MANAGEMENT SCIENCE
Vol. 30, No. 8, August 1984
Printed in U.S.A.

THE FEASIBILITY OF ONE-OFFICER PATROL IN NEW YORK CITY*

LINDA GREEN AND PETER KOLESAR
Graduate School of Business, Columbia University, New York, New York 10027

How many patrol cars staffed with a single police officer are needed to provide equivalent police service to an existing system with n two-officer patrol cars? This question is explored for New York City using a multiple patrol car per call priority queueing model. It is shown that a one-officer patrol program is feasible, yet pitfalls exist which could adversely affect its performance. The paper details the process of data analysis and model building and emphasizes the subjective elements that remain in a highly technical OR study. Speed of response to emergency calls from the public was the key performance characteristic considered. The analysis also raised issues related to the safety of police officers in one-officer cars.
(GOVERNMENTAL SERVICES, POLICE; QUEUES—MULTI-CHANNEL, PRIORITY, APPLICATIONS; URBAN POLICY ANALYSIS)

1. Introduction

In June 1981, the authors were retained by the New York City Office of Management and Budget (OMB) to study the feasibility of changing from two to one police officer per patrol car. One-officer patrol cars were being explored as part of a productivity program in New York's municipal services. The City's Sanitation Department had already replaced many old three-man "back-loader" garbage trucks with new two-man "side-loaders." An agreement between the City and the Uniformed Sanitationman's Association allocated part of the savings realized by the two-man trucks as bonuses to the sanitation workers. The purpose of our study was to help determine whether the City should pursue a similar gainsharing program in upcoming negotiations with the Patrolmen's Benevolent Association (PBA), the union representing New York City police officers.

Feasibility of one-officer patrol cars involves many difficult issues (Kaplan 1979). While it was clear to all that the City would need to field more one-officer cars than two-officer cars in order to maintain adequate protection, it was unknown how many more one-officer cars would be needed to assure the safety of the public and of the police officers themselves. Our explicit assignment from OMB was to answer the question: "How many one-officer cars would the City have to field to achieve the same average dispatch delay to emergency calls as with the current two-officer system?"

1.1. *Setting*

New York City is divided into 73 geographical police patrol commands called precincts. Patrol cars, called radio motor patrols (RMPs), are fielded in each precinct for three 8-hour tours of duty: midnight–8 AM, 8 AM–4 PM, and 4 PM–midnight. RMPs have the primary responsibility of responding to and handling emergencies and crimes reported to the police. They are the main police presence on the City streets and they accounted for about half of the New York City Police Department's (NYCPD) annual budget of $800 million in 1981.

* Accepted by Warren E. Walker; received November 1, 1982. This paper was with the authors 3 months for 2 revisions.

0025-1909/84/3008/0964$01.25

Most calls come to the police via the 911 emergency telephone system. There were about 6.6 million such calls in 1981 resulting in about 2.9 million "radio runs" by patrol cars. At the time of our study, with two officers in a car, approximately 30% of all calls for service in New York received a multiple car response, and at a considerable number of these calls, several cars would be busy for an extended period. Thus, if the City changed to one-officer per car, more cars would be needed at some incidents in order to provide adequate police manpower.

The politics of one-officer patrol are complex. Three major parties emerged during our study: the Mayor's Office (represented largely by the Office of Management and Budget), the Police Department management, and the Patrolman's Benevolent Association leadership. OMB's intention was to try to negotiate an agreement with the PBA and begin partial implementation of a one-officer program in the fall of 1981 if the economic advantages seemed significant. The PBA seemed interested in a one-officer program for the gainsharing dollars they could claim for the rank and file. The Police Department management appeared cautious if not actually opposed to the idea. Our relationships with the various parties were complex for, although we had been conducting research for the Police Commissioner on related problems, we were brought into the one- vs. two-officer controversy by OMB and were, in effect, imposed on the Department by them. We had no contact with the union.

1.2. *Methodology*

The patrol system is a complex stochastic service system. Although, it is useful to conceptualize the system as a queueing model—the main approach in this study—one must recognize that no model can capture all the dominant aspects of the problem. Even the most complex simulation of police patrol can model only a portion of the system (Kolesar and Walker 1974, Larson 1972).

The most commonly used measure of system performance in quantitative analysis of police patrol is speed of response to emergency calls. (See Chaiken and Larson 1972, Kolesar 1982, Larson 1972 for background on analysis of police patrol.) Specifically, the NYCPD has for some years used "dispatch queueing delay"—the interval between the arrival of a call to the 911 emergency telephone system and the time at which the first patrol car is dispatched—as the primary measure of performance. In a one-officer system, the time until the *first* car is dispatched is no longer the only dominant factor —the *level* of response is an equally important consideration. To insure police officer safety, an adequate number of cars must respond quickly to potentially dangerous incidents. Police officer safety was a major issue in the negotiations between the City and the PBA (see *New York Times* October 22, 1981) and was a reason that NYCPD management was cautious about implementing a one-officer program.

Our primary tool of analysis was an extension of a queueing model that had already been developed by Green (1980). In the model, customers (calls for police service) require simultaneous service from a random number of servers (patrol cars), so the model seemed suited to the fundamental objective of the project—analyzing the effect of differing numbers of servers per emergency under one- and two-officer patrol operations. The existing model did not include a priority structure, but some work had been started on adding priorities to the model. The version of the "multiple car dispatch" model that was used in our analysis (which includes priorities and other enhancements) is reported in Green (1984).

Deadlines were tight and our analysis was performed in four weeks during June and July 1981. Although the frequency, severity, and duration of emergencies varies significantly among the 219 (73 × 3) precinct-tours of the City, the time available for the study made it impossible to perform 219 separate analyses. Therefore, we decided to analyze three representative precincts in detail and, after doing so, try to extrapolate the results City-wide.

This paper details the assumptions made, the analysis performed, and the conclusions reached. It also provides some insights into the difficulties of modelling a highly-complex, loosely-managed existing system and an incompletely-defined proposal in a very short period of time in the midst of a sensitive political situation.

