positive and statistically significant effect of PBC on the MTBUR of a product, i.e., performance-based contracts induce improvements in product reliability in our sample. Our estimates indicate reliability improvements under PBC in the 20-40% range, in comparison to traditional T&MCs. We also show that endogeneity of contract choice is clearly an issue in our case, and our proposed statistical framework account for it explicitly. Indeed, we note that the impact estimated by our analysis is about double the impact suggested when endogeneity is not considered. These findings are supported by numerous robustness checks under a number of alternative model specifications and modeling approaches.

Our analysis focuses on the marginal effect of performance-based contracts on product reliability. Our results provide a first step towards understanding the overall impact of performance-based contracts, and our approach was, to a large extent, driven by data availability. The availability of richer data about customers, financial and managerial information, and the specific contract terms between the customer and the supplier, would enable a more complete analysis to cover a number of open questions. Such analysis, could lead to a deeper understanding of the benefits of PBC contracts, e.g., what drives reliability improvement? Is this reliability improvement profitable to the supplier? Does the cost of pre-emptive maintenance exceed the benefits due to reliability improvement? Is the price charged to the customer appropriate? How do specific contract terms moderate the impact on reliability? These questions remain open for future research.

This paper is one of the few studies that empirically estimate the impact of a performance vs. non-performance contract type and other causal factors on supply chain outcomes. While we cannot claim that the conclusions obtained in our study of a particular case are applicable in all settings, we are confident that our findings will be of interest not only to the aircraft repair and maintenance service industry but also to all industries that provide after-sales support for mission critical products. The results are especially relevant for practitioners since this is the first attempt to test the reliability improvement hypothesis for performance contracting based on transactional data. While there are hundreds of papers that propose sophisticated analytical models of various supply chain contracts, there is little empirical evidence of the impact of such contracts on supply chain outcomes. We thus believe that our paper makes a step in closing the gap between theoretic modeling and empirical evidence.

Acknowledgement

This research was partially funded by the Fishman Davidson Center for Service and Operations Management, The Wharton School, University of Pennsylvania. The authors would also like to acknowledge the financial support and active participation in the project by Rolls-Royce, who provided the data used in this study. We would, in particular, like to acknowledge the assistance of Miles Cowdry and W. Bunker Snyder who supported the project as well as James J. Olsen, Lindsey E. Shackelford and Jarod Downing who assisted us in collecting the data. Finally, we acknowledge the financial support and input from the US Air Force, (in collaboration with the University of Tennessee), who is funding our ongoing research in this area. We are also thankful to the following individuals for their extensive comments on the earlier version of this paper: Donald Lee, Richard Lai, Marcelo Olivares, Justin Ren, Enno Siemsen, Rachna Shah, Manuel Sosa and Owen Wu.

References

Ackerberg, D., M. Botticini. 2002. Endogenous matching and the empirical determinants of contract form. *Journal of Political Economy*, 110(3), 564-591.

Banker, R., S.Y. Lee, G. Potter, D. Srinivasan. 2001. An empirical analysis of continuing improvements following the implementation of a performance-based compensation plan. *Journal of Accounting and Economics*, 30(3), 315-350.

Blossfeld, H-P, K. Golsch, G. Rohwer. 2007. *Event history analysis with Stata*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Bolton, P., M. Dewatripont. 2005. *Contract theory*. The MIT Press, Cambridge, MA.

Brijwaard, G. E. 2007. Instrumental variable estimation of treatment effects for duration outcomes. Working paper.

BusinessWire. 2008. Rolls-Royce Nets CorporateCare Contract with NetJets Middle East. November 19, 2008. http://www.businesswire.com/news/home/20081119005755/en.

Cachon, G. 2003. Supply chain coordination with contracts. In: S. Graves and T. de Kok, Editors, *Handbooks in Operations Research and Management Science: Supply Chain Management*, North Holland, Amsterdam, The Netherlands.

