MANAGEMENT SCIENCE
Vol. 54, No. 3, March 2008, pp. 594–607
ISSN 0025-1909 | EISSN 1526-5501 | 08 | 5403 | 0594

informs

DOI 10.1287/mnsc.1070.0750

Allocation Models and Heuristics for the Outsourcing of Repairs for a Dynamic Warranty Population

Li Ding
Durham Business School, Durham University, Durham DH1 3LB, United Kingdom, li.ding@durham.ac.uk

Kevin D. Glazebrook, Christopher Kirkbride
Department of Management Science, Lancaster University, Lancaster LA1 4YX, United Kingdom
{k.glazebrook@lancaster.ac.uk, c.kirkbride@lancaster.ac.uk}

We consider a scenario in which a large equipment manufacturer wishes to outsource the work involved in repairing purchased goods while under warranty. Several external service vendors are available for this work. We develop models and analyses to support decisions concerning how responsibility for the warranty population should be divided between them. These also allow the manufacturer to resolve related questions concerning, for example, whether the service capacities of the contracted vendors are sufficient to deliver an effective post-sales service. Static allocation models yield information concerning the proportions of the warranty population for which the vendors should be responsible overall. Dynamic allocation models enable consideration of how such overall workloads might be delivered to the vendors over time in a way which avoids excessive variability in the repair burden. We apply dynamic programming policy improvement to develop an effective dynamic allocation heuristic. This is evaluated numerically and is also used as a yardstick to assess two simple allocation heuristics suggested by static models. A dynamic greedy allocation heuristic is found to perform well. Dividing the workload equally among vendors with different service capacities can lead to serious losses.

Key words: approximate dynamic programming; greedy heuristics; index policies; outsourcing; warranty repairs

History: Accepted by Wallace J. Hopp, stochastic models and simulation; received June 28, 2005. This paper was with the authors 2 years and 2 months for 3 revisions. Published online in *Articles in Advance* November 30, 2007.

1. Introduction

The recent past has seen considerable growth in the outsourcing by equipment manufacturers (particularly those in the computing industry) of the work involved in undertaking repairs to products during their warranty period. Opp et al. (2003) quote a Merill Lynch report (Serant 2001) to the effect that this trend represents a one hundred billion dollar opportunity for subcontractors and service vendors. Such outsourcing enables manufacturers to focus on their core business and saves the costs involved in maintaining in-house repair facilities. However, it does expose manufacturers to the risk of poor post-sales service resulting in customer dissatisfaction which may be expressed, inter alia, in lost sales. Contact by colleagues with a large equipment manufacturer uncovered a situation in which several external vendors were contracted to undertake warranty repairs. Such a situation is not uncommon.

A range of recent articles and surveys have highlighted risk reduction, geographic coverage, and bearing down on cost as important factors in the recent trend toward the use of multiple vendors when outsourcing business processes. Briskman (2005, p. 2) states that "multivendor situations can lower risk.... Certainly there are cases where more than one vendor is selected for a single service." A report in ZDNet News (2005, p. 1) quotes a Deloitte Consulting LLP study to the effect that "73% of the participants are working with multiple vendors to reduce vendor dependency." It is certainly the case that in an area as sensitive to customer relations as warranty repairs, a manufacturer may not wish to take the risk of being totally reliant on a single service vendor. See, for example, the related comments in the Aberdeen Group's 2005 study "Best Practice in Strategic Service Management" (cited by Violino 2006). Further, in their study of a model for the outsourcing of warranty repairs, Buckowski et al. (2005) cite a desire to increase the geographic reach of (high-quality) service as a rationale for contracting several repair vendors. This was also cited by McDougall (2005, p. 1) as a factor in ABN Amro's move to multiple vendors. He adds that a company wants "enough service providers familiar with the company and its business so that it's possible to shift work among them and keep all the vendors competing for new work." More simply than geographic coverage, the volume of a

company's output may be such that there may not be a single vendor capable of handling all warranty repairs. Business Outsourcing Corporation (2006) cite a case study in which a major computer company contracted two service vendors to undertake a large volume of emergency repairs to machines under warranty.

As far as the equipment manufacturer cited in the opening paragraph was concerned, many of the decisions relating to the management of the repair workload appeared to be taken in a somewhat ad hoc fashion. In this paper, we propose analyses which shed light on a range of important questions which could inform the manufacturer's decision making in this area. These include: What level of service capacity among contracted vendors needs to be available to meet the anticipated demand for the manufacturer's post-sales repair service effectively? Given that the manufacturer's contracted vendors do possess sufficient service capacity, how should the repair work be best distributed among them? How much might the manufacturer be losing (economically and in customer goodwill) by maintaining an existing suboptimal approach to workload distribution? It is the second of these questions which holds the key to answering the other two and whose study is the central focus of this paper. One way of thinking about this workload distribution question is in two stages: at the first stage, consider the (simpler static optimization) problem of determining the proportions of the warranty population for which the respective vendors should be responsible overall. At the second stage consider the (more complex dynamic optimization) problem of how those overall proportions might be delivered to the vendors in a way which reduces the extent to which their workloads vary over time. This will in turn reduce the chance of excessive repair queue lengths causing unacceptable response times for customers.

In response to the first-stage problem above, Opp et al. (2003) and Ding and Glazebrook (2005) have formulated simple static allocation models in which it is supposed that there is a fixed number (K) of items under warranty for all time. These are to be divided between the (V) vendors, vendor v receiving a fixed allocation (k_v) of the items which will be under the vendor's care. The problem of determining vendor allocations to minimize an overall cost rate is formulated as a static optimization problem whose objective is typically convex and separable, and hence which is often solvable by a greedy heuristic. See, for example, Gross (1956), Ibaraki and Katoh (1988), and Fox (1966). However, such static formulations do not do justice to the dynamic and stochastic nature of the population of items under warranty where new items arrive when purchased (whether singly or in batches) and depart when their warranty periods expire. In other words, static formulations cannot shed any direct light on the second-stage dynamic optimization problem identified at the conclusion of the preceding paragraph.

The scenario we consider is described in detail in §2 and formulated as Model 1. In outline, new equipment purchases are made according to a compound Poisson process. All items within a single order are assumed to be allocated to a single vendor who will carry out all repairs on those items until the expiry of their warranty period when they leave the system. We discuss how this assumption may be relaxed in §4 after the main analysis. The date and size of each order and the vendor to which it is allocated are all logged, and these data form the basis of all subsequent allocation decisions. In particular, whenever an allocation decision is to be made, the number of items under warranty at each vendor is known along with the amount of time remaining of the warranty period for each item. However, the decision maker is not able to observe the current repair queue at each vendor which is the locus at which costs are actually incurred. Continuous observation of all vendor repair queues would involve a substantial administrative overhead but results in a simpler fully observable stochastic dynamic optimization problem (Opp et al. 2005). That analysis is of a relatively conventional model concerning the dynamic routing of incoming items to alternative service stations to minimize average cost rates over an infinite horizon. While such problems are known to be very difficult, there is at least a substantial literature devoted to them. See, for example, Hordijk and Koole (1990), Weber (1978), and Winston (1977).

As will become clear in §2, Model 1 gives rise to a nonstandard stochastic and dynamic optimization problem which is challenging to solve. It has a number of features which make conventional use of dynamic programming (DP) for its solution unrealistic. Following a brief discussion in §3 of a static allocation model (Model 2) and of dynamic allocation heuristics which may be inferred from it, our primary analysis is contained in §4. Here we adopt a two-stage approach to design an allocation heuristic which makes full use of system-state information. At the first stage we design an optimal static allocation using an approximating model (Model 3). This establishes an appropriate proportion of work overall which should be directed to the respective vendors. At the second stage, we apply a single DP policy improvement step. The resulting dynamic heuristic makes allocation decisions in light of values of calibrating indices for the vendors which are functions of all of the available data. This index heuristic has the effect of allocating newly arriving items to the

vendors whose current workload is relatively low when due account is taken of their service capacity. It is both a policy of interest in its own right and also provides a benchmark against which other simpler procedures are assessed. A simple dynamic greedy policy is found to perform strongly in a simulation study reported in §5. We also assess the sensitivity of our conclusions to model assumptions.

2. An Allocation Model for a Dynamic Warranty Population (Model 1)

Purchases of a finished good (item) occur according to a compound Poisson process (η, F). The positive real η is the rate at which orders for the item occur, while F is the cumulative distribution function (c.d.f.) of the order size. We use $X \sim F$ for a generic order size, a positive integer-valued random variable with mean θ and finite second moment θ_2. Upon receipt of an order, a decision must be made concerning which one of V vendors should be responsible for repairing all the items in that order during the ensuing warranty period (W years). Later in the paper, we consider the possibility that items in an order might be sent to several vendors. Once a decision is made, the order size, purchase date, and vendor chosen are recorded and are available to inform future decisions.