2. An Overview of Police Patrol and the 911 Emergency Response System in New York City

Calls for police service are received centrally at Police Headquarters in Lower Manhattan where each is automatically routed to a bank of operators who handle the geographic area (borough) from which it originates. The system is engineered and staffed to insure that approximately 99% of all telephone calls are answered by an operator within 30 seconds. Once the operator has verified that the call needs a police response, he or she types the time, location, type of incident, and other information into the 911 computer system (SPRINT), which then transfers the call electronically to the dispatcher for the appropriate precinct. Each dispatcher is responsible for all patrol forces in two or four adjacent precincts (a division). Except in dire emergencies, each precinct operates separately; RMPs are typically not assigned across precinct boundaries. The dispatcher's major functions are to keep track of each RMPs status and to assign them to the 911 emergency calls—usually called "jobs" by the police. Jobs are characterized by one of 144 incident codes grouped into 7 priority levels. The dispatcher uses this priority scheme and other job specific information to determine the order in which jobs are assigned as cars become available. A separate job queue is displayed on the dispatcher's screen for each precinct.

The major functions of RMPs are to respond to 911 calls, to perform preventive patrol, and to react to incidents encountered during patrol ("patrol initiated activities," which are also entered into SPRINT). During their tours of duty RMPs frequently go "out-of-service". Out-of-service time includes such activities as precinct assignments, car repairs, and meal breaks, and can account for 20–60% of a car's time. Thus, the number of cars actually functioning in a precinct at any given time is often far less than the number assigned.

Official NYCPD policy is that one RMP is dispatched to a job and this "primary" car may request one or more "backup" cars either at dispatch or upon arrival on the scene. This nominal policy does not itself constitute a basis for modelling and analyses, since patrol cars often "assign themselves" to back up a job when they hear radio transmission from the dispatcher to the primary car.

In understanding the command and control process of the patrol force, it is important to note that the dispatchers do not supervise or command the RMPs. Field supervision is nominally performed by sergeants assigned to patrol. However, they are often unavailable during all or much of a tour of duty. It is not uncommon to have only one sergeant available in an entire division. Superior officers are rarely seen in the field, and, as a consequence, the patrol system is rather loosely managed.

The sequence of events between the occurrence of a hypothetical emergency that gets two patrol cars and the arrival at the scene of both patrol cars assigned is illustrated in Figure 1. The figure will clarify some of the concepts and terminology employed in our analysis. (The durations of the various time intervals are not drawn to scale.) In particular:

(a) For many calls, the interval that defines the potential effectiveness of police response is the total elapsed time between the occurrence of the incident and the arrival of adequate police forces on the scene. There is no effective control of the length of the reporting delay, which may be quite long (Kansas City Police 1977).

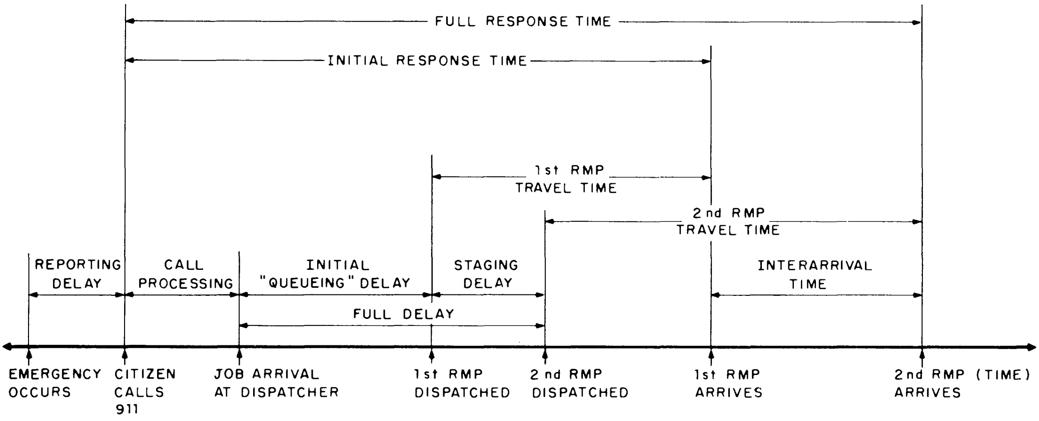

FIGURE 1. Schematic Diagram of the Sequence of Events During Dispatch and Response of a Typical 911 Call for Emergency Police Service.

(b) The queueing model estimates only those delay components due to RMP unavailability. This encompasses the intervals called "initial delay," "full delay" and "staging delay." The actual duration of these delays may also be affected by the dispatcher's ability to handle the communication-control workload with all the cars in the several precincts he or she handles. It is very likely that the complexity of these tasks would be affected by a change from two officers to one officer in a patrol car. *Full delay*, the interval between job arrival at the dispatcher's terminal and dispatch of the last RMP assigned to the job, proved to be the key performance measure in most of our analysis. *Initial delay*, the interval between job arrival and dispatch of the first RMP, is also an important measure and is related to the public's perception of police responsiveness. *Staging delay*, the interval between the dispatch of the first RMP and the dispatch of a back-up, is particularly important to the safety of officers in a one-officer program.

(c) Travel times are important but are not explicitly part of the queueing analysis. Management concerns, time constraints, and the unavailability of data did not allow for a complete analysis of this component of delay. Some travel time results were calculated based on the results of the queueing analysis. As indicated in Figure 1, full and initial response times incorporate call processing, dispatch delay, and travel times and are more complete measures of performance.

3. The Model

Since 1974, the NYCPD had been using a computer program developed by The New York City-Rand Institute to help allocate patrol cars among precincts. The program, a precursor of The Patrol Car Allocation Model (PCAM) of Chaiken and Dormont (1978), was based on the M/M/c nonpreemptive priority queueing model of Cobham (1954). NYCPD management were aware that a serious shortcoming of their model was its inability to explicitly represent multiple car dispatches, a critical issue in comparing one- and two-officer programs. We therefore developed and employed the multiple car dispatch (MCD) priority queueing model described below. Work on the robustness and validity of the final version of the MCD model is reported in Green and Kolesar (1984, 1983).

The MCD queueing model represents an individual precinct-tour. All queueing statistics calcuated are steady state results—effectively viewing each precinct-tour as being of infinite duration. The model estimates delays due only to patrol car unavailability. This was thought to be the largest component of the total delay of a 911 job and the component most sensitive to the change from two officers to one officer per car. The key statistics calculated are the initial delay, staging delay and full delay illustrated in Figure 1 and discussed in §2. The assumptions of the MCD model are that:

(1) There is a fixed number, s, of identical cars on patrol duty at all times.

(2) Jobs arrive according to a Poisson process at a constant rate.

(3) A job belongs to priority class $k = 1, 2$ with probability p_k, and class 1 jobs have nonpreemptive priority over class 2 jobs. Jobs are dispatched in the order of their arrival (FIFO) within each priority class.

(4) A class k job requires service from i cars with probability $c_k(i)$.

