UNC SCHOOL OF LAW

NORTH CAROLINA LAW REVIEW

Volume 60 | Number 1 Article 9

10-1-1981

Predicting and Reducing Court-Case Time through Simple Logic

Stuart S. Nagel

Recommended Citation

Stuart S. Nagel, *Predicting and Reducing Court-Case Time through Simple Logic*, 60 N.C. L. Rev. 103 (1981).
Available at: http://scholarship.law.unc.edu/nclr/vol60/iss1/9

ESSAY

PREDICTING AND REDUCING COURT-CASE TIME THROUGH SIMPLE LOGIC

STUART S. NAGEL†

Recent state speedy trial acts, such as chapter 15A, section 701 of the North Carolina General Statutes,†† *have increased greatly the pressure on an already overburdened court system. One possible approach to this problem is the use of statistical techniques to schedule criminal and civil cases to reduce the amount of time consumed in processing the case from complaint or indictment to its final disposition. In this Essay, Professor Nagel examines and critiques several of these methods for statistical prediction of time consumption and backlogs of court cases and suggests that further research is needed to implement the ideas generated by such analysis. Perhaps the most important aspect of Professor Nagel's essay is not the equations that he develops but the overall concept of using mathematics, together with the deductive process, as a tool for dealing with court delay and the need for greater judicial efficiency.*

A substantial quantity of literature has been published praising the potential relevance of queueing theory to predicting and reducing time consumption in the courts.[1] That kind of theory has been helpful in both predicting and in reducing time consumption in many other situations, such as tollgates, telephone exchanges and manufacturing processes.[2] There is a need for better prediction and planning in the processing of court cases, and also a need for reducing the amount of time consumed from the filing of complaints in civil and criminal cases to their final dispositions. There may, however, be some

† Professor of Political Science, University of Illinois at Urbana-Champaign. B.A. 1957, J.D. 1958, Ph.D. Political Science, 1961, Northwestern University; member, Illinois Bar. The author wishes to thank the following people for their helpful comments on an earlier verison of this Article: John Reed of the Department of Administrative Science at Clarion State College, Thomas Saaty of the Department of Operations Research at the University of Pittsburgh, and Hans Zeisel of the University of Chicago Law School.

†† N.C. Gen. Stat. § 15A-701 (Interim Supp. 1980).

[1] Literature praising the potential relevance of queueing theory to the courts includes: H. Bohigian, The Foundations and Mathematical Models of Operations Research with Extensions to the Criminal Justice System (1971); D. Greenberg, Mathematical Criminology (1979); J. Reed, The Application of Operations Research to Court Delay (1973); Nagel & Neef, Time-Oriented Models and the Legal Process: Reducing Delay and Forecasting the Future, 1978 Wash. U.L.Q. 467; Nagel, Neef & Munshaw, Bringing Management Science to the Courts to Reduce Delay, 62 Judicature 128 (1978); and Reed & Slivka, Operations Research and the Courts, in Modeling the Criminal Justice System 159 (S. Nagel ed. 1977).

[2] On queueing theory in general, see D. Gross & C. Harris, Fundamentals of Queueing Theory (1974) [hereinafter cited as Gross]; A. Lee, Applied Queueing Theory (1961); and S. Richmond, Operations Research for Management Decisions 405-438 (1968); and T. Saaty, Elements of Queueing Theory: With Applications (1961).

question as to the applicability of management science methods like queueing theory to aid in resolving those judicial process problems.[3]

I. Basic Purposes and Concepts

A. Basic Purposes

A key purpose of this Essay is to analyze how one might go about predicting time consumption and backlogs for court cases or administrative matters. A second key purpose is to contrast statistical prediction with deductive prediction, and then to contrast within deductive prediction what might be called common-sense deduction versus queueing-theory deduction.

One purpose in predicting the amount of time consumed by court cases or other government cases is to allow courts to give priority to cases that are predicted to consume less processing time in order to reduce the average total time of the set of cases being processed, somewhat like an express line in a supermarket. One has to be able to predict how much time each case will consume before the cases can be assigned to different waiting lines.[4]

A second way in which time prediction can be useful is that by learning which variables are good predictors of time consumption, this may tell us something about where to concentrate our resources in order to reduce total time consumption. If, for example, we can develop a model for predicting time consumption from various variables, then we can possibly do some simulation work, and by manipulating some of those variables, predict how much time can possibly be saved.[5]

A third purpose to which time prediction can be put is to enable lawyers to reach better decisions on whether to accept an out-of-court settlement or to go to trial. Knowing how long a wait one will have before trial can be useful in discounting the value of the predicted damage-award. A distant damage award of $5,000 may be worth less than an immediate offer of $4,000. That kind of prediction use, however, will not be discussed in this Essay because it is not as related to the policy problem of time reduction with which we are primarily concerned. Lawyers already know how long the average case takes to come to trial in the courts in which they operate. For their bargaining purposes, they do not generally need to know how much trial time is consumed by various types of cases, or the relation between various judicial procedures and delay reduction.[6]

Why is it necessary to reduce time consumption in the first place? One answer might emphasize the harm that delay causes, in the context of both

[3] Some court researchers have questioned the applicability of queueing theory to the courts, but not in a very systematic manner. Flanders, Modeling Court Delay, 2 Law & Policy Q. 305 (1980); and Johnson, Analytic Tools from Other Fields, in The Use/Nonuse/Misuse of Applied Social Research in the Courts 41-43 (M. Saks & C. Baron eds. 1980). See note 33 infra.

[4] On priority sequencing to reduce delay, see Reducing Delay, supra note 1, at 474.