The following information (current at the time of receipt of any order) is available for the allocation decisions described above:

(i) the size of incoming order (x);

(ii) the number of items currently under warranty at vendor v (N_v) along with the durations $(t_n^v, 1 \le n \le N_v)$ of their unexpired warranties, $1 \le v \le V$.

Given that items within the same order have identical unexpired warranties, this information may alternatively be presented as

$$(\mathbf{x}^v, \mathbf{t}^v) \equiv \{(x_1^v, t_1^v), (x_2^v, t_2^v), \ldots, (x_{M_v}^v, t_{M_v}^v)\}, \quad (1)$$

where the x_m^v, $1 \le m \le M_v$, are order sizes and the t_m^v, $1 \le m \le M_v$, are the corresponding durations of unexpired warranties, numbered such that

$$W \ge t_{M_v}^v > t_{M_v-1}^v > \cdots > t_2^v > t_1^v \ge 0, \quad 1 \le v \le V.$$

We have

$$N_v = \sum_{m=1}^{M_v} x_m^v,\ 1 \le v \le V, \quad \text{with } N = \sum_{v=1}^{V} N_v,$$

the total number of items under warranty. All of the quantities $\mathcal{N} \equiv \{N_v, (\mathbf{x}^v, \mathbf{t}^v), 1 \le v \le V\}$ evolve through time driven by the dynamics of the order process (η, F) and the allocation decisions. Standard results indicate that, once the above system has been in operation for (at least) time W, the mean and the variance of the total number of items under warranty are given by

$$E(N) = \eta W \theta \quad (2a)$$

and

$$\mathrm{var}(N) = \eta W \theta_2. \quad (2b)$$

The breakdown/repair process for items at each vendor is Markovian and not observable by the decision maker. We suppose that at time t, $N_v(t)$ items are under warranty at vendor v, with $D_v(t)$ the number awaiting or undergoing repair (down) and $U_v(t) = N_v(t) - D_v(t)$ the number which are functioning satisfactorily (up). In the absence of new arrivals at vendor v, the rate associated with the transition $\{D_v(t), U_v(t)\} \to \{D_v(t) + 1, U_v(t) - 1\}$ (a breakdown) is $\lambda U_v(t)$, while that associated with the transition $\{D_v(t), U_v(t)\} \to \{D_v(t) - 1, U_v(t) + 1\}$ (a repair) is $\mu_v \min\{\sigma_v, D_v(t)\}$. Equivalently, items draw successive up-times independently from an exp(λ) distribution, with $\lambda > 0$ the breakdown rate for individual items. Further, repairs are effected at vendor v by σ_v repairers working in parallel, each at rate $\mu_v > 0$, $1 \le v \le V$. We suppose that repairs are carried out on a first-come-first-served basis and that any item which breaks down during its warranty period will have its repair completed. See Opp et al. (2003) and Ding and Glazebrook (2005) for a discussion of the above model assumptions.

Should an item under warranty at vendor v break down and experience a response time (time between its breakdown and the ensuing completion of its repair) of r, then a cost $c_v(r)$ is incurred. This may include repair costs (parts and labour) in addition to costs which assess the impact on the manufacturer of lost customer goodwill when repair times are long. In their static model, Opp et al. (2003) assume the linear form

$$c_v(r) = c_v + hr, \quad (3a)$$

while Ding and Glazebrook (2005) consider models for which

$$c_v(r) = c_v + hI(r > \tau), \quad (3b)$$

$$c_v(r) = c_v + h(r - \tau)^+, \quad (3c)$$

where $I(\cdot)$ is the indicator function. In (3a), a single time unit spent by a single item awaiting repair incurs a fixed goodwill cost of h. In (3b), a goodwill cost of h is incurred for those items whose response times exceed some service quality threshold τ, while in (3c), a cost of h is incurred for every unit of time by which the response time exceeds τ. In both (3b) and (3c), goodwill costs are only incurred if a manufacturer guarantee of service quality fails to be met.

The above costs are aggregated over all repairs (across all vendors) and averaged over time. The goal

of analysis is to develop policies for the allocation of incoming orders to vendors on the basis of the data in $\mathcal{N}$ which will minimize (or come close to minimizing) the resulting average cost rate incurred over an infinite horizon. Good allocation policies will take account of the work already committed to each vendor in relation to its service capacity. Vendors which become overloaded are likely to produce large response times and high costs. However, how the current vendor loads should be used to support allocation decisions is far from clear. Particular difficulties for an analysis of this model (hereafter referred to as Model 1) based on stochastic DP are as follows:

(a) The breakdown/repair processes which generate the costs are not observable. Costs for vendor v may only be inferred from its current state $\{N_v, (\mathbf{x}^v, \mathbf{t}^v)\}$.

(b) This state is itself complex, being both continuous and of high (and variable) dimension.

Despite these formidable difficulties, we will succeed in developing effective allocation procedures. These will be described in the upcoming sections and will be subject to numerical evaluation.

Before proceeding further, we offer an explanation of some terms. The descriptors "static" and "dynamic" are applied in the paper both to *models* of allocation problems and to *policies* for making allocations. A *static model* is one in which the warranty population is taken to be constant over time, while in a *dynamic model*, the population varies in size. A *static allocation policy* is one which takes no account of the system state information in $\mathcal{N}$ in making decisions, while a *dynamic policy* does take such account.

3. Heuristics Developed from Static Models

The *static model approach* to the design of allocation heuristics ignores the dynamic and stochastic nature of the warranty population. A static model (called Model 2) considers a fixed item population of size K and designs a collection $\mathbf{k} = (k_1, k_2, \ldots, k_V)$ of fixed allocations to vendors to minimize a resulting cost rate. The static model approach infers heuristics for Model 1 in §2 from the results of analyses of static Model 2.

We write $g_v(k_v)$ for the average cost rate incurred at vendor v when it has a fixed number k_v of items in its warranty population *for all time*. The allocation problem for Model 2 may be expressed as

$$\text{(P)} \quad \min \sum_{v=1}^{V} g_v(k_v) \equiv \bar{G}(K)$$
$$\text{s.t.} \sum_{v=1}^{V} k_v = K,$$
$$k_v \in \mathbb{N}, \quad 1 \le v \le V.$$

In outline, to develop the cost rate g_v, we consider D_v the (random) number of down items at the vendor v. For a fixed vendor population of size k_v, D_v evolves as a birth-death process with state space $\{d;\ 0 \le d \le k_v\}$ whose stationary distribution $\{\Pi_{vd}(k_v);\ 0 \le d \le k_v\}$ is straightforward to compute. See, for example, Opp et al. (2003) and Ding and Glazebrook (2005). See also Taylor and Karlin (1998) for a discussion of birth-death processes. Now write

$$\bar{c}_v(d) = E\{c_v(r) \mid D_v = d\}, \quad 0 \le d \le k_v - 1,\ 1 \le v \le V,$$

for the conditional expected cost incurred when an item breaks down at a time at which d other items are already queued for repair at vendor v. Because the repair times of individual items at vendor v are independent and have an exponential distribution with rate μ_v, it is a straightforward matter to compute $\bar{c}_v(\cdot)$ for the cost models (3a)–(3c). An appropriate cost rate for vendor v may now be developed as

$$g_v(k_v) = \sum_{d=0}^{k_v} \lambda(k_v - d)\Pi_{vd}(k_v)\bar{c}_v(d), \quad k_v \in \mathbb{N},\ 1 \le v \le V, \tag{4}$$

for use in the optimization problem (P).

Note that for a wide range of plausible cost models, we may expect the vendor-specific cost rates $g_v\colon \mathbb{N} \to \mathbb{R}^+$, $1 \le v \le V$, to be increasing convex in k_v or nearly so.

Example 1 (Opp et al. 2003). For the linear cost model in (3a), the associated vendor cost rate g_v will be increasing convex in k_v for all values of the model parameters, provided only that $h > \lambda c_v$.

Example 2 (Ding and Glazebrook 2005). When single repairer approximations (in which the service rate of the single repairer is taken to be $\mu_v\sigma_v$) are deployed for the cost models in (3b) and (3c), the resulting vendor cost rates g_v are increasing in k_v for all values of the model parameters. They are also close to convex in a sense which is made precise in Ding and Glazebrook (2005). Note that for realistic scenarios, single repairer approximations have been found to be adequate. They are, moreover, appropriate in situations where the manufacturer has knowledge of each vendor's effective service rate but not of its number of repairers.

The significance of the above is that when each of the g_vs is increasing convex in k_v, then the optimization problem (P) is solved by a greedy algorithm. This solution was first proposed by Gross (1956); see also Fox (1966). The greedy algorithm which solves (P) may be described as follows:

Greedy Algorithm for (P)

Step 0. Set $k_v = 0$, $1 \le v \le V$.

Step 1. Choose any $w \in \arg\min_{1 \le v \le V}\{[g_v(k_v + 1) - g_v(k_v)]\}$.