(5) If a job arrives at the dispatcher and enough cars are available it begins service immediately, i.e., the required number of cars is dispatched to the scene. Otherwise, when a job becomes first in queue, it is assigned cars as they become free, these cars are held for the job, and service does not begin until the required number are available.

(6) Completion times for each car on a job are independent, identically distributed (i.i.d.) exponential random variables.

Each of these assumptions is untrue to some extent. We examine each in turn and discuss its potential effect on the accuracy of the analysis.

(1) Cars go out-of-service at various times during the tour and fluctuations in the number of effective cars—cars actually operating—can be significant. The pattern of out-of-service times is unpredictable. Since delay curves are convex in the number of cars available, the actual expected delay will thus exceed the expected delay that would be experienced if the number of effective cars were constant. Thus our steady-state model will tend to underestimate delays. But this occurs in both a one- and two-officer system and there is no a priori reason to believe that the results will be biased in either direction.

(2) Recent tests on the distribution of 911 calls throughout the day indicate that call arrivals can reasonably be modelled as a time-dependent Poisson process (see Green and Kolesar 1983). However, the marked nonstationarity may cause greater delays than if the arrival rate were smooth. This would be true for both modes of patrol car operation.

(3) The NYCPD had 7 dispatch priority levels. Since the primary objective of the study was to compare the performance of the two modes of operation with respect to the highest priority jobs, NYCPD officials had grouped all incidents into two classes—high and low. Although preemption of lower priority jobs is sometimes possible, it rarely occurs. Thus a two-class nonpreemptive priority model appeared appropriate.

(4) There was no definitive objective policy regulating the number of cars to be dispatched for a given type of incident. The dispatcher, officers in patrol cars, and the level of system congestion determined how many cars would respond to a given job. This will be discussed further in §5.

(5) As soon as a car is assigned to a job, it travels to the scene. Although in some cases the first car on the scene must wait for a second or third car before action can be taken, in many multiple car responses some "service" begins as soon as the first car arrives. Yet, the point at which service begins is not clearly identifiable; there is no recorded information on when cars arrive at the scene and start "working."

(6) Originally, we considered this assumption most suspect. Besides the exponentiality itself, we were concerned over two independence assumptions: (a) RMP service time completions for the same job are independent of one another, and (b) service time is independent of the incident type and number of responding units. In §5 we show that these assumptions are reasonable reflections of reality. More details are given in Green and Kolesar (1983).

4. Data Needs and Availability

As with most queueing systems, two parameters are needed to characterize a precinct-tour: the *overall call rate*, which is a fundamental measure of the level of demand for service; and the *service rate*—the average service time per car per call. But the MCD model also requires the proportion of calls by priority class and the distribution of the number of cars needed by these priority classes. This two-way characterization is what we call the *severity* of a precinct-tour. In understanding the role of severity, it is important to bear in mind that it simultaneously reflects several more fundamental characteristics: crime severity, patrol car workload, and potential danger to patrol officers. Much of our analysis centered on characterizing precincts by both call rate and severity. (Service rate, we shall show, may reasonably be modelled as being constant across precinct-tours.)

These data are needed for each precinct-tour modelled. The major source of data on NYCPD patrol activities is the 911 SPRINT computer system. For every patrol car dispatched, a computer record is made containing, along with other information, the arrival time of the job, its location, its incident code, the time of dispatch, and the time of service completion of the car. A separate record exists for each car that responded to the job. Unfortunately, this data base does not contain explicit information about the number of cars on patrol or when cars are out-of-service.

Seeking to finesse the need to create 219 detailed models, we asked OMB and NYCPD to select three precincts and represented typical "light," "moderate," and "heavy" precincts in terms of the amount and severity of 911 calls and in terms of danger to the officers. (We were purposely vague in this request for we felt that police judgment and intuition were important in the selection.) The "light" precinct selected is a largely middle- to upper-class, moderate population density, residential neighborhood in Queens[1] that has an area of 5.2 square miles and a relatively low incidence of personal crime. The precinct chosen as "moderate" is in lower Manhattan, covers 0.52 square miles and is a partly commercial, partly residential neighborhood with moderate levels of both personal and property crimes. The "heavy" precinct is in densely populated central Brooklyn. It has an area of 1.73 square miles and high levels of crime in all categories. NYCPD supplied us with the SPRINT tapes for June and July 1980 for these three precincts. These months were suggested by NYCPD as a period with high levels of activity. The tapes contained about 20,000 incidents.

The NYCPD provided for each of the 144 incident codes, a designation as either high or low priority and a requirement of either one- or two one-officer cars for *initial* dispatch. (This was a tentative policity developed largely for the purpose of negotiations with the PBA.) Thus, we were able to group incidents into four *aggregate categories*—$H2$, $H1$, $L2$, $L1$. H jobs have the highest dispatch priorities and L jobs have the lowest while the number following H or L indicates the number of one-officer cars to be initially dispatched. $H2$ jobs are, as expected, the most severe incidents and include many types of crimes in progress such as robberies, burglaries, and assaults as well as reports of shots fired and calls to assist a police officer. $H1$ jobs are typically ambulance cases and residential incidents for which there is less potential danger to responding officers. $L2$ is the smallest category, including only pickups of emotionally disturbed persons. The $L1$ category includes incidents such as past burglaries and larcenies, reports of a disorderly person or noise, and street accidents.

This 4-way grouping did not provide information on the *total* number of cars that *should* respond to each incident type, nor were the OMB or NYCPD prepared to take an official position on this sensitive issue (see *New York Times* November 2, 1981).

[1] The identities of the sample precincts have been masked in order to preserve confidentiality.

This dispatch policy information was, however, crucial to our analysis since it reflects the key difference between a one- and two-officer program. This information gap forced us to develop dispatch policies ourselves for both programs in order to carry out the analysis. This is discussed in the next section.

In addition to detailed data for the three precincts, we obtained the most current aggregate statistics used in the NYCPD computerized patrol car allocation program. This gave us, for each of the 73 precincts and the three tours of duty, the call rate and the percentage of calls in the three priority classes used in that program—high, medium and low. (In §6 we describe how we linked this 3-way grouping to the 4-way grouping of $H2$, $H1$, $L2$, $L1$ in order to estimate City-wide parameters, so we could make extrapolations from the detailed analysis of the three selected precincts.)

