Improving the New York City Arrest-to-Arraignment System

RICHARD C. LARSON
Queues Enforth Development, Inc.
432 Columbia Street
Cambridge, Massachusetts 02141

and
Operations Research Center
E40-149a
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

MICHAEL F. CAHN
Queues Enforth Development, Inc.

MARTIN C. SHELL
Queues Enforth Development, Inc.

In 1988, New York City’s arrestees were in-custody waiting to be arraigned for an average of 40 hours, occasionally for more than 70. Moreover, they were held in crowded, noisy conditions that were emotionally stressful, unhealthy, and often physically dangerous. The city competitively awarded a contract to Queues Enforth Development, Inc. to address the problems of arraignment delays and the staggering associated costs of operating the arrest-to-arraignment system. The project team mounted an extensive two-year effort that included developing comprehensive, easy-to-use system-performance and cost models for the city’s four major boroughs and using the models to recommend sweeping operational and policy changes. The recommendations implemented since 1990 have saved the city tens of millions of dollars and have helped to lower average arraignment times to 24 hours or less.

New York City has had a long-standing tradition of utilizing the techniques of management science and operations research (MS/OR) in planning and operating many of its complex urban service systems. In the late 1960s, the Rand Corporation of California opened Manhattan offices, operating the New York City Rand Institute until 1975. Many of the Rand Institute’s research products, for instance planning models for the New York City Fire and Police Departments, remain

Copyright © 1993, The Institute of Management Sciences
0091-2102/93/2301/0076$01.25

JUDICIAL/LEGAL—CRIME
QUEUES—SIMULATION

INTERFACES 23: 1 January–February 1993 (pp. 76–96)
NEW YORK CITY

in use today [Chaiken 1971, Kolesar and Blum 1973, and Larson 1971a, 1971b, 1972b, 1974, 1975]. After 1975, individual city departments issued requests for proposals (RFPs) to carry out additional agency-specific management science projects. Among the agencies involved in major management science projects were the Department of Sanitation [Larson, Gregory, and Minkoff 1988], the Department of Transportation [Cahn, Larson, Neitnerman, and Schaack 1987], the Department of Health and Hospitals [Larson 1982], the Department of Environmental Protection [Larson 1988], the Office of Management and Budget (OMB) [Larson and Rich 1987], and the Department of Probation.

Collectively, all of this work has saved the city tens of millions of dollars and improved services to the citizens of New York. The work has been acknowledged in the MS/OR profession, too, with numerous refereed published papers and at least two books deriving from the work. The Operations Research Society of America recognized this early work with two Lanchester Prize awards [Larson 1972a and Kolesar and Walker 1974].

It has been harder to track the MS/OR work performed for the city since 1975 due to the multiple agencies involved at different times, with different individuals and groups doing the studies. The city, encouraged by the success of much of the earlier work, has invested in creating its own in-house MS/OR groups now active in the New York Fire Department, the New York Police Department (NYPD), the Emergency Medical Services of the Department of Health and Hospitals, the Department of Sanitation, the Department of Transportation, and the OMB. The transportation commissioner, Lucius Riccio, has a PhD in operations research.

The work we describe, perhaps the single most ambitious MS/OR project undertaken by the city to date, involves the arrest-to-arraignment (ATA) system of the city. This "system" encompasses multiple city, state, and private agencies and, prior to the study, had virtually no central accountability or coordination. Professionals operating the ATA system are responsible for the activities that follow the arrest of a suspected criminal offender from the point of arrest to the point of first appearance before a magistrate judge, who determines probable cause and sets (or denies) bail. Many spatially dispersed activities must transpire during this process, many in a required time sequenced order. To make matters even more complex, each of the four boroughs studied—the Bronx, Brooklyn, Manhattan, and Queens—had over the years developed its own idiosyncratic ways of operating "its" ATA system, so in effect by the late 1980s the city had four different ATA systems that operated almost independently.

New York City Faces a Criminal Justice Crisis

During the 1980s, the time it took to move a defendant from the point of arrest to her first court appearance before an arraignment judge grew longer. While there had been no statutorily mandated time
limit, it is generally conceded that holding a defendant for more than 24 hours without a court hearing to establish probable cause for the arrest constitutes an unreasonable delay. In 1988, average delays commonly exceeded 40 hours, with a number of defendants waiting more than 72 hours for their arraignments. On March 23, 1990 (before our project’s recommendations were implemented), the New York Times ran a front-page story on a woman who spent 45 hours in pre-arraignment detention in the Bronx under the headline, “Trapped in the terror of New York’s holding pens.” Arrestees were denied a speedy court appearance, and the lengthy delays greatly diminished the efficiency of justice system personnel and facilities.

Cahn et al. [1989a] describe the complexities of the New York City ATA system in exhaustive detail, and Figure 1 shows a generic flow chart of the system as it functioned in 1988. It is “generic” because in reality each of the four major boroughs of New York City (the Bronx, Brooklyn, Manhattan, and Queens) has its own ATA system. Each borough has up to 25 police precincts, as well as its own criminal court, central booking facility and district attorney’s complaint room. The process begins with an arrest by either a New York City Police Department (NYPD) officer or an officer of one of the several other law enforcement agencies with jurisdiction in New York City (Figure 1). Transit and housing authority police department arrests dominate the non-NYPD sources. In the days prior to our study, the arresting officers (A/Os) typically escorted their prisoner(s) to their affiliated police precinct where the prisoner was searched, fingerprinted, and detained while the police completed an arrest report and a variety of other paperwork. The A/O then escorted her prisoner to the borough’s central booking facility, at which point the paths of the A/O, the arrestee, and the paperwork diverged. Central booking serves as a concentration point for holding large numbers of arrestees (up to 100) in a single location and as a centralized facility for screening and processing fingerprints and other arrest-related paperwork.