Step 2. Set $k_w = k_w + 1$.
Step 3. If $\sum_{v=1}^{V} k_v < K$, go to Step 1; otherwise stop.

The reader should note that if we introduce into (P) a capacity constraint (of the form $k_v \le B_v$) for each vendor, then we only require each g_v to be convex up to the capacity B_v for a greedy approach to provide an optimal solution. See Ding and Glazebrook (2005).

The above discussion of static Model 2 yields a natural allocation heuristic for dynamic Model 1 of §2 for cases in which the fixed allocation, vendor-specific cost rates g_v are increasing convex or nearly so. Suppose that an order of size x has arrived and is awaiting allocation and that the current system state is given by $\mathcal{N}$. For each vendor v, compute the average cost-rate escalation

$$g_v(N_v + x) - g_v(N_v) \tag{5}$$

experienced when a (fixed) warranty population at vendor v is increased from N_v to $N_v + x$. The reader should note from (4) that in the computations of $g_v(N_v)$ and $g_v(N_v + x)$, different stationary distributions $\{\Pi_{vd}(N_v), 0 \le d \le N_v\}$ and $\{\Pi_{vd}(N_v + x), 0 \le d \le N_v + x\}$ are used. The *dynamic greedy allocation heuristic* (denoted GRE) will assign the incoming order x to any vendor for which the quantity in (5) is minimal. Please note that GRE makes no use of the time remaining under warranty of the items at each vendor, only how many there are. As will become clear, numerical evidence suggests that this simple heuristic performs outstandingly well.

We now describe another way of using static Model 2 to develop dynamic allocation heuristics. Suppose that the mean size of the warranty population $\eta W\theta$ is an integer (and otherwise take the nearest integer to it). Now consider the static optimization problem (P) with $K = \eta W\theta$. Use

$$\mathbf{k}(\eta W\theta) = \{k_1(\eta W\theta), k_2(\eta W\theta), \ldots, k_V(\eta W\theta)\}$$

for an optimal set of vendor allocations. We propose a heuristic for Model 1 which dynamically tracks this static solution by allocating an incoming order to any vendor for which the difference

$$N_v - k_v(\eta W\theta) \tag{6}$$

is minimal. Note that this *tracking heuristic* (denoted TRA) takes no account of the order size x.

Comment. The models discussed in this paper suppose that the administrative overhead involved rules out observation of the repair queues at the vendors when allocation decisions are made. The work of Opp et al. (2005) considered static models with costs given by (3a) in which utilisation of information regarding the lengths of the repair queues reduced overall costs by between 0% and 18%, with figures of 1%–5% being typical. We would expect a similar degree of cost improvement to be available for our dynamic warranty populations.

4. An Approximate DP Approach to Heuristic Development

We shall now proceed to develop an allocation heuristic for Model 1, which makes full use of system state information, unlike the heuristics GRE and TRA developed in §3. To do so, we must overcome the two challenges posed by Model 1 of partial observability and system-state complexity, and expressed in (a) and (b) of §2. We shall use an *approximate DP approach*, which has proved effective in other application domains. See, for example, Glazebrook et al. (2004), Krishnan (1987), and Tijms (1994). This is a two-stage approach to policy development in which the problem's inaccessible value function is approximated by deployment of an assumption that all decisions beyond the current one are made according to a strongly performing static (state-independent) allocation policy. The first stage of the approach concerns the development of such a static policy. This policy indicates the proportion of workload overall which should be allocated to each vendor. An effective dynamic heuristic is then developed at the second stage by the application of a single DP policy improvement step. The resulting heuristic will be used both as a benchmark by which other simpler policies may be judged and as a policy of interest in its own right.

Stage 1: Initial Static Policy. We write $\mathbf{p} = (p_1, p_2, \ldots, p_V)$ for any static policy for Model 1 which independently allocates each incoming order to vendor v with probability p_v, where

$$p_v \ge 0, \quad 1 \le v \le V, \quad \text{and} \quad \sum_{v=1}^{V} p_v = 1.$$

At the first stage of our approximate DP approach, we choose $\mathbf{p}$ to minimize the system cost rate. Recall that costs in Model 1 depend on the unobserved response times of repaired items. Write $H(\mathbf{p})$ for the system cost rate incurred under $\mathbf{p}$, with $\hat{\mathbf{p}}$ such that

$$H(\hat{\mathbf{p}}) = \min_{\mathbf{p}} H(\mathbf{p}).$$

The partially observed nature of the system make the cost rates $H(\mathbf{p})$ difficult to compute. We overcome this by making use of a fully observed approximating model (called Model 3) which modifies Model 1 by assuming that costs are incurred at (observable) rate $\sum_{v=1}^{V} g_v(N_v)$ when the system is in state $\mathcal{N} = \{N_v, (\mathbf{x}^v, \mathbf{t}^v), 1 \le v \le V\}$. Here the g_v, $1 \le v \le V$, are vendor cost rates inferred from Equation (4) in §3. In words, in this approximation to Model 1, a cost rate is assumed for the system in state $\mathcal{N}$ which would be exact if the respective vendor populations were to remain fixed at their current size. We write $G(\mathbf{p})$ for

the system cost rate under static policy $\mathbf{p}$ for approximating Model 3.

While the cost rates $G(\mathbf{p})$ for approximating Model 3 are easily available, simulation evidence suggests that they slightly overestimate the true cost rates $H(\mathbf{p})$. The extent of this overestimate was 1.07% on average for the cases studied involving singleton orders and 6.07% on average for the cases in which order sizes were random. Therefore, to achieve greater precision in our approximate DP approach, we considered the introduction of a tuning parameter α into approximating Model 3 (which in all cases studied was set at a value between 0.985 and 1.00) such that the system cost rate in state $\mathcal{N}$ is adjusted to $\sum_{v=1}^{V} g_v(\alpha N_v)$. This reflects the fact that the true cost rate under any static policy is approximated with greater accuracy if the vendor populations are slightly reduced in approximating Model 3. Details of how α is chosen are given in the discussion of our numerical results. However, in no case did this α-adjustment in Model 3 result in a dynamic allocation heuristic whose cost performance was significantly changed thereby. We therefore conclude that the unadjusted version of approximating Model 3 (i.e., with $\alpha = 1$) is perfectly adequate for our approximate DP approach. Nonetheless, to enable a full discussion of the issues, we describe the development of our dynamic heuristic with the α-adjustment in place.

In steady state (namely, after time W has elapsed from the beginning of the process), under static policy $\mathbf{p}$ the number of orders allocated to vendor v whose warranties have yet to expire is $\bar{M}_v \sim \text{Poisson}(\eta W p_v)$, $1 \le v \le V$. Moreover these random variables are independent. Now write $S_n = X_1 + X_2 + \cdots + X_n$, $n \in \mathbb{N}$, where the X_i are i.i.d. with $X_i \sim F$, the order size distribution, $1 \le i \le n$. Further, write

$$\gamma_v(n) = E\{g_v(\alpha S_n)\}, \quad 1 \le v \le V,\ n \in \mathbb{N}, \tag{7}$$

where in (7), g_v is the vendor-specific cost rate function developed in (4). The quantity $\gamma_v(n)$ is the mean cost rate for vendor v while responsible for items from n orders under approximating Model 3 with tuning parameter α. The approximating average system cost rate incurred when policy $\mathbf{p}$ is applied is then given by

$$G(\mathbf{p}) = \sum_{v=1}^{V} E_{\eta W p_v}\{\gamma_v(\bar{M}_v)\}, \tag{8}$$

where the subscript $\eta W p_v$ in (8) denotes an expectation taken under the assumption that $\bar{M}_v$ has the Poisson distribution with this mean. An intermediate goal of analysis is the search for a static policy $\mathbf{p}^*$ such that

$$G(\mathbf{p}^*) = \min_{\mathbf{p}} G(\mathbf{p}). \tag{9}$$

Following discussion of the g_v, $1 \le v \le V$, in §3 it is important to point out (and straightforward to show) that if the g_v are all increasing convex, then so are the γ_v, $1 \le v \le V$. In this event, the optimization problem in (9) is convex and separable, and simple efficient algorithms exist for its solution. The quantity $G(\mathbf{p}^*)$ is an accessible upper bound on the average cost rate incurred when an optimal (dynamic) policy is applied to approximating Model 3.

Stage 2: DP Policy Improvement Step. Suppose now that at time zero an order of size x arrives and awaits allocation to a vendor when the system state is $\mathcal{N} = \{N_v, (\mathbf{x}^v, \mathbf{t}^v), 1 \le v \le V\}$. Under DP policy improvement (PI), a decision regarding this allocation is made on the assumption that all subsequent allocations are made according to static policy $\mathbf{p}^*$. This assumption enables us to make a suitable approximation to optimal (future) costs. More specifically, we shall use $C(v, \mathbf{p}^*, T \mid \mathcal{N}, x)$ to denote the expected cost incurred under approximating Model 3 over the horizon $T \ge W$ when the initial allocation is made to vendor v and all subsequent allocations are according to $\mathbf{p}^*$. Our PI heuristic will, in any state $\mathcal{N}$, choose to allocate an order of size x to any vendor $v(x, \mathcal{N})$ satisfying

$$C\{v(x, \mathcal{N}), \mathbf{p}^*, T \mid \mathcal{N}, x\} = \min_{1 \le v \le V} C(v, \mathbf{p}^*, T \mid \mathcal{N}, x), \quad T \ge W. \tag{10}$$

It will follow from the analysis below that there is indeed such a choice of vendor.