Cameron, A. C., P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Cohen, M. A., N. Agrawal, V. Agrawal. 2006. Winning in the Aftermarket. *Harvard Business Review*, 84(5), 129-38.

Cohen, M. A., V. Agrawal, Y-S Zheng. 1997. Service parts logistics: a benchmark analysis. *IIE Transactions*, 29(8), 627-639.

Crocker, K., K. Reynolds. 1993. The efficiency of incomplete contracts: an empirical analysis of Air Force engine procurement. *RAND Journal of Economics*, 24(1), 126-146.

Deshpande, V., M. A. Cohen, K. Donohue, 2003. An empirical study of service differentiation for weapon system service parts. *Operations Research*, 51(4), 518–530.

Geary, S., K. Vitasek. 2008. *A Contractor's Guide to Life Cycle Product Support Management*. Supply Chain Visions.

Government Accountability Office. 2008. Improved analysis and cost data needed to evaluate the cost-effectiveness of Performance Based Logistics. GAO-09-41.

Gopal, A., K. Sivaramakrishnan, M. S. Krishnan, T. Mukhopadhyay. 2003. Contracts in offshore software development: An empirical analysis. *Management Science,* 49(12), 1671–1683.

Gopal, A., K. Sivaramakrishnan. 2008. On vendor preferences for contract types in offshore software projects: the case of fixed price vs. time and materials contracts. *Information Systems Research,* 19(2), 202-220.

Greene, W. H. 2008. *Econometric Analysis*. Sixth edition, Prentice-Hall, Upper Saddle River, NJ.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica*, 47(1), 153-161.Heinrich, C. J. 2002. Outcomes-Based performance management in the public sector: Implications for government accountability and effectiveness. *Public Administration Review*, 62(6), 712-725.

Heinrich, C. J., Y. Choi. 2008. Performance-Based Contracting in Social Welfare Programs. The *American Review of Public Administration*, 37(4), 409-435.

Ho, J., K. Ho, J. Mortimer. 2008. Welfare effects of full-line forcing contracts in the video rental industry. Working paper.

Holstrom, B. 1979. Moral hazard and observability. *The Bell Journal of Economics*, 10(1), 74-91.

Hubbard, T. N. 1988. An empirical examination of moral hazard in the vehicle inspection market. *The RAND Journal of Economics*, 29(2) 406-426.

Ioannou, I., J. Mortimer, R. Mortimer. 2009. The effects of capacity on sales under alternative vertical contracts. Working paper.

Kesavan, S., V. Gaur, A. Raman. 2008. Incorporating price and inventory endogeneity in firm-level sales forecasting. Working paper, University of North Carolina.

Kirk, R. L., T. J. DePalma. 2005. Performance-based logistics contracts: a basic overview. CNA Corporation. November 2005. Available at http://www.cna.org/documents/D0012881.A2.PDF (accessed April 23rd 2009).

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

Kim, S.-H., M. A. Cohen, S. Netessine. 2009a. Reliability or inventory? Analysis of product support contracts in the defense industry. Working paper.

Kim, S.-H.., M. A. Cohen, S. Netessine, S. Veeraraghavan. 2009b. Contracting for infrequent restoration and recovery of mission-critical systems. Working paper.

Lazear, E. P. 2000. Performance pay and productivity. *American Economic Review*, 90(5), 1346-1361.

Lafontaine, F., M. Slade. 2001. Incentive contracting and the franchise decision. In *Advances in Business Applications of Game Theory,* K. Chatterjee and W. Samuelson (eds.), Kluwer Academic Press.

Lafontaine, F., K. L. Shaw. 1999. The dynamics of franchise contracting: evidence from panel data. *The Journal of Political Economy*, 107(5), 1041-1080.

Lee, D. K. K. and S. A. Zenios. 2007. Evidence-based incentive systems with an application in health care delivery. Working paper.

Lee, L. F. 1983. Generalized economic models with selectivity. *Econometrica*, 51(2), 507-512.