After being searched again, the arrestee was held at central booking for a bail interview and eventually lodged at the borough’s criminal courthouse (space permitting) or more likely transported to an outlying police precinct to await arraignment. Her fingerprints were faxed to the New York City Police Department in those cases. The arrestee was then transported to the courthouse, where an attorney or defender for the Bronx, Brooklyn, Manhattan, or Queens was available to represent him or her at arraignment.

![Figure 1](image-url)

**Figure 1:** This flow chart depicts the New York City generic arrest-to-arraignment system as it operated in 1988. There have always been differences—some minor and others major—among the boroughs in the way they process arrests. The times in parentheses indicate approximate total elapsed time from the time of arrest.

INTERFACES 23:1 78
York State Division of Criminal Justice Services (DCJS) at the state capitol in Albany for a positive identification and the return of the arrestee's criminal history (rap sheet). In the meantime, the A/O proceeded to the district attorney's complaint room to be interviewed by an assistant district attorney who would draw up the formal sworn complaint document. On the average, it took 18 hours to complete all the paperwork and 39 hours to get the arrestee to the courthouse so that an arraignment could take place—more than 40 hours after the initial arrest event. In fact, approximately 30 percent of the arrestees waited 50 hours or more to be arraigned (Figure 2). During this lengthy period, the arrestees were confined in a series of very noisy, crowded, and often unhealthy or dangerous holding facilities.

The arrest figures amply demonstrate the magnitude of the problem (Table 1). In 1988 in the four major boroughs alone, the total pool of arrests for which the defendants could be detained awaiting arraignment was over 300,000. Of those, only 12 percent were issued desk appearance tickets (DATs) permitting them to leave with an obligation to return to court in 30 days for a scheduled arraignment. The remaining "on-line" arrestees spent an average of about 40 hours in the various holding spaces awaiting a court appearance. The limited use of DATs was rooted in police fears of defendants failing to appear in court and of releasing defendants for whom there might be outstanding warrants. Unlike most jurisdictions, where misdemeanors outnumber felonies, in New York City felonies predominate, many of them involving violence or illicit drugs. Thus, an arrestee might find herself in the same holding cell with violent repeat offenders or one or more defendants in the midst of drug episodes.

Many diverse organizations participate in the ATA system beside the NYPD and the other law enforcement agencies responsible for arrests. These include the New York City Department of Correction which supervises some of the pre-arraignmentees; the offices of the (independently elected) district attorneys in each borough; the New York State Office of Court Ad-

Figure 2: In 1988, the delay between the time a police officer made an arrest and the time the arrestee was arraigned before a magistrate ranged from a few hours to 100.
LARSON, CAHN, SHELL

<table>
<thead>
<tr>
<th>Borough</th>
<th>On-Line Felonies</th>
<th>On-Line Misdemeanors</th>
<th>On-Line Infractions and Violations</th>
<th>Desk Appearance Tickets</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bronx</td>
<td>38,660</td>
<td>16,855</td>
<td>740</td>
<td>5,985</td>
<td>62,240</td>
</tr>
<tr>
<td>Brooklyn</td>
<td>47,650</td>
<td>22,780</td>
<td>1,560</td>
<td>10,185</td>
<td>82,175</td>
</tr>
<tr>
<td>Manhattan</td>
<td>51,795</td>
<td>58,660</td>
<td>3,880</td>
<td>17,275</td>
<td>131,610</td>
</tr>
<tr>
<td>Queens</td>
<td>29,330</td>
<td>15,780</td>
<td>1,440</td>
<td>6,750</td>
<td>53,300</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>167,435</strong></td>
<td><strong>114,075</strong></td>
<td><strong>7,620</strong></td>
<td><strong>40,195</strong></td>
<td><strong>329,325</strong></td>
</tr>
</tbody>
</table>

Table 1: In 1988, New York City arrests for the four major boroughs totaled 329,325. Among the charge classes, on-line arrests refer to those in which the defendant is held pending arraignment. Desk appearance tickets (DATs) are issued to those arrested for lesser offenses; such arrestees are not held in custody but given a date to appear in court for arraignment.

administration (OCA), which furnishes all city court personnel; the State DCJS, the central repository of criminal histories and fingerprints; the Legal Aid Society, which is the principal public defender organization; the Criminal Justice Agency (CJA), which contracts with the city to provide bail interviews; and the various county bar associations to which private criminal attorneys typically belong.

In response to this problem, the city had instituted a variety of procedural changes and constructed a number of new facilities designed to reduce ATA delays and to improve the overall ATA process. Prominent among these had been the institution of central booking, which became operational in all five boroughs by 1978, and the city’s expenditure of millions on arrest-related police overtime (more than $25 million in 1988). Unfortunately, the combination of arrests increasing in numbers and severity (mirroring an illicit drug use epidemic) and facilities that suffered from both age and space limitations had more than offset these attempts at improvement. As a result, the situation continued to deteriorate to the point where in 1985 the Legal Aid Society brought a federal class action law-suit that challenged the legality of detaining pre-arraignment defendants for more than 24 hours. Despite the fact that a 1987 lower court ruling that held that such delays are unconstitutional was eventually overturned on appeal, the problem remained one of crisis proportion. (In 1990, after conclusion of the ATA project, the New York State Supreme Court held that the city was compelled to give the court a satisfactory explanation whenever the arraignment of a defendant exceeded 24 hours. If it could not, the defendant would have to be released.)