If the incoming order at time zero is allocated to vendor v, then, from (1), the vendor v state undergoes a transition $(\mathbf{x}^v, \mathbf{t}^v) \to (\mathbf{x}^v, \mathbf{t}^v)^x$, where

$$(\mathbf{x}^v, \mathbf{t}^v)^x \equiv \{(x_1^v, t_1^v), (x_2^v, t_2^v), \ldots, (x_{M_v}^v, t_{M_v}^v), (x, W)\}. \tag{11}$$

Subsequent evolution under policy $\mathbf{p}^*$ is independent for distinct vendors. Further, at any time $T > W$, the time zero allocation of the size x order does not impact the system state then. The latter has been evolving under allocation $\mathbf{p}^*$ for the previous W time units. Hence,

$$\begin{aligned} C(v, \mathbf{p}^*, T \mid \mathcal{N}, x) &= C_v\{p_v^*, W \mid (\mathbf{x}^v, \mathbf{t}^v)^x\} - C_v\{p_v^*, W \mid (\mathbf{x}^v, \mathbf{t}^v)\} \\ &\quad + \sum_{w=1}^{V} C_w\{p_w^*, W \mid (\mathbf{x}^w, \mathbf{t}^w)\} \\ &\quad + (T - W) \sum_{w=1}^{V} E_{\eta W p_w^*}\{\gamma_w(\bar{M}_w)\}, \\ &\qquad 1 \le v \le V,\ T \ge W. \end{aligned} \tag{12}$$

Expression (12) partitions the expected system cost during $[0, T]$ between costs incurred during $[0, W]$ and those incurred during $[W, T]$ (the final term

in (12)). Costs incurred during $[0, W]$ are partitioned between those incurred at the chosen vendor v under $\mathbf{p}^*$ and those at other vendors. The following result follows immediately from (10) and (12).

THEOREM 1 (CHARACTERIZATION OF THE PI HEURISTIC). *The allocation policy obtained when a single DP policy improvement step is applied to static policy* $\mathbf{p}^*$ *operates as follows*: *If an order of size* x *arrives when the system state is*

$$\mathcal{N} = \{N_v, (\mathbf{x}^v, \mathbf{t}^v), 1 \le v \le V\},$$

it should be allocated to any vendor v *which has a minimal value of the calibrating index*

$$I_{xv}\{(\mathbf{x}^v, \mathbf{t}^v)\} \equiv C_v\{p_v^*, W \mid (\mathbf{x}^v, \mathbf{t}^v)^x\} - C_v\{p_v^*, W \mid (\mathbf{x}^v, \mathbf{t}^v)\}. \quad (13)$$

It is a straightforward matter to compute the calibrating indices in (13). To develop appropriate formulae, we expand the notation in (7) such that for any $y \in \mathbb{N}$, $n \in \mathbb{N}$ and any $1 \le v \le V$, we have that

$$\gamma_v(y, n) = E[g_v\{\alpha(y + S_n)\}].$$

It will clarify matters if we now focus on indices for an individual vendor and drop the vendor identifier v. Hence, we write

$$I_x\{(\mathbf{x}, \mathbf{t})\} = C\{p^*, W \mid (\mathbf{x}, \mathbf{t})^x\} - C\{p^*, W \mid (\mathbf{x}, \mathbf{t})\}, \quad (14)$$

where

$$(\mathbf{x}, \mathbf{t}) \equiv \{(x_1, t_1), (x_2, t_2), \ldots, (x_M, t_M)\}. \quad (15)$$

The following result gives an expression for the index in (14). Note that we now adopt the notational conventions $t_0 = 0$, $t_{M+1} = W$.

LEMMA 1 (VENDOR-SPECIFIC CALIBRATING INDICES). *The vendor-specific indices which determine the PI allocation heuristic are given by*

$$I_x\{(\mathbf{x}, \mathbf{t})\} = \sum_{m=0}^{M} \int_{t_m}^{t_{m+1}} E_{\eta t p^*}\left\{\gamma\left(x + \sum_{m+1}^{M} x_r, \bar{M}\right) - \gamma\left(\sum_{m+1}^{M} x_r, \bar{M}\right)\right\} dt \quad (16)$$

for all values of the arguments concerned.

PROOF. Consider the computation of

$$C\{p^*, W \mid (\mathbf{x}, \mathbf{t})^x\}. \quad (17)$$

Choose some fixed time t such that $t_m < t \le t_{m+1}$, $0 \le m \le M$. If the state at time zero is $(\mathbf{x}, \mathbf{t})^x$, then by time t the items represented by the pairs (x_i, t_i), $1 \le i \le m$, will no longer be under warranty at the vendor. Of those items which had been under warranty at time zero, a total of $x + \sum_{m+1}^{M} x_r$ will remain so. These items will have been joined at t by a Poisson distributed number of orders (with mean $\eta t p^*$) which have been allocated to the vendor since time zero. Hence, in the computation of (17) for approximating Model 3, the expected instantaneous cost rate at t is given by

$$E_{\eta t p^*}\left[g\left\{\alpha\left(x + \sum_{m+1}^{M} x_r + S_{\bar{M}}\right)\right\}\right] \equiv E_{\eta t p^*}\left[\gamma\left(x + \sum_{m+1}^{M} x_r, \bar{M}\right)\right]. \quad (18)$$

If we now consider the computation of $C\{p^*, W \mid (\mathbf{x}, \mathbf{t})\}$, we can derive equivalent expected instantaneous cost rates by inserting $x = 0$ into the expressions in (18). The expression in (16) is then derived from (14) by integration of the appropriate instantaneous cost rates. This concludes the proof. □

The next result has three parts, each of which describes aspects of the behaviour of the calibrating indices introduced in Theorem 1 and Lemma 1. The first part will be utilized in comment 2 following the result which concerns problems where we drop the assumption that each incoming order is sent to a single vendor. The remaining two parts assert that the indices increase with vendor workload (in a suitably defined sense) for the important cases in which the vendor cost rate functions g are increasing convex. Hence, the heuristic described in the statement of Theorem 1 favours vendors with smaller relative workloads when making allocations, as would seem sensible. Theorem 2, Parts (ii) and (iii) confirm that the calibrating indices have the kind of properties that the decision maker might wish for. First, we need to introduce for the vendor state $(\mathbf{x}, \mathbf{t})$ the related function $(\mathbf{x}, \mathbf{t})(t)$: $[0, W] \to \mathbb{N}$, where $(\mathbf{x}, \mathbf{t})(t)$ is the number of items in the state $(\mathbf{x}, \mathbf{t})$ whose unexpired warranties exceed t. We have that

$$(\mathbf{x}, \mathbf{t})(t) = \sum_{m+1}^{M} x_r, \quad t_m < t \le t_{m+1},\ 0 \le m \le M.$$

We now say that the system state $(\mathbf{x}, \mathbf{t})$ *dominates* system state $(\mathbf{y}, \mathbf{s})$ if and only if

$$(\mathbf{x}, \mathbf{t})(t) \ge (\mathbf{y}, \mathbf{s})(t), \quad 0 \le t \le W. \quad (19)$$

The proof of Theorem 2 may be found in the online appendix to this paper (provided in the e-companion).[1]

THEOREM 2 (PROPERTIES OF THE VENDOR-SPECIFIC CALIBRATING INDICES). *If the vendor-specific cost rate* g *is increasing convex, then the calibrating index* $I_x(\mathbf{x}, \mathbf{t})$ *has the following properties*:

(i) $I_x(\mathbf{x}, \mathbf{t})$ *is increasing and convex in* x *for all* $(\mathbf{x}, \mathbf{t})$;

(ii) $I_x(\mathbf{x}, \mathbf{t})$ *is increasing componentwise in both* $\mathbf{x}$ *and* $\mathbf{t}$ *for all* x;

[1] An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

(iii) *If* $(\mathbf{x}, \mathbf{t})$ *dominates* $(\mathbf{y}, \mathbf{s})$, *then, for all* x,

$$I_x(\mathbf{x}, \mathbf{t}) \geq I_x(\mathbf{y}, \mathbf{s}).$$

Comments.