5. Analysis of the Three Sample Precincts

5.1. *Call Rates and Severity Distributions*

We first analyzed the data from the SPRINT computer for the three sample precincts (light, moderate, and heavy) to determine the call rates and the severity distribution of jobs over the four aggregate code groups. These appear in Table 1 for each precinct-tour. While the distributions are not dramatically different among the precincts, we see that for each tour, the call rates, the total percentages of high priority jobs, and the proportions of $H2$'s—the most serious incidents—are ordered among the precincts as expected.

5.2. *Service Times*

We were interested in answering several questions about service times:

(1) *How well do the service time distributions fit the model's assumption of i.i.d. exponential random variables*? This, of course, is important in judging the accuracy of the model's predictions.

TABLE 1

Severity Distribution for Three Sample Precincts

	% of Jobs in Each Group				Average No. of Calls Per Hour
	H_1	H_2	L_1	L_2	(λ)
"Light" Precinct					
Tour 1	15.2	24.8	55.4	4.5	2.1
2	18.0	14.8	64.3	2.9	2.7
3	14.5	17.6	64.1	3.8	4.0
Average	15.7	18.5	62.1	3.7	2.9
"Moderate" Precinct					
Tour 1	13.1	26.6	56.2	4.0	4.3
2	16.8	15.9	65.1	2.2	4.2
3	14.5	19.6	63.5	2.3	6.4
Average	14.7	20.6	61.9	2.8	5.0
"Heavy" Precinct					
Tour 1	16.2	31.9	50.0	1.9	5.3
2	17.0	21.4	59.5	2.1	6.9
3	16.5	24.4	57.2	1.9	9.7
Average	16.6	25.3	56.2	2.0	7.3

TABLE 2

Average Service Time per Car by Job Group and Number Working

	Number of Cars Working		
	1 Car Jobs	2 Car Jobs	3 + Cars
Job Group: *H*1			
1st Car	29	39	47
2nd Car	—	19	33
3rd Car	—	—	16
Avg.	29	29	32
Job Group: *H*2			
1st Car	23	27	31
2nd Car	—	17	23
3rd Car	—	—	13
Avg.	23	22	22
Job Group: *L*1			
1st Car	28	36	57
2nd Car	—	14	37
3rd Car	—	—	14
Avg.	28	25	36
Job Group: *L*2			
1st Car	28	41	53
2nd Car	—	19	27
3rd Car	—	—	14
Avg.	28	30	30

(2) *How do service times per car vary by aggregate job code*? Since the model assumes that the average service time per car is the same for all jobs, this is also relevant to the model's reliability. Yet, since job codes are a reflection of job severity, differences among them were expected.

(3) *How do service times per car vary among precincts and tours*? Differences in precinct size and traffic congestion might cause differences in travel times, which are part of service times. Higher levels of job severity and danger in "heavy" precincts might result in longer service times than in "light" precincts. Greater availability of cars on some tours might result in longer average service times.

Tables 2 and 3 contain summaries of service time statistics. Sample sizes are large enough that the patterns are not statistical artifacts—whatever their explanation, the patterns are real.

Table 2 gives average service times per car on jobs that had a one-car response, a two-car response and a three- or more car response for each of the four aggregate code groups over all precinct-tours. Data are presented for the longest working car, next longest working car, etc. We observe that patterns are quite similar across aggregate job codes with the exception that, counter to intuition, *H*2's—the most serious calls—have the lowest per car service times. An important observation is that the

TABLE 3

Average Service Time Per Car by Precinct and Tour (Minutes)

	Tour 1	Tour 2	Tour 3	All Tours
Light Precinct	24.0	28.2	29.6	27.8
Moderate Precinct	20.8	25.3	23.0	23.0
Heavy Precinct	28.0	29.7	28.4	28.7
All Precincts	24.6	28.1	26.9	27.1

average service time per car is not dependent on the number of cars responding to a job.

The spread of the mean service times by car on multiple-car jobs is generally close to that predicted by the assumption of independent, identically distributed exponential service times. For example, if service time per job were exponentially distributed with mean $= 1/\mu = 30$ minutes, then on a two-car job, the first car to free up would have an average service time of $1/2\mu = 15$ minutes and the second car's expected duration would be $3/2\mu = 45$ minutes. Similarly, for a three-car response, the mean service times would be 10, 25 and 55 minutes. In particular, the actual average service times of the $L2$ jobs in Table 2 have an average service time per car of about 30 minutes and their spreads are remarkably close to the theoretical values. Histograms of service times showed the characteristic exponential pattern—a mode near zero and a long tail to the right. Coefficients of variation were very close to one. For more support of the exponential assumption, see Green and Kolesar (1983) which details our later and more extensive analyses of this issue.

Table 3 shows the mean service time per car by precinct-tour. Counter to our prior expectations, the results show that Tour 1, which has the lowest average call rate, also has the lowest average service time. This is particularly surprising since this tour also has the highest proportion of $H2$ jobs (see Table 1). The only other consistent pattern is that the average service time is lowest for the moderate precinct.

Before drawing modelling conclusions from these data, we must recall how service times are reported and what they actually represent. Service time as recorded on the SPRINT system begins when the dispatcher assigns the car to an incident or (for self-assignment) when the car notifies the dispatcher that it is responding. Service time ends when the car radios the dispatcher that it is again available for assignment or, alternatively, that it is going out-of-service. There can be inaccurate reporting on both ends. For example, a car may forget to notify the dispatcher of the termination of work on a job, or cars may respond as backups without informing the dispatcher. It is important to note that an NYCPD policy mandates the average service time per car be no more than 30 minutes. In an effort to comply with this policy, cars have been observed to report termination on a job before they're really done, and to assign themselves to a job after they've already arrived at the scene.

Given this background, and given the lack of significant explainable patterns of service time by precinct, tour, or job type, the needs of the model, and our time constraints, we decided to use a single value for the mean service time per car in the analysis of the two-officer program. We chose 30 minutes because it was close to the overall average of the 3 sample precincts (as well as other precincts we had examined in our previous work), and because it was consistent with stated NYCPD policy.

Unfortunately, there was no comparable set of data from SPRINT (or indeed from any source) for a one-officer program, though a very limited one-officer program had operated experimentally in several New York City precincts. We therefore had to extrapolate from the existing two-officer service time data. Although call rates and the distribution of incidents into aggregate job groups are independent of the mode of patrol car operations, average service time per car might well change with one-officer operations. NYCPD officials believed that the service time of a single officer working alone would likely be longer than with a partner. They reasoned that with two officers, one could be filling out reports while the other was dealing with the incident itself. This belief was somewhat supported by a study of one-officer patrol conducted in San Diego (Boydstun et al. 1977). For jobs with a multiple-officer response, there was no apparent reason for believing that service times per car would be different with one-officer operations. After discussion with NYCPD officers, but without their official endorsement, we decided to model service times for the one-officer program to be equal to those for the two-officer program for all jobs that received more than one officer, and to allow a 20% increase in the mean service time per car for jobs where an officer worked alone. This could not be done explicitly because the model assumes the same average service time per car for all jobs. However, we could achieve the proper total workload by increasing the overall average. In particular, under the dispatch protocols we used, we approximated a 36-minute average for these jobs and a 30-minute average for all others by using a mean service time per car of 33 minutes.