Recognizing that “easy fixes” were not working and that the scope of this problem required the addition of outside resources to augment its own substantial justice capabilities, the city decided to retain a consultant to conduct a comprehensive study of the ATA process in the four populous boroughs. (Staten Island did not experience the ATA delays of the other boroughs.) In February 1988, the city released an RFP that defined a three-phased project. It called for (1) an in-depth survey of the ATA process culminating in a detailed descriptive document with accompanying cost implications; (2) an analysis of the
ATA system employing modern technical methods; and (3) the recommendation of policy changes designed to reduce ATA delays and, where possible, reduce the costs of the ATA system to the taxpayers of New York City.

After a competitive procurement process, the city selected Enforth Corporation (now, Queue Enforth Development, Inc.) to conduct the ATA study. The study began in October 1988 under the joint supervision of the Office of the Criminal Justice Coordinator (now, Office of the Deputy Mayor for Public Safety) and the OMB. As the RFP correctly observed, the boroughs of the Bronx, Brooklyn, Manhattan, and Queens differ in their ATA processes. They differ in arrest volumes and patterns, procedures, facilities, and other characteristics. Accordingly, we were asked to consider each borough independently—in essence, to conduct four parallel studies.

**Adopting a Suitable Approach**

The ATA process consists of three interrelated components: moving the defendant from the point of arrest to a precinct, central booking, and finally to the courthouse; producing a criminal history for the arrestee based on a positive fingerprint identification; and swearing out a complaint (Figure 1). Since a delay in any of these three components could delay the completion of the entire arraignment process, we had to develop a complete qualitative and quantitative understanding of each component to model accurately the system as a whole. Confounding factors in this analysis were the number of independent agencies involved in the process in each of the four boroughs and the differences from borough to borough in geography, facility complement, and system policies.

To meet our contractual mandate “to provide a set of recommendations based upon accepted operations research principles,” we had to isolate those issues best addressed through mathematical modeling and determine the most appropriate sources of data. In addition, we needed to develop a qualitative understanding of the issues that underlie the functioning of the entire system. These issues included the statutory requirements for arraignment, the budgetary constraints under which all participating agencies were operating, the historical context that affected the manner in which participating agencies interacted, and the administrative structures of the agencies involved. It was critical that we come to grips with these issues early in the project so that we could identify the range of potential policy alternatives. To accomplish this, in consultation with client project management we developed a three-phased approach, each phase resulting in important project products.

**Phase One: Initial Analysis**

In the first phase of the study, the 12-member project team attempted to become familiar, in exhaustive detail, with the functioning of the ATA systems in each of the four boroughs. Perhaps most crucial to this familiarization was the observation in person of the conduct of all phases of the ATA process. We visited every type of facility in every borough during all three duty shifts and on both weekdays and weekends. The many hours we spent in precincts, central booking, complaint rooms, courthouses, and other venues afforded us an understanding of the microstructures of the system essential to its ac-
curate modeling. On one occasion, we were present at 3:00 AM in a courthouse when a clamor arose in the holding cells because a female defendant was discovered to be smoking crack cocaine. She had undergone searches at the precinct of her arrest and at central booking. On another occasion, we met a Latino defendant at central booking who had just been discovered to have spent 48 hours in the holding cell awaiting a bail interview. His paperwork had been "lost," and speaking no English he was unable to make his custodians aware of his plight. No oral or written description of the ATA process could possibly have substituted for this kind of first-hand knowledge gained by "living in the system."

Concurrently, team members were examining the system infrastructure (for example, facility locations, capacities, conditions, and staffing) and identifying relevant data gathered by existing manual and automated information systems. The product of this phase was a detailed description of the ATA process [Cahn et al. 1989a]. This 400-page document consisted of separate chapters describing the ATA system as it functioned in each of the four boroughs. In addition to a narrative discussion of the process, the chapters included flow charts of that process, and statistical tables and charts focusing on characteristics of the arrests and ensuing arraignments. When the project team first began its work, there had been no single coherent description of the city's ATA system. Our report has filled that void and is often used to orient new justice system employees.

In addition to analyzing the ATA systems in the four boroughs, our project team studied all the support systems, such as the fingerprint identification process, the warrant management system, and the NYPD's on-line booking system (OLBS). Containing records of every city arrest, the OLBS is of critical importance—it is the transaction processing system that records the status and outcome of most of the stages in the ATA process. It served as the primary data source for much of the quantitative information developed during the first project phase.

We gathered information in order to understand the operations and performance of the ATA system. Our analyses focused on the steps necessary to move a defendant from the point of arrest to an arraignment courtroom and to meet the statutory obligations for her arraignment. While excessive delays were a key factor in initiating this study, the high costs associated with the ATA process were an additional motivating factor. After initial familiarization with the structure of the ATA system, part of the project team began allocating costs to each component of the ATA process. We determined the costs of operating the system as configured and began to develop a cost model that would later enable us to estimate the cost consequences of alternative ATA policies and procedures.

Gathering the data necessary to complete this detailed description of the ATA system proved invaluable. Over the first weeks of the study, we met with officials
from all of the agencies involved in the ATA process, who expressed a wide range of perspectives on the ATA system. While each agency was able to furnish a fine-grained description of its own role in the system, no participant offered a systemic overview that accurately represented those parts of the system for which that agency was not directly responsible. As a result, the views of one participating agency were often dismissed by others as lacking objectivity, and cooperative solutions to system problems were difficult to achieve.