1. To give the reader more sense of the likely form of the key calibrating indices in (16), we consider a toy example in which all orders are of size 1 ($x_m^v = 1$ for all m, v), and all vendor-specific cost rates are quadratic such that

$$g_v(k) = a_v + b_v k + c_v k^2, \quad 1 \leq v \leq V,\ k \in \mathbb{N}, \tag{20}$$

where in (20) and what follows we restore the vendor identifier v. It is possible to show that the vendor-specific calibrating indices are then given by

$$\alpha^{-2} I_v(\mathbf{t}^v) = c_v \left(2 \sum_1^{M_v} t_m^v - \eta p_v^* W^2 \right) + const, \quad 1 \leq v \leq V, \tag{21}$$

where in (21) *const* denotes a constant common to all indices. The expression in (21) has a simple interpretation: $\sum_1^{M_v} t_m^v$ is the total of all unexpired warranties at vendor v, and $\eta p_v^* W^2/2$ is the mean value of this quantity under the static policy $\mathbf{p}^*$. Hence, an allocation policy based on the indices in (21) (see Theorem 1) will favour vendors whose current commitments (as measured by $\sum_1^{M_v} t_m^v$) are most below those indicated by the optimal static policy, the difference being factored by the (positive) cost constant c_v. In this sense, the PI heuristic can be said to *dynamically track* the optimal static solution. This could be thought of as a more refined version of the tracking heuristic TRA described at the end of §3.

2. A variety of elaborations of the material in this section are readily available, including to models in which warranty periods may differ between orders. We can also consider alternatives to the assumption that each order should be allocated to a single vendor. Suppose, for example, that we allow each order to be distributed between vendors in any fashion. In that event, the result of Stage 2 of the above approximate DP analysis is modified as follows: Should an order of size x arrive when the system state is $\mathcal{N} = \{N_v, (\mathbf{x}^v, \mathbf{t}^v), 1 \leq v \leq V\}$, a PI allocation heuristic would propose that the order be divided between the vendors such that vendor v receives y_v items, where $\mathbf{y} = \{y_1, y_2, \ldots, y_V\}$ is chosen to

$$\begin{aligned} &\text{minimize} \ \sum_{v=1}^{V} I_{y_v v}\{(\mathbf{x}^v, \mathbf{t}^v)\} \\ &\text{subject to} \ \sum_{v=1}^{V} y_v = x. \end{aligned} \tag{22}$$

Further, the result of Theorem 2(i) implies (via the result of Gross 1956) that in the important cases for which the vendor-specific cost rates g_v are all increasing convex, then the minimization in (22) is achieved by a greedy procedure. Note that we can also define versions of the heuristics GRE and TRA, developed at the end of §3, which are suitable for this version of the allocation problem. Early numerical results suggest that allowing an order to be placed with several vendors instead of just one results in a modest reduction in the cost rates (of 0.38% on average in the cases studied) incurred by good heuristics.

3. It is a straightforward matter to compute the vendor-specific calibrating indices online as the system evolves over time. To see this, note that standard integration results mean that the quantity in (16) may be re-expressed as

$$\begin{aligned} &I_x\{(\mathbf{x}, \mathbf{t})\} \\ &\quad = \sum_{m=0}^{M} \sum_{n=0}^{\infty} \left\{ \gamma\left(x + \sum_{m+1}^{M} x_r, n \right) - \gamma\left(\sum_{m+1}^{M} x_r, n \right) \right\} \\ &\qquad \times \left\{ \sum_{s=n+1}^{\infty} (\eta p^*)^{s-1} (t_{m+1}^s e^{-\eta p^* t_{m+1}} - t_m^s e^{-\eta p^* t_m})(s!)^{-1} \right\}. \end{aligned} \tag{23}$$

To compute this accurately and quickly, it is enough to create a library of values of the quantities $\gamma(y, n)$ for the range $0 \leq y \leq \eta W\theta + 3\sqrt{\eta W \theta_2}$, $0 \leq n \leq \eta W + 3\sqrt{\eta W}$, and import these values, as appropriate, into the expression in (23). In practice, the two infinite sums converge quickly and may be well approximated by taking the upper limit of summation to be $\eta W + 3\sqrt{\eta W}$.

4. It has already been pointed out that $G(\mathbf{p}^*)$ is an accessible upper bound on the average cost rate incurred when an optimal (dynamic) policy is applied to approximating Model 3. To obtain a useful approximative lower bound, use $\tilde{G}(K)$, $K \in \mathbb{R}^+$, for the optimal value of the optimization problem (P) without the integrality constraints, namely,

$$\begin{aligned} (\tilde{\mathrm{P}}) \quad \min \ & \sum_{v=1}^{V} g_v(k_v) \equiv \tilde{G}(K) \\ \text{s.t.} \ & \sum_{v=1}^{V} k_v = K, \\ & k_v \geq 0, \quad 1 \leq v \leq V. \end{aligned}$$

The proof of Lemma 2 may be found in the online appendix.

Lemma 2. *When the vendor-specific cost rates* g_v, $1 \leq v \leq V$, *are convex,* $\tilde{G}(\alpha\eta W\theta)$ *is a lower bound on the average cost rate achieved by any allocation policy for approximating Model* 3.

5. Simulation Study

The proposals for allocation procedures made in the previous two sections have been assessed by an extensive simulation study. Questions to which we seek answers include the following:

1. When does the static Model 2 of §3 provide an adequate basis for the development of good allocation policies? In particular, are the very simple heuristics derived from static Model 2 competitive with the PI heuristic, given that the latter makes much more extensive use of system state information?

2. What cost penalties might be incurred if the warranty repair work is divided equally between vendors in contexts where their service capacities are different?

3. How sensitive are our conclusions to model assumptions?

4. What do our analyses and numerical results tell us about the research questions posed in the introduction?

To shed light on such matters, we conducted simulation studies of an extensive range of systems, all having four vendors ($V = 4$) to conduct the warranty work. The means and standard deviations of the total population of items under warranty at any time for the cases reported here are given in Table 1. The results of part of this study are summarized in Tables 2 and 3. We adopt single-repairer approximations for vendor dynamics throughout. All items have a warranty period of two years ($W = 2$) and a breakdown rate of 1.2 per year ($\lambda = 1.2$). The repair cost model (3c) is used, with $c_v = 0$, $1 \le v \le V$, $h = 1$, and $\tau = 0.04$. Hence, goodwill costs grow linearly (at the rate of one unit per year) once the response time for a repair exceeds 10 working days.

Table 2 concerns problems in which items are purchased as singletons at rates of 50 per year (Table 2(a)) and 250 per year (Table 2(b)), respectively. Throughout Table 3, orders are placed (and require allocation to a vendor) at a rate of 25 per year. The order size X is such that $X - 1$ has a Poisson distribution with mean $\theta - 1$, where θ takes values 1 (Table 3(a)), 5 (Table 3(b)), 11 (Table 3(c)), 15 (Table 3(d)), and 19 (Table 3(e)). Note that the introduction of the random order size into the examples in Table 3 increases the variability of the population size N, with the standard deviation of N growing approximately in a linear fashion with its mean. Finally, for problem specification, each of the seven subtables within Tables 2 and 3 considers six vendor scenarios, in all of which the total of the vendor service rates are equal, but where the degree of inequality between the vendors increases from scenario 1 (where the service rates are all equal) through scenario 6 (where vendor 1 has a service rate which is eight times that for vendor 4). More specifically, in scenario 1, service rates have the form $\mu_{v1} = (62.5)H$, $1 \le v \le 4$, while in scenario $j > 1$, vendor v has service rate $\mu_{vj} = 250Hx_j^{v-1}(1 - x_j) \cdot (1 - x_j^4)^{-1}$, $1 \le v \le 4$, $2 \le j \le 6$, where $x_j = 1 - 0.1 \cdot (j - 1)$, $2 \le j \le 6$. The constant H is adjusted to give a reasonable service capacity for the problem concerned and takes the values 0.7 (Tables 2(a) and 3(a)), 1.5 (Table 3(b)), 2.5 (Table 2(b)), 3.0 (Table 3(c)), 3.8 (Table 3(d)), and 4.7 (Table 3(e)). Rows in the table correspond to a given scenario (i.e., choice of service rates). The column heads in the tables are as follows:

$\bar{G}(\eta W\theta)$: These columns contain the values of the optimization problem (P) evaluated for the relevant mean population size. Because the vendor-specific cost rates for our examples are close to convex, then by Lemma 2, these values are close to the lower bounds on achievable cost performance for approximating Model 3 when $\alpha = 1$.