5.3. *Dispatching Protocols*

To determine a protocol for dispatching under a two-officer program, we first examined the distribution of the number of cars actually responding to calls. Differences between precincts and tours were small and did not follow consistent patterns. Variations among the aggregate job groups made more sense and are shown in Table 4. Note that $H2$'s have the highest proportion of multiple-response jobs and $L1$'s the lowest while $H1$'s and $L2$'s have very similar distributions.

These numbers, however, could not be taken at face value. From our previous experience we knew that historical response levels reflected a variety of nonpolicy factors that sometimes had undesirable results. For example, data showed that low crime precincts in Staten Island had a higher average response per job than high crime precincts in central Harlem. A likely explanation is that in Staten Island, where both the call rate and severity of incidents is very low but large areas must be patrolled, cars are usually available and eager to respond to incidents, even when not needed. In Harlem, cars spend a greater fraction of time working at jobs and are often unavailable as backups, even when needed. In summary, the number of cars responding to a job was very state-dependent. This was due partly to loose management of the patrol system, partly to geographical differences, and partly to the existing allocation model, which did not properly account for the number of cars required per job. Thus, direct use of these historical distributions would have been inappropriate and misleading.

In determining our two-officer dispatch "protocol," we took as starting points the NYCPD minimum initial dispatch rules for one-officer cars, and the historical distributions of cars used. We modified these based on comments from NYCPD officials and on our own field experience and judgment as follows: for $H1$ and $L1$ jobs we reasoned that if official NYCPD policy was an initial dispatch of one car under one-officer operations, then one car with two-officers should be adequate for most of these jobs. $L2$ jobs, which are narrowly defined and few in number, pose little possibility for the unexpected and the official initial dispatch of two-officers seemed very reasonable as a standard for total response. $H2$ jobs, which range from past robberies to bank robberies in progress, are the most difficult to model. In general, they are incidents in progress that have an element of danger. We concluded that the two-officer initial dispatch was not an adequate level of total response and our dispatch protocol for them under a two-officer program was two cars—a total of four officers. In summary, our two-officer protocol sends 2 cars to $H2$ jobs, 1 car to $H1$

TABLE 4

Distribution of Number of Cars Used by Job Type

		Number Used		
		2 Car		Average
Job Type	1 Car	(Percent)	3 + Cars	No. of Cars
$H2$	61	26	13	1.56
$H1$	76	18	6	1.33
$L2$	72	21	7	1.43
$L1$	86	12	2	1.20

jobs, 1 car to $L2$ jobs, and 1 car to $L1$ jobs, and it became known as the 2-1-1-1 protocol.

We also needed to specify a dispatch protocol for one-officer cars. For the low priority jobs, the official initial dispatch policy of 2 cars for $L2$ and 1 car for $L1$ seemed a good standard. We reasoned that sending 2 officers to $L2$ jobs was appropriate since both officers might be necessary to handle a very disturbed person. For $L1$'s, which are typically incidents such as a report of a past burglary or a noisy party, one officer seemd to be quite adequate. The existence of large numbers of these $L1$ jobs was, of course, the main motivation for considering a one-officer program in the first place.

The only reasonable candidates for the number of cars that should be assigned to $H1$ jobs were one or two. These are ambulance cases where the police act primarily to control onlookers and traffic. Since it is easy to imagine situations in which a large crowd gathers, we decided to be conservative and specify a two-car response as the standard.

To be consistent with a two-car response to $H1$'s under two-officer operations, the one-officer car response for $H2$'s had to be either three or four cars. We decided on three cars for the following reasons:

—The three one-officer cars bring three officers to a job, which is close to the average number responding historically under two-officer operations (1.56 cars = 3.12 officers).

—Four cars would imply an average service time for the primary car of 62.5 minutes and a total service time per job of over 3 hours, both of which were excessive as compared to current experience.

—Since the primary measure of performance was delay until the *last* car dispatched is available, this would imply that no service would begin until the fourth car was assigned—a very unlikely situation.

Our dispatching rule for the one-officer model became known as the 3-2-2-1 protocol.

Although neither of these dispatching protocols was officially accepted by either NYCPD or OMB (nor could they be), there was no objection to them as reasonable policies for assigning cars to jobs under a more tightly managed system.

5.4. *Effective Cars*

The queueing model assumes that a car is either busy on a job or available for assignment, yet we know that cars go out-of-service for meals, gas, precinct assignments and the like. The only data on out-of-service times came from an experiment conducted in one precinct in 1973 by The New York City-Rand Institute, which showed that cars were out-of-service about 30% of the time. This figure was being used by NYCPD in its own planning and we adopted it for all precincts and tours for both one- and two-officer modes of operation. We made the assumption that out-of-service time is uniform across the tour. This is only a rough approximation since out-of-service times cluster at the very beginning, very end, and middle of the tour. We use the term "fielded cars" to indicate the number of cars actually assigned to patrol and the term "effective cars" to indicate the net numbers of cars available after accounting for out-of-service time. Thus, 0.7 × fielded cars = effective cars, or fielded cars = 1.43 × effective cars.

5.5. *Results*

The queueing model was run in one-officer and two-officer mode for various numbers of servers in each of the 3 precincts to generate tables of queueing delays and

TABLE 5

Queueing Results for "Heavy" Sample Precinct

		Two-officer RMP				One-officer RMP					
		RMPs				RMPs					
Tour	Average No. of Jobs Per Hour	Eff.[1]	Field.[2]	Aver. Avail.[3]	Priority Delay[4] (mins.)	Eff.[1]	Field.[2]	Aver. Avail.[3]	Priority Delay[4] (mins.)	% Incr. in RMP's	% Decr. in Off.
1	5.3	5	7.8	1.3	10.6	8	11.3	3.1	6.7	61	20
		6	8.7	2.4	4.1	9	12.8	4.2	3.2	50	25
		7	10.1	3.4	1.7	10	14.2	5.3	1.5	43	29
2	6.9	5	7.2	0.8	15.8	8	11.3	2.3	9.8	61	20
		6	8.7	1.8	6.3	9	12.8	3.3	4.9	50	25
		7	10.1	2.9	2.6	10	14.2	3.5	2.4	43	29
		8	11.5	3.5	1.0	11	15.6	5.4	1.4	38	31
3	9.7	7	10.1	9.9	11.6	10	14.2	1.7	12.4	43	20
		8	11.6	2.0	5.6	11	15.6	2.8	6.9	38	31
		9	12.8	3.0	2.8	12	17.0	3.9	4.0	33	34
		10	14.2	4.1	1.4	13	18.5	4.5	2.3	30	35

[1] Effective cars = average number of cars operating "on queue" during the tour.