As outside consultants, we had the advantage of not being perceived as allied with any particular agency. However, some staff members expressed skepticism over our ability to solve the ATA problems. As a result, we structured the first phase analyses to serve three crucial functions: (1) to provide us with a detailed understanding of the intricacies of the ATA systems; (2) to offer an initial statistical characterization of the performance measures associated with the upcoming modeling effort; and (3) to produce for the participating agencies an objective encyclopedic account of all phases of the ATA system, illustrating the depth of the team’s knowledge. Without this early proof of objectivity and familiarity, we would have had difficulty gaining client acceptance for the mathematical modeling results or credibility for our recommendations.

Approximately half of the time markers indicated in Figure 1 were available in machine-readable format from the OLBS arrest records. By analyzing these data, and developing detailed process flow charts, we were able to come to some preliminary conclusions and obtain direction for the more detailed analyses we performed in the second phase of the study.

**Phase Two: Analysis and Modeling**

In the second phase of the study, we examined critically the products of the first phase, formulated and initiated implementation of a modeling strategy and began to examine alternative system configurations and policies that might merit consideration in the modeling effort.

At the outset, we recognized that the system performance data available from OLBS did not provide a complete enough picture of the operations of the ATA system. We obtained additional data regarding the movement of arrestees within the courthouses from the Office of Court Administration (OCA), and we acquired data on the disposition of cases from the Criminal Justice Agency (CJA). To complete the quantitative picture of the movement of arrestees through the ATA system, we had to encode manually hundreds of pages of barely legible handwritten logs.

Examination of the rich ATA data base now available to us resulted in some important conclusions. On average, the defendant paperwork (sworn complaint and rap sheet) was completed well before the defendant arrived at the courthouse (Figure 1). Thus, in general, delays in identifying fingerprints and delays in generating complaints did not significantly delay arraignments; the majority of arrestees could not be brought to the courthouse until well after their paperwork was ready. It also
became apparent that arresting officers were spending an average of more than eight hours from the time they departed central booking until they swore out the complaint ending their role in the process. While this might not delay the arraignment, paying police officers to wait in line for eight hours or more (much of it on overtime) to perform a 30-minute task was expensive and excessive.

Mathematically, we can state the temporal relationship among the steps in the ATA process as follows:

\[ T_a > \max \{ T_c, T_r, T_d \} \]

where

\[ T_a = \text{elapsed time from arrest to arraignment}, \]
\[ T_c = \text{elapsed time from arrest to swearing of the complaint}, \]
\[ T_r = \text{elapsed time from arrest to return of the "rap sheet," and} \]
\[ T_d = \text{elapsed time from arrest to arrival of the defendant at the courthouse}. \]

There had been constant debate in the ATA community as to whether delays in generating complaints, in producing rap sheets, or in transporting defendants were the primary cause of the excessively lengthy ATA times. The analysis showed no single culprit. Rather, the interactions among the three processes for all defendants in the system created a form of "gridlock." Frequently, among a group of 15 to 20 defendants brought to the courthouse, there would be some for whom complaints and rap sheets were ready while others would have to wait for hours before they could be arraigned. Thus, the effective capacity of the system was greatly reduced. These and other interdependencies among ATA processes and defendants greatly increased the complexity of the system.

In selecting and implementing a modeling strategy, we needed to satisfy a number of criteria. The models had to address both the performance and cost consequences of a wide range of policy and work-load scenarios. Also, the models had to be validated with historical data, so as

---

**The cost of police overtime related to arrests exceeded $25 million.**

---

to be credible. As the models themselves were project deliverables, to remain with the client for continuing use as decision support tools, they had to be accessible to criminal justice planners who had limited technical training. Due to the highly dynamic and stochastic nature of the ATA system, we chose Monte Carlo simulation as the modeling approach. Also, we selected Apple Macintosh™ hardware for its friendliness and user familiarity and the Extend™ simulation software for its relative ease of use. Finally, because of the idiosyncrasies in the ATA process in each of the four boroughs, we developed four separate and distinct models [Cahn, Larson, Shell, and Rich 1989b and 1990].

The simulation model was to serve as a mathematical representation of the entire ATA system in each of the four boroughs that would allow users to experiment with what-if scenarios involving combinations of work loads and arrest processing policies. As we completed our on-site analyses
of the existing ATA systems, team members recognized a number of opportunities for reducing arraignment delays and system costs. These ranged from short-term modifications, such as increasing the number of arraignment courtrooms or adding complaint writing staff at the district attor-

Paying police officers to wait in line for eight hours or more was excessive.

neys’ complaint rooms, to much longer-term modifications involving the construction of new facilities or alterations in processing policy that would entail statutory authorization.

The simulation software allowed us to use a modular simulation approach permitting each of the steps in the ATA process to be represented as a single “block.” The blocks included
—The arrest,
—Precinct processing,
—Central booking,
—Fingerprint identification,
—Complaint generation,
—Precinct lodging,
—Courthouse holding, and
—Arraignment.

On the computer screen, the model looks very much like a flow chart. Blocks represent the stages in the process, while the lines linking the blocks represent the flows of arrestees, paperwork, and information. Underlying each block is a program written in a structured “C-like” language embodying the processing that occurs in that stage in the process as well as the dependencies of that stage on others. Users can examine and modify the parameters relevant to each block by double-clicking on a particular block on the screen, producing a “dialog box” containing user-selectable parameters and options (Figure 3).