Table 1 **Means and Standard Deviations of N, the Total Number of Items Under Warranty, for Examples in Subsequent Tables**

	2(a)	2(b)	3(a)	3(b)	3(c)	3(d)	3(e)
$E(N)$	100	500	100	300	600	800	1,000
$\sqrt{\text{var}(N)}$	10.00	22.36	15.81	45.28	88.03	116.40	144.74

Table 2 **Results of a Simulation Study of the Comparative Performance of Five Allocation Heuristics When Orders Are Singletons**

Scenario	$\bar{G}(100)$	$G(\mathbf{p}^*)$	PIH	APIH	GRE	TRA	SMA
				(a)			
1	2.976	3.624	3.153	3.153	3.159	3.159	3.153
			(0.007)	(0.007)	(0.006)	(0.006)	(0.007)
2	2.957	3.599	3.128	3.134	3.136	3.138	3.512
			(0.007)	(0.007)	(0.007)	(0.008)	(0.008)
3	2.886	3.513	3.065	3.054	3.063	3.090	4.960
			(0.007)	(0.007)	(0.007)	(0.006)	(0.013)
4	2.745	3.348	2.914	2.922	2.919	2.926	8.671
			(0.006)	(0.007)	(0.006)	(0.007)	(0.021)
5	2.517	3.081	2.697	2.689	2.691	2.720	16.480
			(0.006)	(0.006)	(0.006)	(0.006)	(0.033)
6	2.183	2.690	2.357	2.361	2.360	2.410	46.743
			(0.006)	(0.006)	(0.006)	(0.006)	(1.041)
Scenario	$\bar{G}(500)$	$G(\mathbf{p}^*)$	PIH	APIH	GRE	TRA	SMA
				(b)			
1	11.224	15.109	12.291	12.291	12.303	12.303	12.291
			(0.033)	(0.033)	(0.034)	(0.034)	(0.033)
2	11.186	15.057	12.302	12.255	12.267	12.283	20.262
			(0.030)	(0.031)	(0.032)	(0.030)	(0.049)
3	11.042	14.879	12.115	12.119	12.125	12.138	45.997
			(0.030)	(0.033)	(0.031)	(0.031)	(0.070)
4	10.765	14.535	11.857	11.861	11.847	11.862	86.468
			(0.030)	(0.032)	(0.030)	(0.032)	(0.093)
5	10.315	13.973	11.405	11.397	11.403	11.427	140.494
			(0.031)	(0.033)	(0.033)	(0.032)	(0.125)
6	9.642	13.135	10.736	10.717	10.700	10.813	∞
			(0.028)	(0.029)	(0.033)	(0.031)	

Note. See the text for further details.

Table 3 Results of a Simulation Study of the Comparative Performance of Five Allocation Heuristics When Order Sizes Are Random

Scenario	$\bar{G}(100)$	$G(\mathbf{p}^*)$	PIH	APIH	GRE	TRA	SMA
				(a)			
1	2.976	4.860	3.390	3.381	3.400	3.400	3.381
			(0.010)	(0.010)	(0.010)	(0.010)	(0.009)
2	2.957	4.829	3.363	3.363	3.379	3.386	3.781
			(0.010)	(0.010)	(0.010)	(0.010)	(0.011)
3	2.886	4.720	3.290	3.288	3.298	3.316	5.321
			(0.010)	(0.010)	(0.011)	(0.010)	(0.016)
4	2.744	4.510	3.150	3.157	3.162	3.200	9.147
			(0.009)	(0.010)	(0.009)	(0.010)	(0.026)
5	2.517	4.171	2.926	2.927	2.940	3.038	16.949
			(0.009)	(0.009)	(0.009)	(0.010)	(0.040)
6	2.183	3.672	2.617	2.607	2.619	2.801	51.456
			(0.009)	(0.009)	(0.010)	(0.010)	(1.150)
Scenario	$\bar{G}(300)$	$G(\mathbf{p}^*)$	PIH	APIH	GRE	TRA	SMA
				(b)			
1	11.548	30.412	16.709	16.719	16.775	16.775	16.698
			(0.077)	(0.071)	(0.075)	(0.075)	(0.075)
2	11.512	30.333	16.670	16.701	16.755	16.761	20.641
			(0.070)	(0.072)	(0.073)	(0.074)	(0.080)
3	11.379	30.601	16.599	16.576	16.671	16.789	33.564
			(0.075)	(0.073)	(0.073)	(0.073)	(0.105)
4	11.116	29.537	16.428	16.371	16.490	16.763	56.276
			(0.076)	(0.078)	(0.077)	(0.075)	(0.130)
5	10.694	28.682	16.117	16.060	16.180	16.848	93.139
			(0.076)	(0.070)	(0.074)	(0.077)	(0.280)
6	10.066	27.405	15.687	15.579	15.709	17.238	∞
			(0.076)	(0.076)	(0.078)	(0.079)	
Scenario	$\bar{G}(600)$	$G(\mathbf{p}^*)$	PIH	APIH	GRE	TRA	SMA
				(c)			
1	10.793	52.522	22.113	22.098	22.227	22.227	22.095
			(0.137)	(0.144)	(0.137)	(0.137)	(0.141)
2	10.753	52.386	22.113	22.061	22.221	22.263	31.349
			(0.142)	(0.143)	(0.143)	(0.141)	(0.156)
3	10.611	51.918	22.069	22.006	22.169	22.447	60.359
			(0.139)	(0.139)	(0.139)	(0.143)	(0.199)
4	10.336	51.014	21.882	21.779	22.045	22.533	107.938
			(0.135)	(0.142)	(0.138)	(0.139)	(0.259)
5	9.887	49.544	21.734	21.589	21.838	23.072	178.670
			(0.132)	(0.136)	(0.138)	(0.140)	(0.482)
6	9.214	47.350	21.413	21.092	21.527	24.124	∞
			(0.135)	(0.133)	(0.129)	(0.151)	
Scenario	$\bar{G}(800)$	$G(\mathbf{p}^*)$	PIH	APIH	GRE	TRA	SMA
				(d)			
1	19.056	84.432	39.928	39.882	40.096	40.102	39.953
			(0.236)	(0.239)	(0.240)	(0.241)	(0.234)
2	19.007	84.258	39.941	38.890	40.061	40.098	54.693
			(0.235)	(0.233)	(0.236)	(0.239)	(0.258)
3	18.830	83.660	39.864	39.997	40.115	40.403	93.887
			(0.242)	(0.244)	(0.242)	(0.243)	(0.303)
4	18.483	82.504	39.836	39.997	40.012	40.829	163.249
			(0.241)	(0.244)	(0.239)	(0.241)	(0.356)
5	17.916	80.626	39.762	40.151	39.999	41.632	289.603
			(0.241)	(0.242)	(0.248)	(0.248)	(1.959)
6	17.060	77.824	39.692	40.425	39.885	43.477	∞
			(0.246)	(0.249)	(0.247)	(0.260)	

Table 3 (Continued)

Scenario	$\bar{G}(1{,}000)$	$G(\mathbf{p}^*)$	PIH	APIH	GRE	TRA	SMA
				(e)			
1	22.406	108.173	51.082	51.058	51.358	51.327	51.113
			(0.327)	(0.325)	(0.325)	(0.329)	(0.321)
2	22.356	107.961	50.970	51.061	51.263	51.369	70.364
			(0.323)	(0.318)	(0.318)	(0.323)	(0.336)
3	22.166	107.233	51.028	50.900	51.221	51.623	124.578
			(0.319)	(0.317)	(0.320)	(0.318)	(0.371)
4	21.798	105.826	51.127	50.908	51.363	52.264	208.585
			(0.317)	(0.325)	(0.325)	(0.328)	(0.430)
5	21.194	103.541	51.183	50.865	51.406	53.524	393.394
			(0.321)	(0.325)	(0.321)	(0.332)	(4.113)
6	20.278	100.135	51.335	50.837	51.574	55.338	∞
			(0.324)	(0.317)	(0.319)	(0.338)	

Note. See the text for further details.

$G(\mathbf{p}^*)$: These columns contain the values of the average cost rates incurred when an optimal static policy is applied to approximating Model 3 with $\alpha = 1$.

PIH: This is the PI heuristic of Theorem 1/Lemma 1 developed for the case $\alpha = 1$. These columns contain simulation-based estimates of the average cost rates incurred when PIH is applied to Model 1.

APIH: This is a PI heuristic as in Theorem 1/Lemma 1, but now developed for an α-value chosen to yield greater precision in our approximate DP methodology. A single α-value applies to each of the seven subtables within Tables 2 and 3. To obtain the chosen α, for each subtable a simulation-based estimate was obtained of the overall overestimate of true costs, which resulted when an optimal static policy was applied in all scenarios of approximating Model 3 without adjustment ($\alpha = 1$). The α-value was then chosen to secure the appropriate percentage cost reduction for the simple scenario 1 in which all optimal static policies are such that $p_v^* = 0.25,\ 1 \le v \le 4$. Details of the αs chosen and the degree of agreement post adjustment between approximating (Model 3) and true (Model 1) costs may be found in Table 4. The columns headed APIH in Tables 2 and 3 contain simulation-based estimates of the average cost rates incurred when this adjusted PI heuristic is applied to Model 1.