[2] Fielded cars = effective cars × 1.43.

[3] Average RMP availability = average number of cars on patrol.

[4] Average priority delay = average full delay (time until all cars are sent) for calls in categories $H1$ and $H2$.

other performance measures. As an example, Table 5 presents results for the "heavy" precinct. The table is arranged so that "comparable" levels of expected full delays to high priority jobs generally appear on the same line for both modes of operation. For example, on Tour 1, with 6 effective two-officer cars, the expected delay to high priority jobs is 4.1 minutes, while 9 effective one-officer cars are required to achieve a 3.2 minute average delay. One more effective car in each mode produces delays of 1.7 and 1.5 minutes respectively. The right-most column of the table shows the percentage increase in one-officer RMP's necessary to achieve roughly the equivalent level of performance as with two officers in a car.

Several pertinent observations can be made from the tables. First, the required increase in one-officer cars above the "equivalent" number of two-officer cars varies from 30% to 60% depending on the performance level chosen—that is, on the expected delay to high priority jobs selected as a benchmark. Similiar results were obtained for the other sample precincts. Another significant result is that the average number of RMPs available for patrol is considerably higher for one-officer cars than for two-officer cars at every staffing level. Note that the ratio of one-officer to two-officer cars decreases with increasing level of performance. At the time, the NYCPD was fielding 8, 8, and 11 two-officer patrol cars on Tours 1, 2, and 3, in the "heavy" precinct. These levels were quite low, resulting in high delays and high required increases for one-officer cars.

At this point in the study OMB was quite optimistic since these were the first results they had seen which indicated that a one-officer program could be economically attractive. We proceeded to extend our analysis City-wide.

6. Extension and Extrapolation to a City-Wide Model

Our goals were to estimate the City-wide number of one-officer cars needed to achieve an equivalent level of performance to two-officer cars, and to identify those precinct-tours in which a one-officer program would be most effective. Our analysis of the 3 sample precincts had indicated that the distinguishing characteristics of a precinct-tour were its call rate and severity distribution. So the next task was to categorize each of the City's 219 precinct-tours by these two characteristics. As mentioned in §4, the only City-wide data available was a listing of the most recent inputs used in the NYCPD patrol car allocation program, which included, for each of the 219 precinct-tours, the call rate and the fraction of calls that fell into NYCPD's definition of high, medium and low priority. Our problem was to use these data to develop approximate severity distributions that could capture the diversity of all 219 of the City's precinct-tours, yet would require only a modest amount of computing with the MCD queueing model.

We decided to develop three prototypical precinct models for severity distribution—light, moderate, and heavy, and use the city-wide NYCPD priority breakdowns to group all 219 precinct-tours into these three categories. To accomplish this, we first examined the NYCPD City-wide priority breakdown graphically and visually derived a light, moderate, and heavy distribution of the three priority classes. Our intention was to define a moderate distribution that would approximate the average over all precinct-tours, and light and heavy prototypes that would be symmetric about the average and fall somewhere in the tails of the distribution of severity across all precinct-tours. The problem of representing a distribution of 219 vector-valued data points is a difficult and subjective task, and our procedures were somewhat intuitive.

We then converted these prototypical priority distributions to severity distributions. Using the data from the 9 sample precinct-tours, we regressed the fraction of $H2$ jobs against the fraction of high priority jobs, the fraction of $L1$ jobs against the fraction of low priority jobs, and the fraction of $H1$ plus $L2$ jobs against the medium priority jobs. The results produced good predictor equations for $H2$ and $L1$, but no relation for the $H1 + L2$ grouping. Using the regression equations, we derived equivalent fractions of $H2$ and $L1$ jobs for the light, moderate, and heavy scenario from the original prototype distributions. We then used the severity distribution data from the 3 sample precincts to divide the remaining percentage of jobs between the $H1$ and $L2$ categories according to $P_{L2}/(P_{L2} + P_{H1})$ where P_N is the proportion of the jobs with aggregate code N. We thus obtained the prototype severity distributions given in Table 6.

We grouped the 219 precinct-tours into the three severity categories by plotting a scattergram of every precinct-tour on a coordinate system with one axis the proportion of $L1$ jobs and the other axis the proportion of $H2$ jobs as obtained from the regression equations. We also plotted our prototype precincts on this graph and we designated each precinct-tour as "light," "moderate," or "heavy" according to its position on the scattergram relative to the prototype distributions. The scattergram is shown in Figure 2. It shows that our prototypes fairly capture the full range of severity distributions.

We used the queueing model to generate results for each prototype severity distribution under the dispatch protocols developed earlier for one- and two-officer operations. For each severity mode-protocol combination, the queueing model was run for call rates ranging from 0.5 calls per hour to 12 calls per hour. This covered the range of call rates in the City-wide data. For each call rate, tables of delays and car utilizations were

TABLE 6

Severity Distributions by Precinct-Tour Category

Precinct-Tour	Job Type			
Category	$H2$	$H1$	$L2$	$L1$
Light	14%	12%	3%	71%
Moderate	22%	18%	3%	57%
Heavy	29%	22%	3%	46%