The “script,” or program, associated with each block depicts, in as much detail as possible, the functioning of that stage of the process. From the data analyses underlying the project’s first two reports [Cahn et al. 1989a and Cahn, Larson, Shell, and Rich 1989b], we understood the statistical distributions of the system’s key parameters. We embedded these distributions in each block, with user-modifiable parameters. For example, we modeled the arrest rate in each borough as a time-varying Poisson process with individual means for each hour of the day and day of the week. A user can modify any of the arrest rates or model an overall increase or decrease in the arrest rate by specifying an upward or downward “scale factor.” Each block in the model embodies one or more processes modeled stochastically.

We could model some stages of the process as simple single or multiple server queuing systems. Others, however, proved more complex. The “precinct lodging” block, for example, simulates several interdependent processes. This block models the selection of precincts in which to lodge van loads of prisoners, determining the time required to transport them to and from the precinct and the sequence in which to transport groups of defendants back to the courthouse. With up to 15 lodging precincts in each borough, separate facilities for male and female defendants, and the linkage between scheduling transportation to the courthouse and the status

January–February 1993 85
of the defendants' paperwork, this was the most difficult subsystem to model.

We were able to model many of the policy options under consideration simply by altering one or more parameters in the existing model's structure. For example, to eliminate the need to lodge arrestees in precincts before moving them to the courthouse, one sets the courthouse holding capacity sufficiently large so that the model "chooses" to move arrestees directly from central booking to the courthouse. We incorporated other alternatives in the form of optional parameters added to existing blocks. At the time of the project, the city had been exploring the practice of having A/Os write their own complaints at the precinct, thereby eliminating the need for a complaint room interview. This option was built into the precinct processing block and could be applied to a user-specified fraction of all arrests. More complex policy options required the addition of special purpose blocks to the model's structure. One of the policy options that required an additional block was the use of a centralized pre-arraignment mass holding facility. However, the inherently modular structure of the performance model greatly facilitated such changes.

The model generates several types of output. On performance, it generates aggregate performance measures, such as average overall ATA time and the average times required for completing individual stages of the process, as well as time-series graphs of system performance (Figure 4). In addition, the model also outputs a file
Figure 4: Example of the output available from the ATA simulation model. These are average daily values for two key elapsed time statistics: the elapsed time from arrest to generation of a sworn complaint and the elapsed time from arrest to arraignment.

containing system configuration parameters that are available as input to a separate spreadsheet-based cost model. This architecture allows the user to create families of performance and cost measures for each borough under wide ranging work loads and policy alternatives. By implementing a static cost model, we gave justice and financial planners the ability to estimate rapidly the comparative cost consequences of a broad spectrum of system alternatives. Inherent in the spreadsheet cost model (implemented in Microsoft Excel™) are a number of “back-of-the-envelope” queuing models as well as probabilistic models of the elapsed times between stages in the ATA process.

Through repeated use of the model (in production runs), we discovered that by simulating the equivalent of several days of system operation, the oscillatory behavior of the model would tend to stabilize. As a general practice, we ran production runs that simulated at least four weeks of system operations to reduce run-to-run variability and to ensure that observed differences among alternative scenarios were attributable to differences in the underlying policy assumptions rather than to statistical variability.

Because of the complexity, nonlinearity, stochasticity, and nonsteady state nature of the ATA system, we had to employ a Monte Carlo simulation. However, we buttressed this approach with “back-of-the-envelope” models so that the totality of our management science approach embodied both Monte Carlo simulation and analytical modeling.

The models were developed to serve two primary purposes: to help us to compare work load and policy options during the study period; and to be retained by the clients and used as a flexible, intuitive de-
cision support tool to help the city’s criminal justice planners well into the future.

During the study period, we used the model to estimate system performance and cost under a wide variety of promising scenarios which evolved from the extensive on-site and statistical analyses we conducted during the first two phases of the study. These scenarios varied in arrest volume, staffing levels at each stage of the process, physical capacities of holding areas at central booking, and numbers of lodging cells in precincts and in the courthouse. Some scenarios incorporated new facilities or new procedures. Toward the end of the study, we presented the models to the client organizations whose employees had been trained in their use. Criminal justice planners in the city can now propose new facilities or procedures in the arrest-to-arraignment process and quickly and easily determine their cost and potential performance.

Phase Three: Findings and Recommendations

When it initiated the project in late 1988, New York City had been experiencing lengthy ATA delays for more than a decade. While many planners and managers in participating agencies had been involved in the process for years, there was little consensus as to the sources of delay. Virtually every agency and every phase of the process was held responsible for the delays by at least one of the other system participants. Delays in identifying fingerprints, generating complaints, or transporting arrestees, “slow” arraignment judges, and many others factors were blamed. As outside management science practitioners, we were able to draw a more informed and objective picture of the system.

One of our more surprising observations was the pattern of arrests over a typical week in 1988. One might assume that the occurrence of crimes would peak on weekends; in fact, the weekly arrest peak occurred more often in the middle of the week, on Wednesday or Thursday. We discovered that, to some extent, arrests were discretionary events. Related criminal acts exist, but at certain times of the week police officers appeared less inclined to make arrests. We believed this phenomenon to be an artifact of the delays in the ATA process. During busy periods, an officer could be occupied with a single arrest for 24 hours or more. This discouraged them from making arrests close to or on the weekends, despite the overtime pay premium.