GRE, TRA, SMA: These columns report on the cost performances of, respectively, the dynamic greedy allocation heuristic (GRE) described at the conclusion of §3 near (5), the tracking heuristic (TRA) described near (6), and a heuristic (SMA), which allocates each incoming order to any vendor whose current committed workload (i.e., total of all unexpired warranty times) is smallest. The policy SMA is included in the study to give an indication of how a simple, indeed naive, allocation rule that treats all vendors equally performs. Each column contains the appropriate simulation-based estimates of the average cost

Table 4 The α Values Use for the APIH Heuristics of Tables 2 and 3 Together with the Resulting Optimal Static Cost Rate for Approximating Model 3 and Model 1

Scenario		2(a) $\alpha = 0.999$	2(b) $\alpha = 0.998$	3(a) $\alpha = 0.991$	3(b) $\alpha = 0.990$	3(c) $\alpha = 0.985$	3(d) $\alpha = 0.988$	3(e) $\alpha = 0.989$
1	Model 3	3.616	14.827	4.693	28.954	48.321	79.343	101.904
	Model 1	3.617 (0.009)	14.842 (0.041)	4.692 (0.017)	29.026 (0.096)	48.379 (0.209)	79.317 (0.311)	101.832 (0.398)
2	Model 3	3.593	14.785	4.661	28.872	48.187	79.202	101.688
	Model 1	3.598 (0.008)	14.784 (0.038)	4.660 (0.016)	28.917 (0.100)	48.268 (0.201)	79.293 (0.301)	101.671 (0.386)
3	Model 3	3.504	14.608	4.549	28.593	47.691	78.553	100.869
	Model 1	3.504 (0.009)	14.579 (0.036)	4.555 (0.015)	28.611 (0.094)	47.774 (0.202)	78.641 (0.312)	101.001 (0.394)
4	Model 3	3.339	14.250	4.339	28.035	46.718	77.252	99.272
	Model 1	3.338 (0.008)	14.286 (0.034)	4.357 (0.015)	28.035 (0.104)	46.847 (0.200)	77.463 (0.307)	99.361 (0.379)
5	Model 3	3.073	13.686	4.008	27.179	45.226	75.354	97.041
	Model 1	3.066 (0.008)	13.691 (0.034)	4.022 (0.016)	27.263 (0.099)	45.391 (0.211)	75.620 (0.312)	97.353 (0.389)
6	Model 3	2.682	12.856	3.520	25.956	43.192	72.686	93.819
	Model 1	2.689 (0.008)	12.880 (0.039)	3.531 (0.015)	25.979 (0.105)	43.357 (0.207)	73.143 (0.326)	94.244 (0.394)

rates incurred when the heuristic concerned is applied to Model 1.

In the tables, the standard errors of the corresponding cost-rate estimates appear in parentheses.

The following comments relate to questions 1–4 posed at the start of this section.

Question 1. A major feature of the results in Tables 2 and 3 is the strong performance of the heuristics GRE and TRA based on solutions to static Model 2. For the singleton order problems of Table 2, the performance of GRE is almost indistinguishable from that of the PI heuristics. In Table 3, where orders are of random size, the PI heuristics tend to outperform the others but the margins are often small and in most cases fail to be statistically significant. In the worst case, for GRE its average cost rate exceeds that of APIH by 2.06%. We conclude that in most environments in which the warranty population is subject to moderate temporal variability, the dynamic greedy heuristic will perform well. The equivalent worst-case performance for TRA is 14.38%. With regard to the latter allocation procedure, there is some evidence of deteriorating performance as the differences between vendor service rates increase, especially so for the bulk order cases of Table 3. It appears that TRA's failure to take account of order size in making allocations may lead it on occasion to overload vendors with small service capacity. Finally, we note that there is no real evidence in the tables that the introduction of the tuning parameter and the corresponding modification of PIH to APIH has made any significant impact on the resulting heuristic's cost performance. Relative costs range from a 1.85% advantage in favour of PIH to a 1.52% advantage in favour of APIH. In most cases, the difference in cost rate between the two is very small and of no practical significance. We conclude that the simple version of approximating Model 3 (with $\alpha = 1$) is perfectly adequate for our approximate DP approach. For clarity, future comments concerning the PI approach will focus exclusively on the heuristic PIH.

An interesting inference from the above strong performance of GRE is that in the current models, once the number of items at each vendor is known, further information regarding unexpired warranty times adds relatively little to effective decision making. To understand why, see Table 5, which records the means and standard deviations of the times between successive allocations of newly purchased items to vendor 1 for the six scenarios of Table 2(a) under the heuristics PIH, GRE, and the optimal static policy $\mathbf{p}^*$ (computed for $\alpha = 1$). Note that, while the average rates at which the three heuristics send items to the vendor are equal (within sampling error), the dynamic heuristics impose greater regularity on the allocations as reflected in the smaller standard deviations for the interallocation times. We conclude that, if past allocations have been made effectively, it will be rare to encounter a situation in which, from the perspective of making an optimal allocation decision, one vendor dominates another with regard to numbers of items (N_v) but is dominated with regard to (any reasonable measure of) the unexpired warranties ($\mathbf{t}^v$). Note that all of this relates to the Poisson assumption (with uniform rate) concerning arriving orders. Should the arrival process be, for example, nonhomogeneous Poisson with substantial fluctuations in the arrival rate, then we would expect the patterns of unexpired warranties to be necessarily much more irregular and potentially more informative for (good) allocation decisions. Further consideration of this issue will be the subject of a later paper.

Question 2. One way of understanding Tables 2 and 3 is as follows: Our simulations show that the overall proportions of items allocated to the vendors under static policy $\mathbf{p}^*$ agree well with those allocated under the heuristics PIH, APIH, GRE, and TRA.

Table 5 Estimated Means and Standard Deviations of the Interallocation Times at Vendor 1 Generated by Three Heuristics for the Scenarios in Table 2(a)

	Sample mean	Sample standard deviation	Sample size
Scenario 1			
PIH	0.0786	0.0399	1,017
GRE	0.0774	0.0668	1,030
$\mathbf{p}^*$	0.0826	0.0808	967
Scenario 2			
PIH	0.0678	0.0372	1,180
GRE	0.0675	0.0597	1,184
$\mathbf{p}^*$	0.0641	0.0638	1,247
Scenario 3			
PIH	0.0543	0.0307	1,474
GRE	0.0541	0.0517	1,477
$\mathbf{p}^*$	0.0558	0.0572	1,433
Scenario 4			
PIH	0.0446	0.0264	1,794
GRE	0.0445	0.0424	1,978
$\mathbf{p}^*$	0.0454	0.0454	1,760
Scenario 5			
PIH	0.0391	0.0267	2,044
GRE	0.0389	0.0348	2,054
$\mathbf{p}^*$	0.0398	0.0402	2,010
Scenario 6			
PIH	0.0322	0.0233	2,485
GRE	0.0320	0.0306	2,501
$\mathbf{p}^*$	0.0313	0.0306	2,557

See Table 5 and the above comments. The dynamic heuristics improve the cost performance of the static policy $\mathbf{p}^*$ by offering these proportionate workloads to the vendors in a manner that reduces variability, and hence, lessens the chance of excessive queue lengths for repair. It is clear from the tables that this dynamic workload management can be important in cost terms. In most of Tables 3(c)–3(e), for example, the cost rates associated with these dynamic heuristics are little more than $0.5G(\mathbf{p}^*)$. A point to note is that Tables 2 and 3 are broadly consistent in indicating that a good dynamic allocation heuristic can achieve a cost rate, which moves from $G(\mathbf{p}^*)$ around 70% of the way toward $\bar{G}(\eta W\theta)$. Hence, these simply calculated values give a good prior indication of the cost savings achievable from dynamic workload management.

The heuristic SMA departs from this general pattern in being committed to splitting the workload evenly between the vendors, notwithstanding any differences in the vendor service rates. The cost consequences of this are evident from the tables. While in Tables 2 and 3 it is the results for scenario 6 that are most dramatic, it may be that the results for scenario 2 carry more weight. For the latter problems, the between-vendor differences are fairly small (with the smallest vendor service rate being 73% of the largest), and yet the cost rate for SMA can exceed that from PIH by nearly 65% (see Table 2(b)) in individual cases. Note that in some cases of scenario 6, the amount of congestion experienced by the least capable vendor under SMA was so severe that we could not obtain good cost rate estimates in simulation runs of reasonable length. In these cases, the cost rate has been entered as ∞.

Question 3. We conducted a number of simulation experiments to explore how the performance of the allocation heuristics stood up to a range of modest departures from model assumptions. In Table 6, we find that simulation-based estimates of average cost rates incurred under PIH and GRE for the following cases:

(I) PIH and GRE are derived for the model considered in Table 2(b), but are applied in an environment in which new orders arrive as a nonhomogeneous Poisson process with arrival rate 240 for the first three quarters of each year and 280 for the final quarter.

(II) As in case I, but the annual arrival rate is now 240 for quarters 1 and 3 and 260 for quarters 2 and 4 of each year.