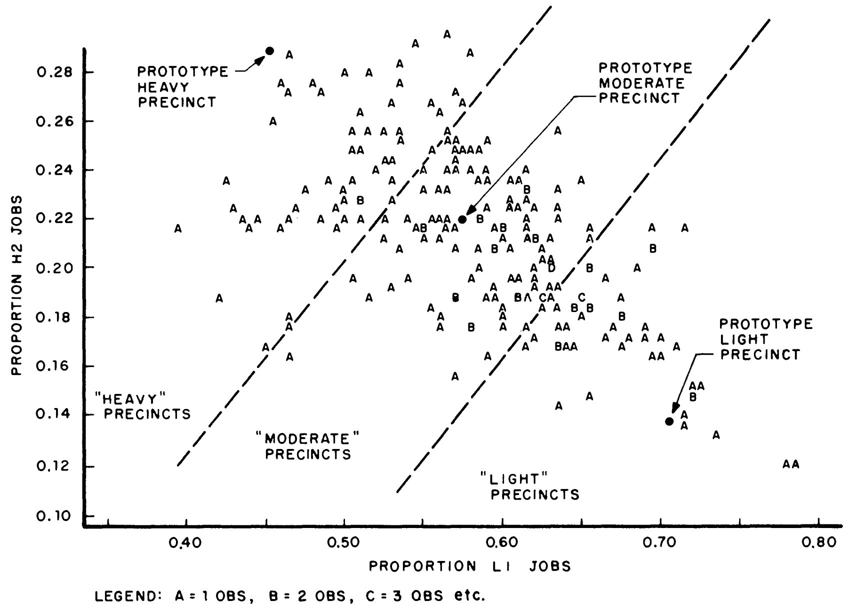

FIGURE 2. Scattergram of 219 Precinct-Tours.

generated for a range of staffing levels from the lowest for which a solution was possible up to one that produced essentially zero average delay.

7. City-Wide Results

The City-wide distribution of the 219 precinct tours by call rate and severity type is shown in Table 7, which provides a hypothetical illustration of our final results.[2]

Table 7 was created as follows: From the queueing tables generated for the prototype precincts, we determined for both one- and two-officer dispatch protocols the number of effective cars needed to insure a maximum expected delay to high priority jobs of (in this case) 3 minutes for all relevant call rates. This number is shown in the table along with the additional cars needed for the one-officer program. The number of cars needed to be fielded was obtained as before by multiplying effective cars by the out-of-service factor of 1.43. The results show that, compared to a two-officer program, the one-officer program requires about a 35% increase in the number of cars fielded and yields a decrease of about 32% in police officers in patrol cars.

To test the sensitivity of these results to the standard of performance, similar calculations were performed for several other levels of delay. These analyses showed that for a range of maximum expected delay to high priority jobs of 1/2 to 10 minutes one-officer programs were always attractive vis a vis two-officer programs. In this way, we answered the primary question of how many additional one-officer cars were needed to achieve the same level of performance as with two-officers in a car. The results also indicated those precincts in which a one-officer program would have the biggest economic impact. Contrary to prior beliefs of both OMB and NYCPD, Table 7 shows that the greatest advantages can be realized in the precincts with the highest call rates and that this is true for each severity class.

In addition to expected delay to high priority jobs, we compared the performance of the two systems with respect to several other measures. We found that with the

[2] The actual results have been coded to preserve confidentiality.

TABLE 7

City-Wide Requirements for Maximum Expected Delay of 3 Min. to High Priority Jobs

	"Light"			"Moderate"			"Heavy"		
# Jobs/	# of	Eff. Cars		# of	Eff. Cars		# of	Eff. Cars	
Hr.	Pct-Tours	2	1	Pct-Tours	2	1	Pct-Tours	2	1
1	2	4	5	2	4	5	6	4	5
2	4	5	6	18	5	6	17	5	7
3	3	5	7	25	5	7	18	6	8
4	7	6	8	23	6	8	9	6	9
5	8	6	9	15	7	9	4	7	10
6	8	7	9	12	7	10	3	8	11
7	3	7	10	9	8	11	1	8	11
8	3	8	11	2	8	11	2	9	12
9	2	8	12	5	9	12			
10	1	9	12	2	10	13	1	10	14
11				1	10	14			
Total	41	259	354	114	713	953	61	359	496

		Effective	Fielded
Cars Needed:	Two-officer	1331	1901
	One-officer	1803	2576
Difference:			+ 675 (35%)
Officers Needed:	Two-officer	3802	
	One-officer	2576	
Difference		− 1226 (32%)	

Note: The table contains results for 216 of the City's 219 Precinct-tours. The model was not applied to the 3 precinct-tours with call rates lower than 1 call per hour.

number of cars needed to achieve a specified expected delay to high priority jobs, other measures including overall delays, initial delays, proportion of jobs delayed, and patrol car utilization did not deteriorate as a consequence of the change from two- to one-officer per car. An example of such comparisons for a particular precinct-tour is shown in Table 8. Except for the travel times, all the numbers reported there are produced directly by the queueing model. Thus, we find that for the six calls per hour moderate severity precinct illustrated with six two-officer cars, average delay to high priority jobs is 3.9 minutes, while the nine one-officer cars the average delay to high priority jobs is 2.9 minutes.

The effect on travel times of changing from six two-officer to nine one-officer cars was estimated from the queueing statistics by using the well established relationship that expected travel distances are approximately inversely proportional to the square root of the expected number of available patrol cars and that travel time is approximately proportional to travel distance (Chelst 1981, Kolesar and Blum 1973, Larson 1972). (In a recent paper Chelst 1981 applies these ideas to analyze the impact on response times of a one-officer program. While his work does not carry over exactly, our approach and conclusions are similar.) The square root law for travel times states that $ET_i \approx k_i/\sqrt{\alpha}$ where ET_i is the expected travel time of the ith closest car to respond, α is the expected number of cars available to respond and k_i is a scale parameter depending on the geometry of the response environment. Both theoretical and empirical evidence indicates that $k_2 \approx 2k_1$, i.e. the second of two arriving cars travels roughly twice as far as the first car to arrive. Noting from Table 8 that α for two-officer cars is 2.3 and α for one-officer cars is 4.0 we estimate that the expected travel time of the first car in a one-officer program is roughly $\sqrt{2.3/4.0}$ or 76% of the expected travel time for the first car for the two-officer program and consequently the

TABLE 8

Detailed Performance Characteristics of a Selected Precinct-Tour

	Two-officer Cars	One-officer Cars
Effective cars	6.0	9.0
Fielded cars	8.6	12.9
Officers needed	17	13
Available cars	2.3	4.0
Car utilization	61%	55%
High Priority Jobs		
Percentage delayed	31%	25%
Average delay	3.9 min..	2.9 min.
Avg. initial delay	2.7 min.	1.5 min.
Avg. delay of delayed jobs	12.6 min.	11.6 min.
Low Priority Jobs		
Percentage delayed	25%	14%
Average delay	4.9 min.	2.2 min.
Avg. delay of delayed jobs	19.6 min.	15.7 min.
Average Travel Distance*		
First car	1.00	.76
Second car		1.52
Average of both cars	1.00	1.14

*Relative to first two-officer car.

expected travel time of the second one-officer car to arrive is twice that of the first or 52% longer than the first two-officer car. Thus the first one-officer car arrives substantially sooner, but the second substantially later. The significance of these results with respect to police officer safety is unclear. It depends largely on the policy that would be adapted for sending one-officer cars to the scene of a potentially dangerous incident—that is, whether cars would rendezvous first or arrive at the scene one at a time, and what actions the first-arriving officer would take prior to the arrival of a backup.