A second important discovery was the identification of the “critical path” in the ATA process. On average, fingerprints were identified and complaints generated well before defendants were brought to the courthouse (Figure 1). While these processes varied, most arraignments were held up by delays in moving the defendant rather than delays to the paperwork. An “obvious” solution appeared to lie in increasing courthouse throughput. If more arrestees could be arraigned each hour, courthouse holding space would free up more readily, and more arrestees could be moved to the courthouse. However, when this was attempted on an ad hoc basis, it accomplished very little in the absence of current and reliable information on defendant and paperwork status.

To take advantage of increased throughput in the courthouse, the system had to
move arrestees through the process at an accelerated pace, giving preference to those whose paperwork was ready ahead of those whose paperwork was not. Each arrestee who was moved to the courthouse holding area before her paperwork was ready reduced its effective holding capacity. If this happened to a great extent, “gridlock” often ensued, with arrestees ready to be arraigned unable to move to the courthouse, and those in the courthouse not ready to be arraigned clogging the system. This phenomenon contributed to some of the worst delays we observed.

During the early stages of the study, members of our project team observed a complaint room queue of more than 150 police officers waiting either to be interviewed or for a complaint to be typed. Many of these officers were being paid at overtime rates, and none of them was available for her or his regular assignment. (Institutional and logistical barriers precluded the scheduling of interviews as a solution to the queuing problem.) It became clear that the marginal cost of increasing the number of “servers” (that is, complaint writers and typists) would be modest compared to prospective savings in police overtime costs from shortening the complaint room queue. Reducing the time A/Os took to process an arrest appeared to have several positive potential effects. In addition to reducing the cost of processing arrests, it would also minimize the likelihood that arraignments would be delayed by the lack of a sworn complaint. It would also tend to remove some of the disincentives influencing the observed time patterns of arrests.

Individuals charged with most categories of misdemeanors, or with lesser (“Class E”) felonies were eligible for DATs. Instead of being held in custody after arrest, they would be taken to a police precinct and fingerprinted, then given a document requiring them to appear for arraignment at a pre-specified date and time. These individuals could be scheduled to appear in court during relatively low-volume time periods. However, the NYPD exercised this option infrequently, giving DATs to a small fraction of the misdemeanor arrestees and to virtually none of the lesser felony arrestees. While surprising, the limited use of this option was attributable to a number of factors. For one, it was not always possible to verify that an arrestee had no outstanding warrants, inasmuch as the police could not make a positive fingerprint identification prior to release. Also, the police feared that some of these individuals would not appear in court on their scheduled arraignment dates. In some cases, officers made arrests because of the long delays, knowing that the time waiting for arraignment might be the only time someone accused of a minor offense would be likely to spend behind bars.

In general, we observed that the system in each borough was a hybrid of centralized and decentralized elements. While booking and the courthouse were centralized, the lodging of defendants at police precincts and the initial arrest processing at the precinct of arrest were decentralized operations. This mixture presented obsta-
cles to both tracking and accountability.

After analyzing the data and modeling the ATA system, we proposed a series of recommendations [Cahn, Larson, Shell, and Rich 1990] to the city. These were divided into three categories: (1) modifications essential to the baseline system, (2) short-term, high-payoff modifications in operations, and (3) longer-term high-payoff modifications to facilities and operations.

**Essential Baseline Modifications**

Throughout our analyses of the ATA system we were struck by the fact that a number of independent agencies ran the ATA process with no central administrative control and no common base of system information. While the ATA simulation model gave the city a tool for making long-term strategic decisions, it lacked its tactical analog: a real-time system providing detailed status updates and incorporating the tools to help it deploy its resources. Such a system would also serve as a repository for the data needed to update the strategic planning models as work loads and policies shifted.

In addition to the tactical planning tool, the city clearly needed an ATA administrative structure. We recommended that the city appoint a city-wide ATA director (reporting to the deputy mayor for public safety) and four borough ATA directors. This group's most important function would be to coordinate the activities of the multitude of agencies that used the strategic and tactical planning tools.

With these systems in place, the city could implement a number of general "system-tightening" initiatives. These are low-cost improvements in such areas as inter-agency coordination, transportation and other logistics, and fingerprint quality.

**Short-Term, High-Payoff Operational Modifications**

The simplest way to reduce the delays in a multiple-queue system such as the ATA system is to reduce the number of "customers." By issuing DATs to a greater fraction of those arrested for misdemeanors and lesser felonies, the city would reduce the system work load.

Once the city had implemented the information and administrative infrastructures, it could make two system improvements that promised to reduce delays and costs. First, we recommended increases in both complaint room staff and the number of arraignment courtrooms. Increasing staff in the complaint rooms was primarily a cost-reduction measure. The savings in A/O overtime would more than cover the cost of the additional staff and the cost of scheduling additional arraignment courtrooms. Additional courtrooms would increase the courthouse throughput and streamline the movement of defendants through the system.

Alternatively, we recommended that the city consider eliminating the complaint room interview entirely. If the A/O could generate the criminal complaint herself, with review by the district attorney's office, the savings in police officer overtime and for personnel costs in the district attorney's office would be substantial. In tandem, the use of "smart" complaint generation computer software would greatly assist police officers in formulating and typing the complaints.

**Longer-Term High-Payoff Modifications**

The two groups of short-term recom-
mendations we have described required neither the construction of new facilities nor the modification of any legal procedures or statutes. In the long term, however, the city may have to construct new facilities or modify existing statutes to improve the efficiency of the system. As we write, the city is expanding the holding spaces in both the Brooklyn and Manhattan courthouses and has recently completed rehabilitating the Bronx courthouse central booking facility and holding space.