Lu, M., C. Ma, L. Yuan. 2003. Risk selection and matching in performance-based contracting. *Health Economics*, 12(5), 339-354.

Maddala, G. S. 1983. *Limited-Dependent And Qualitative Variables in Econometrics*. Cambridge University Press. Cambridge, UK.

Masten, S., S. Saussier. 2002. Econometrics of contracts: an assessment of developments in the empirical Literature on contracting. Ch.16, In: *The Economics of Contracts-Theories and Applications* (E. Brousseau and J-M Glachant, eds.), Cambridge University Press.

Mortimer, J. H. 2008. Vertical contracts in the video rental industry. *Review of Economic Studies*, 75 165-199.

Novak, S., S. D. Eppinger. 2001. Sourcing by design: Product complexity and the supply chain. *Management Science*, 47(1) 189-204.

Novak, S., S. Stern. 2008. How does outsourcing affect performance dynamics? Evidence from the automobile industry. *Management Science*, 54(12), 1963-1979.

Novak, S., S. Stern. 2009. Complementarity among vertical integration decisions: evidence from automobile product development. *Management Science*, 55(2), 311-332.

Padmanabhan, V. 1995. Usage heterogeneity and extended warranties. *Journal of Economics & Management Strategy*, 4(1), 33-53.

Prendergast, C. 1999. The provision of incentives in firms. *Journal of Economic Literature*, 37(1), 7-63.

Prendergast, C. 2002a. The tenuous tradeoff between risk and incentives. *Journal of Political Economy*, 110(5), 1071-1102.

Prendergast, C. 2002b. Uncertainty and incentives. *Journal of Labor Economics*, 20(2), 1071-1102.

Ramdas, K., T. Randall. 2008. Does component sharing help or hurt reliability? An empirical study in the automotive industry. *Management Science*, 54(5), 922-938.

Rao, H., H. R. Greve, G.F. Davis. 2001. Fool's gold: Social proof in the initiation and abandonment of coverage by Wall Street analysts. *Administrative Science Quarterly*, 46(3), 502-526.

Randall, T., S. Netessine, N. Rudi. 2006. An empirical examination of the decision to invest in fulfillment capabilities: A study of Internet retailers. *Management Science*, 52(4), 567–580.

Rausand, M., A. Høyland. 2004. *System Reliability Theory: Models, Statistical Methods, and Applications*. Wiley, New Jersey.

Shaver, J. M. 1998. Accounting for endogeneity when assessing strategy performance: does entry mode choice affect FDI survival? *Management Science*, 44(4), 571–585.

Shen, Y. 2003. Selection incentives in a performance-based contracting system. *Health Services Research*, 38(2), 535-552.Sherbrooke, C. C. 1992. *Optimal Inventory Modeling of Systems: Multi-Echelon Techniques*. Wiley.

Slade, M. E. 1996. Multitask agency and contract choice: an empirical exploration. *International Economic Review*, 37(2), 465-486.

Standard & Poor's. 2009. Industry Surveys: Aerospace and Defense (February 12th).

Terwiesch, C., J. Ren, T. H. Ho, M. A. Cohen. 2005. An empirical analysis of forecast sharing in the semiconductor equipment supply chain. *Management Science*, 51(2), 208-220.

Thomas, T. J. 2005. Performance logistics now coming to Australia. *Australian Defense Business Review*, 24(7), 23-24.

Toffel, M. W. 2002. Contracting for servicizing. Working paper, Harvard University.

Ton, Z., A. Raman. 2009. The effect of product variety and inventory levels on sales: a longitudinal study. *Production and Operations Management Journal*, forthcoming.

U.S. Government Accountability Office. 2008. Defense logistics: improved analysis and cost data needed to evaluate the cost-effectiveness of Performance-Based Logistics, GAO-09-41.

Wooldrigde, J. M. 2002. *Econometric Analysis of Cross-Section and Panel Data*. MIT Press, Cambridge, MA.