8. Conclusions

We drew three major conclusions from this study:

(1) From the point of view of response time, a one-officer patrol program was feasible in New York City. For most precinct-tours, an equivalent level of response time could be achieved with significantly fewer police officers.

(2) Though we had to some extent considered police officer safety by using as our primary measure the expected full delay, which includes the staging delay, this crucial issue warranted further examination. We had not had time, for example, to explicitly calculate the probability that a high priority, dangerous incident would suffer a staging delay, or the average time to get a backup for such a job.

(3) Our findings should be regarded as tentative because they were made with limited data for the current two-officer system and no data on one-officer performance. Moreover, our model assumed that patrol cars would be more tightly managed and would respond according to the protocols we had defined, and also that no significant additional delays would be introduced in a one-officer program due to increased loads on the dispatchers.

On July 7, 1981, we briefed the Deputy Mayor for Operations, the Police Commissioner, the Budget Director, the Director of the Office of Muncipal Labor Relations, and their staffs. We described the model and the data used, presented the results of our analysis, and stated our conclusions. In summarizing our position, we recommended that the City not implement a large-scale one-officer program without further study, but that it develop and undertake a carefully monitored experimental program. Such a program should serve not only as a means of obtaining data on one-officer patrol, but also as a vehicle for defining and testing policies for dispatching and management of cars that we felt were particularly crucial for a successful program.

Although the Police Department concurred with our concerns on these matters, it appeared for a while in the Fall of 1981 that, under OMB initiative, the City would proceed to implement a one-officer program in 39 precincts without appreciable further study and no experimentation. However, negotiations broke down when the PBA demanded the City guarantee that the number of cars fielded be adequate to assure rapid backup, and the City was unwilling to write such a guarantee into the contract (*New York Times* November 2, 1981).

Negotiations resumed in the summer of 1982. The results of our analysis were used to justify the fiscal attractiveness of the one-officer program and the City's further exploration of it. In the meantime, with outside funding we had continued research on modelling police patrol operations and in the process had collected more data from NYCPD. While confirming the general validity of our analysis, these data also substantiated our feelings that more study was needed prior to implementation of a one-officer program. We found that out-of-service times were substantial (during periods of peak activity they could be more than twice as high as the figure we had been given to use in the feasibility study). We also found that a large fraction of high priority calls were being delayed for substantial periods even when patrol cars were available. (This was apparently due to dispatcher-communications overloads (Green and Kolesar 1983), which would probably worsen with more (one-officer) cars.) We extended the capability of the queueing model to estimate staging delays directly and also collected empirical data on them. This analysis indicated that long staging delays could occur if an appropriate service standard were not used in designing a one-officer program.

After we called these facts to the attention of the City, an extensive study was commissioned to develop a one-officer program (New York Times October 15, 1982).[3] At the time, there was considerable pressure on the City to do so since, according to many analysts, the contract negotiated with the municipal unions in October 1982 put the New York City budget in a precarious state (*New York Times* October 13, 1982). One-officer patrol had been shown by this analysis to be one of the most attractive productivity programs available to reduce the budget pressures.[4]

[3] As of this writing the City is expecting the study report on how such a program might be implemented.

[4] We are very grateful to Dr. Roken Ahmed for his invaluable assistance in performing the extensive computer computations.

References

BOYDSTUN, J. E., M. SHERRY AND N. P. MOELTER, "Patrol Staffing in San Diego One- or Two-Officer Units," Police Foundation, 1977.

CHAIKEN, J. M. AND P. DORMONT, "A Patrol Car Allocation Model: Background," and "A Patrol Car Allocation Model: Capabilities and Algorithms," *Management Sci.*, 24 (1978), 1280–1300.

—— AND R. C. LARSON, "Methods of Allocating Urban Emergency Units," *Management Sci.*, 19 (1972), 110–130.

CHELST, K. R., "Deployment of One- vs. Two-Officer Patrol Units: A Comparison of Travel Times," *Management Sci.*, 27 (1981), 213–230.

COBHAM, A., "Priority Assignment in Waiting Line Problems," *Oper. Res.*, 2 (1954), 70–76.

GASS, S. AND J. DAWSON, "An Evaluation of Policy Related Research: Reviews and Critical Discussions of Policy Related Research in the Field of Police Protection," *Report of Mathematica*, Princeton, N.J., 1974.

GREEN, L., "A Queueing System in Which Customers Require a Random Number of Servers," *Oper. Res.*, 28 (1980), 1335–1346.

———, "A Multiple Dispatch Queueing Model of Police Patrol Operations," *Management Sci.*, 30 (June 1984).

——— AND P. KOLESAR, "A Comparison of the Multiple Dispatch and $M/M/c$ Priority Queueing Models of Police Patrol," *Management Sci.*, 30 (June 1984).

——— AND ———, "Testing the Validity of a Queueing Model of Police Patrol Operations," Columbia University, Working Paper No. 521A, 1983.

Kansas City, Missouri Police Department, *Response Time Analysis Vol. 2*, 1977.

KAPLAN, E. H., "Evaluating the Effectiveness of One-Officer Versus Two-Officer Patrol Units," *J. Criminal Justice*, 7 (1979), 325–355.

KOLESAR, P., "Ten Years of Research On the Logistics of Emergency Services," in J. P. Brans (Ed.), *Operational Research '81*, North-Holland, New York, 1982.

——— AND E. BLUM, "Square Root Laws for Fire Engine Response Distances," *Management Sci.*, 19 (1973), 1368–1378.

KOLESAR, P. AND W. WALKER, "A Simulation Model of Police Patrol Operations," Report R-1625/2, The Rand Corporation, Santa Monica, Cal., 1974.

LARSON, R. C., *Urban Police Patrol Analysis*, MIT Press, Cambridge, Mass., 1972.

New York Times, "Police and City in a Deadlock Over One-Officer Patrol Cars," October 22, 1981.

———, "1-Officer Police Cars: Safety is Issue," November 2, 1981.

———, "City Fiscal Woes May Delay New Police Class," October 13, 1982.

———, "City Restudying One-Officer Cars for Police Patrol," October 15, 1982.