One possible operational improvement that would require a legal change would be the implementation of an “expedited arraignment” system for misdemeanor cases. Individuals arrested for misdemeanors for whom no outstanding warrants are found at the initial precinct warrant check could be taken to a separate location that would serve as an “express line” arraignment facility. This would require the waiver of the requirement for fingerprint-based identification. Defendants could be moved directly from the precinct to a centralized expedited arraignment facility where a complaint would be drawn and the defendant arraigned. This would remove these defendants from the system very quickly and reduce its load. Defendants whose arraignments could not be expedited would experience much shorter delays.

A second improvement in this category would be to centralize the lodging of arrestees in each borough. Constructing a single “mass holding facility” in each borough would greatly reduce costs and delays. Currently, a single borough may maintain a decentralized holding system consisting of as many as 20 separate facilities. Indeed, most or all of the arrest processing functions could be incorporated in a single location. Holding space, complaint generation, fingerprint transmission, even the arraignment courtroom itself could be co-located, thereby reducing costs, increasing system efficiency, and greatly simplifying the logistics of the system.

The Project Serves as a Springboard to System Improvements

On May 30, 1990, New York City’s Mayor David N. Dinkins released the findings of the ATA project at a city hall press conference. He made the following public statement:

The Enforth report contains several cost-effective proposals that can significantly decrease the amount of time that it takes to complete the arraignment process. This alone is good news. But the news gets better. The report also contains recommendations that will result in a savings of millions of dollars to the city and will allow us to free our police officers from the process sooner. This will enable these officers to get back to the business of protecting our citizens. Adoption of these proposals will also improve the fairness of our system.

The mayor went on to charge his deputy mayor for public safety, the Honorable Milton H. Mollen, with responsibility for implementing a number of the project’s recommendations and earmarked $3.8 million in his fiscal year 1991 executive budget for that purpose. The mayor set in motion that day a series of policy changes that have overhauled the city’s ATA system. Its performance has improved, and cost savings have been dramatic.

From a management perspective, the
most important recommendation the city implemented was to establish an ATA administrative infrastructure. Almost immediately, the city hired a citywide ATA director, former Legal Aid attorney Martin Murphy. He then hired ATA borough directors responsible for the ATA systems in the Bronx, Brooklyn, Manhattan, and Queens. Murphy, who reports directly to the deputy mayor, and his staff have be-

Arrestees are no longer “warehoused” under hellish conditions for longer than necessary.

come the central strategic planning arm of the ATA system and the monitors of its day-to-day operations. In addition, they mediate the “finger pointing” interagency disputes that had hitherto crippled the system’s ability to respond rapidly to short-term emergencies. For example, both the NYPD and the city Department of Correction are responsible for supervising arrestees held prior to arraignment. In the past, it had been difficult to resolve conflicts between the two agencies about transferring custody of defendants from one agency to the other. Such conflicts often further delayed an already lengthy pre-arraignment process. The ATA directors are neutral parties to such disputes and have been able to mediate them successfully and to keep the process moving. Finally, Murphy’s organization has been successful in coordinating the implementation of major ATA planning initiatives, such as the rehabilitation or construction of facilities.

Managers need sound management tools. The ATA performance and cost models gave the ATA system precisely the strategic planning tools it had so sorely lacked. Decision makers can now understand the sources of the arraignment delays that had for so long eluded them, and they can ascertain what impact changes in operations or policies will have on both arraignment delays and system costs. This what-if system planning capability bore fruit immediately. The project team had recommended that the district attorneys add staff to their complaint rooms, to shorten the A/O queues and reduce police overtime costs. Using the performance and cost models, OMB undertook a series of production runs which convinced them that the projected savings were realistic and aided them in calibrating the amount of money to be invested in new personnel for the district attorneys. According to OMB, in 1988 the cost to the city of police overtime related to arrests exceeded $25 million. The city’s current annual savings in this category alone are approximately $10 million.

One might think of the new ATA administrative structure and the strategic planning models as two legs of a three-legged management stool, where the third leg would be a tactical planning tool to assist in real-time decision making. Our project team recommended the design and implementation of a resource management system that would incorporate a defendant-tracking component and personnel and other resource deployment tools. The city recognized the need for such a system, but an emerging fiscal crisis restricted its budget flexibility. As a result, it agreed to
fund the tracking-system component as an
add-on to the NYPD’s existing OLBS and
defered the more research-oriented develop-
ment of the resource deployment tools.
To their credit, the police management in-
f ormation system staff have provided ATA
 system managers with the ability to iden-
tify the pre-arraignment holding locations
of arrestees as well as the status of their
arraignment-related paperwork (rap sheets
and complaints). This has aided the sys-
tem, and it has also enabled arrestees’ fam-
ily members and friends to locate them
and to determine where and when their
arraignments are likely to take place. Pri-
 vate attorneys can locate their clients and
arrange to be present at their arraignments.
Further, the special personal needs of vic-
tims, witnesses, and A/ Os can be better
met through the timely status information
and scheduling flexibility the tracking sys-
tem affords.

The city has implemented a number of
the project team’s recommendations for
operational changes. The one that has had
the most dramatic impact on ATA delays is
that of making greater use of DATs by is-
suing them more frequently and by using
them for both misdemeanors and for lesser
felonies. Each arrestee receiving a DAT in
lieu of an on-line arrest is released from
custody prior to arraignment and is
thereby removed from virtually all of the
pre-arraignment queues. Even the DAT-re-
lated paperwork is handled on a low-prior-
ity basis. Currently, about 20 percent of all
arraignments are receiving DATs as opposed
to 12 percent in 1988, and approximately half
of all those arrested for misdemeanor of-
fenses are being given DATs. By imple-
menting this and other recommendations,
the city has reduced average monthly ATA
delays that routinely exceeded 40 hours in
1988 to about 24 hours citywide. Arrestees
gain the right to a speedy trial and are no
longer “warehoused” under hellish condi-
tions for longer than absolutely necessary.
The city has greatly reduced its costs for
prisoner supervision and transportation
(the police department has been able to
reallocate $11 million annually) and seldom
has to release an arrestee because the
prosecutor cannot give the court a satisfac-
tory explanation why a defendant had not
been arraigned within 24 hours.