Table 6 An Exploration of the Sensitivity of the Performance of PIH and GRE to Departures from Model Assumptions

Case	(I)		(II)		(III)		(IV)	
Scenario	PIH	GRE	PIH	GRE	PIH	GRE	PIH	GRE
1	12.283 (0.033)	12.283 (0.033)	12.304 (0.033)	12.303 (0.030)	3.341 (0.008)	3.349 (0.009)	3.152 (0.006)	3.153 (0.007)
2	12.275 (0.031)	12.254 (0.032)	12.268 (0.034)	12.262 (0.032)	3.320 (0.009)	3.329 (0.008)	3.129 (0.007)	3.136 (0.006)
3	12.112 (0.031)	12.134 (0.031)	12.096 (0.031)	12.131 (0.033)	3.244 (0.009)	3.255 (0.009)	3.054 (0.006)	3.060 (0.006)
4	11.846 (0.031)	11.834 (0.031)	11.858 (0.033)	11.850 (0.033)	3.113 (0.009)	3.119 (0.009)	2.918 (0.006)	2.914 (0.006)
5	11.418 (0.033)	11.422 (0.032)	11.406 (0.027)	11.424 (0.034)	2.900 (0.008)	2.897 (0.008)	2.689 (0.005)	2.695 (0.006)
6	10.732 (0.032)	10.716 (0.031)	10.740 (0.032)	10.713 (0.030)	2.575 (0.008)	2.560 (0.009)	2.355 (0.006)	2.353 (0.006)

(III) PIH and GRE are derived for the model considered in Table 3(a), but are applied in an environment in which the order size X is such that $X-1$ has a Binomial$(2, 0.5)$ distribution. Hence, the order size has the same mean, but a variance which is half that for the assumed Poisson model.

(IV) PIH and GRE are derived for the model considered in Table 2(a), but are applied in an environment in which the up-times for items are drawn independently from a gamma distribution with shape parameter 2 and scale parameter 2.4. Hence, the up-times have the same mean, but a variance which is half that of the assumed exponential.

From Table 6, in cases (I) and (II) the cost rates and hence the relative performance of the two heuristics are only slightly changed by the modifications to the arrivals process. In case (III), reducing the variance of the order size distribution has led to reduced cost rates, with the relative performance of GRE marginally strengthened. Recall that neither the process of arriving orders nor the order size distribution plays any role in the development of greedy heuristic GRE.

From (IV), we see that halving the variance of the item up-times while keeping the mean constant has yielded a reduction in cost rates, while leaving the relative performance of the dynamic heuristics largely unchanged. See Ding and Glazebrook (2005) for comments on the likely impact of other departures from the Markovian assumptions made in relation to the breakdown/repair process at the vendors. We believe these will continue to hold good for our dynamic warranty population models.

Question 4. In the introduction, we posed three research questions related to a manufacturer's decision making in managing the warranty repair workload in our multiple vendor context. The second question concerns the distribution of the repair work among the vendors and has been the main theme of the paper. The first research question raised the issue of how one might determine whether a given level of service capacity among the vendors was adequate to deliver a post-sales repair service at an acceptable cost. The results in Tables 2 and 3 suggest that such questions are relatively straightforward in the sense that in all cases studied, the average cost rates incurred by good allocation heuristics (PIH, GRE, TRA) seem insensitive to how any total service capacity is divided between the vendors (i.e., the choice of scenario). It then follows that a good idea of the cost rate achievable from any total service capacity may be obtained from a simulation of the performance of any of our dynamic heuristics in the simplest case in which all vendor service rates are equal (scenario 1). From Tables 2 and 3, it would appear that cost-rate estimates derived for scenario 1 may modestly exceed those for other scenarios in most cases. Plainly, if the indicated achievable cost rate is not acceptable, then an increase in total service capacity is required. Our simulation evidence suggests that, should an increase in the total service rate be needed, it does not matter very much (in terms of cost performance) whether this is achieved via an increase in the committed service capacities of existing vendors or by contracting with additional vendors.

The third research question raises the issue of how much the manufacturer might be losing by maintaining an existing suboptimal approach to workload distribution. The comments in response to Question 2 above make it clear that serious losses can result over time if the overall proportions of the warranty population allocated to the respective vendors are inappropriate. Happily, guidance on what these proportions should ideally be is available from the solution to simple static Model 2. Once appropriate proportions are established, an indication of whether there is potential for further substantial reduction in costs from the adoption of a dynamic approach to workload management may be obtained by comparing appropriate values of $G(\mathbf{p}^*)$ and $\bar{G}(\eta W\theta)$.

6. Conclusions

We have developed heuristic policies for the allocation of newly purchased items to one of a collection of external vendors contracted to an equipment manufacturer to conduct repairs under warranty. The goal of such policies is the minimization of a cost rate in which goodwill penalties incurred for large response times typically predominate. We have seen that a good idea of what overall proportions should be allocated to each vendor is available from simple static models. However, it is often the case that considerable further savings are available from the adoption of a dynamic approach to workload management that seeks to reduce variability in the vendor subpopulations, and thereby lessen the chance of excessive repair queue lengths causing unacceptable response times for customers. The stochastic dynamic optimization problems involved in determining effective approaches are nonstandard, being only partially observed and having a system state of high and varying dimension. We utilize an approximate DP approach based on policy iteration to develop benchmark policies. It emerged from simulation studies that a simple dynamic greedy heuristic performs outstandingly well. An indication of the scope of the cost savings achievable by strongly performing dynamic policies is available from the relative values of quantities that are easy to calculate (denoted $\bar{G}(\eta W\theta)$ and $G(\mathbf{p}^*)$ in the paper). The conclusions of our study are robust to a range of modest departures from model assumptions.

The authors are grateful to Kulkarni (2006) for a wider discussion of the benefits of this work. He pointed out that the diminished variability resulting from a dynamic approach to workload management may have the ancillary benefit to an equipment manufacturer of making it easier for her to keep to a vendor contract that (say) specifies a minimum number of items for repair over the contract period.

7. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

Acknowledgments

The first author gratefully acknowledges support from Edinburgh University and from an Overseas Research Student Award. The work of the second and third authors was supported in part by the Engineering and Physical Science Research Council through the award of Grant GR/S45788/01. All authors are grateful to the associate editor and the referees for comments, which have acted as a stimulus to additional work and has considerably strengthened this paper.

References

Briskman, S. 2005. It's all about right-sourcing. http://services.silicon.com.

Buckowski, P., M. E. Hartmann, V. G. Kulkarni. 2005. Outsourcing prioritized warranty repairs. Technical report, University of North Carolina, Chapel Hill, NC.

Business Outsourcing Corporation. 2006. Case study 2. http://www.businessoutsourcing.com.

Ding, L., K. D. Glazebrook. 2005. A static allocation model for the outsourcing of warranty repairs. *J. Oper. Res. Soc.* **56** 825–835.

Fox, B. L. 1966. Discrete optimization via marginal analysis. *Management Sci.* **13** 210–216.

Glazebrook, K. D., P. S. Ansell, R. T. Dunn, R. R. Lumley. 2004. On the optimal allocation of service to impatient tasks. *J. Appl. Probab.* **41** 51–72.

Gross, O. 1956. A class of discrete type minimization problems. Technical Report RM-1644, RAND Corp., Santa Monica, CA.

Hordijk, A., G. Koole. 1990. On the optimality of the generalized shortest queue policy. *Probab. Engrg. Inform. Sci.* **4** 477–487.

Ibaraki, T., N. Katoh. 1988. *Resource Allocation Problems: Algorithmic Approaches*. MIT Press, Cambridge, MA.

Krishnan, K. R. 1987. Joining the right queue: A Markov decision rule. *Proc. 26th IEEE Conf. Decision and Control, Los Angeles, CA,* 1863–1868.

Kulkarni, V. G. 2006. Personal communication.

McDougall, P. 2005. Division of labour. http://www.informationweek.com.

Opp, M., K. D. Glazebrook, V. G. Kulkarni. 2005. Outsourcing warranty repairs: Dynamic allocation. *Naval Res. Logist.* **52** 381–398.

Opp, M., I. Adan, V. G. Kulkarni, J. M. Swaminathan. 2003. Outsourcing warranty repairs: Static allocation. Technical report, University of North Carolina, Chapel Hill, NC.

Serant, C. 2001. Solectron to provide Xbox support. *Electronic Buyers News*. (October 18).

Taylor, H. M., S. Karlin. 1998. *An Introduction to Stochastic Modelling*, 3rd ed. Academic Press, San Diego, CA.

Tijms, H. C. 1994. *Stochastic Models: An Algorithmic Approach*. Wiley, Chichester, UK.

Violino, B. 2006. What can logistics do for you? *Global Services.* (June 1).

Weber, R. R. 1978. On the optimal assignment of customers to parallel queues. *J. Appl. Probab.* **15** 406–413.

Winston, W. 1977. Optimality of the shortest line discipline. *J. Appl. Probab.* **14** 181–189.

ZDNet News. 2005. Study: Outsourcing falls from favor. (April 22), http://www.zdnet.com.