We recommended an “expedited” ar-
raignment procedure that would, if imple-
mented, have reduced arraignment delays
even further. However, under this proce-
dure some defendants would be arraigned
prior to a positive identification and receipt
of their rap sheets. Although some very
large jurisdictions conduct arraignments
without a print-based identification (most
notably, Los Angeles County), the city was
unwilling to undertake quite so radical a
change in procedure. In our opinion, a
substantial increase in arrest volumes, such
as the city’s newly adopted community po-
licing program might produce, could force
the city to reexamine its stance on expe-
dited arraignments—especially if some of
the recommended and planned expansions
of the courthouse holding spaces are not
yet completed.

During the project, several of the bor-
oughs were looking into new technologies
as a way to expedite the ATA process and
to reduce costs. For example, the borough
of Queens experimented with the idea of
having the police themselves draft criminal
complaints (with the assistance of a com-
puter-based system), thereby eliminating the complaint room queue entirely. Although there was skepticism about the efficacy of this approach, we encouraged city-wide consideration of this strategy, and it is now in operation in both Queens and Brooklyn. Similarly, the Manhattan district attorney had been pilot testing a closed circuit video system that allowed the assistant district attorneys to interview A/Os at the precinct of arrest. With our encouragement, this system has been expanded in Manhattan and is currently operational in Brooklyn, as well. Generally speaking, the project accelerated the pace at which the ATA system is exploring technologically based solutions to its problems of congestion and high cost.

Now that the arraignment work load has been reduced and smoothed, the past practice of adding fully staffed arraignment courtrooms has been eliminated and the numbers of regularly scheduled courtrooms dedicated to arraignments have been reduced. Since each courtroom operates for eight hours and must be staffed by a judge, a clerk and support clerical and security personnel, Legal Aid Society public defenders, district attorneys, and police security officers, the personnel cost savings to the city and to New York State are quite substantial. Our cost model calculates the annual savings from eliminating a single five-day-a-week, eight-hour arraignment courtroom to be $1.7 million and $2.4 million for a courtroom operating on all seven days. Based on these figures, we project annual savings of $3.5 million for the city and $6 million for the state.

In Retrospect

We commend the decision makers responsible for the world’s largest criminal justice system for their willingness to acknowledge their limitations and to reach outside the city for help. While New York City is rich in talented managers and planners, there are still times when problem size and complexity defy traditional one-dimensional attempts at solution. The city also recognized that it would require substantial management science resources to get at the root causes of the ATA systems’ problems and committed the necessary multi-year funding (over a million dollars, in total). It was a courageous decision in the face of substantial political resistance. Several prominent justice officials objected to funding the project on the grounds that “I know what the problems are; just give me the money and I’ll fix them.” However, once the commitment was made, even the most resistant and skeptical managers gave us their fullest cooperation—our team could not have asked for greater support.

We also commend the city, especially the office of the deputy mayor, OMB, and NYPD, for being so receptive to the management science methods and models employed by the project team. “Radical” concepts, such as the nonlinear explosive growth of queuing times close to system saturation or Monte Carol simulation, could not have been more alien to the experiences of many of the justice practitioners. However, they were patient and open to new ideas. One of the project’s fin-
NEW YORK CITY

est moments occurred as we sat 30 justice practitioners (most of whom had never used any computer before) at their own Macintosh computers and trained them in the use of the ATA simulation model.

The city went about implementing many of our recommendations immediately. Among them were those recommendations that entailed straightforward administrative or procedural changes, those that would for a modest outlay produce major cost savings and finally, those that would greatly reduce the lengthy ATA delays. Given the prevailing fiscal crisis, the city has deferred some of the more costly and longer term recommendations until additional resources become available. However, they have conducted feasibility studies of several of our major recommendations and are putting plans in place for future implementation.

References


Chaiken, Jan M. 1971, “The number of emergency units busy at alarms which require multiple servers,” Report R-531, Rand Corporation, Santa Monica, California.


Larson, Richard C. 1971a, “Measuring the response patterns of New York City police patrol cars,” New York City Rand Institute, R-673-NYC/1HUD.


Larson, Richard C. and Rich, T. F. 1987,

Robert N. Kaye, Deputy District Attorney, District Attorney of Kings County, Municipal Building, Brooklyn, New York 11201, writes "The Arrest-to-Arraignment System Study performed by Enforth Corporation under contract to the Mayor's Office of the Criminal Justice Coordinator was enormously valuable to both the City and State of New York. Implementation of a number of the study's recommendations has helped to bring about a dramatic reduction in arraignment delays, has saved the City and the State of New York millions of dollars, and has greatly improved the system's fairness. In addition, Enforth delivered an excellent set of decision support models to the City at the conclusion of the study.

"As Assistant Criminal Justice Coordinator to Mayor Edward I. Koch, I supervised and worked with the Enforth project team for almost two years. In my current prosecutorial capacity, I am immersed in the daily operations of Brooklyn's arrest-to-arraignment system and well aware of related activities in the other boroughs."