[5] On reducing delay by reducing trial time, increasing settlements, or having more judge-time, see H. Zeisel, H. Kalven & B. Buchholz, Delay in the Court (1959) [hereinafter cited as Zeisel].

[6] On time discounting applied to out-of-court settlements, see S. Nagel & M. Neef, Decision Theory and the Legal Process 143 (1979).

civil and criminal cases. In civil cases, delay adversely affects clients who must wait to be compensated for their injuries. It also increases the likelihood that witnesses will forget, disappear, or die before a case reaches trial. Delay may also cause defendants to be held in jail an unreasonable period pending trial. In criminal cases, an innocent defendant may become quite vulnerable to pleading guilty in return for the prosecutor's agreeing to recommend probation or a sentence equal to the time already served in jail awaiting trial. If the defendant is not jailed while awaiting trial, long delay may greatly increase the likelihood that he will fail to appear at the trial or commit further crimes while released.[7] In some situations, however, it is not necessarily desirable to reduce time consumption. The costs may be too high in terms of rushing a case to trial before one side or the other has an adequate time to prepare. The costs incurred in hiring the necessary additional personnel may in fact not be justified by the savings in time.[8]

It can be demonstrated that it is physically possible to reduce time consumption greatly by showing that the average amount of time to process cases (1) was substantially less in the past, (2) is substantially less in other courts or agencies, or (3) is substantially less in some cases than in others within the same court at a given point in time. This is so even after one makes adjustments for differences in the complexity of the cases, as measured by such things as the number of witnesses, exhibits, or transcript pages. That kind of capability analysis is, however, generally not as useful as attempting to show that (1) as time consumption goes up, certain delay costs go up; (2) as time consumption goes down, certain speedup costs go up; and (3) the total costs (consisting of the delay costs and the speedup costs) bottom out at a figure substantially less than the prevailing amount of time consumption.[9]

The important point for our purposes is that regardless of whether cases are moving too fast or too slowly, it is helpful to policy makers to be able to predict how much time various cases will consume in order to categorize them for processing purposes, and in order to generate policies as to what variables should be changed in order to change the amount of time consumption.

B. Basic Concepts

To simplify the discussion, it would be helpful to define a set of concepts and symbols that we will frequently use. The main ones are as follows:

T = time consumption measured in days or parts of days.

T_w = time consumed waiting to have a case processed.

T_p = time consumed by a case while it is being processed.

[7] On the serious nature of the delay problem in the legal process, see H. James, Crisis in the Courts (1971); and The Courts, the Public, and the Law Explosion (H. Jones ed. 1965).

[8] An example of waste that comes from trying to reduce court delay is time wasted by potential jurors who come to court in case they might be needed. See Merrill & Schrage, Efficient Use of Jurors: A Field Study and Simulation Model of a Court System, 1969 Wash. U.L.Q. 151.

[9] On finding an optimum time to consume, see Nagel & Neef, supra note 1, at 490; Nagel, Measuring Unnecessary Delay in Administrative Proceedings: The Actual Versus the Predicted, 3 Policy Sci. 81 (1972).

T_t = total time consumed in both waiting and processing.

T_a = arrival time or how often a new case arrives.

$\bar{T}$ = average time consumption for a set of cases rather than the time consumption for a given case (and likewise with $\bar{T}_w$, $\bar{T}_p$, and $\bar{T}_t$) (pronounced "bar T sub-w").

T'_w = sum of the T_w's across all the cases in a set of cases, which depends on the order in which the cases are heard.

T'_p = sum of the T_p's for a given case or an average case, covering all the processing stages to which the case is subject, not just the trial stage.

N = number of cases at some stage:

N_r = number of cases remaining at the beginning of the time period.

N_a = number of cases arriving during the time period up to the present time.

N_p = number of cases processed during the time period up to the present time.

N_b = number of cases in the backlog waiting to be processed as of a given point in time.

N'_b = number of cases waiting to be processed and currently being processed.

$\bar{N}_b$ = average number of cases in the backlog over many points in time (and likewise with $\bar{N}_r$, $\bar{N}_a$, $\bar{N}_p$, and $\bar{N}'_b$).

R = rate at which cases arrive or are processed per average day:

R_a = arrival rate or the number of cases arriving per average working day.

R_p = processing rate or the number of cases processed per average working day.

R_R = ratio of the arrival rate divided by the processing rate.

$\bar{R}$ = average rate or ratio over a large number of days, rather than the rate or ratio for a specific subset of days.

Statistical Concepts:

X = score on a variable being predicted from.

Y = score an a variable being predicted to.

a = value of Y when all the predictor variables are scored zero.

b = change in Y when X changes one unit.

Other Concepts:

D = working days passed from the beginning of the time period to the present time.

C = percent of a judge-day spent in case-trying time.

J = number of judges or other sets of processors available to process cases.

S = percent of cases in the waiting line that are settled or that otherwise drop out before they come to the head of the waiting line.

$\star$ = optimum or desired value of the variable to which the star is attached, such at $T^{\star}_{w}$.

\$ = dollars spent to improve S, N_b, T_p, C, J, R_a, or R_p in order to reduce waiting time or total time.

In view of the purposes of this Essay, the key variable to predict is T_t or the total time consumed in both waiting and processing. The next most important variable to predict is N_b, the number of cases in the backlog waiting to be processed. The rest of the Essay is concerned with three methods for predicting total time, backlogs, and related variables. The methods are (1) predicting through inductive statistical analysis, (2) predicting through algebraic deduction from processing time (T_p) and backlogs (N_b), which is referred to as the common-sense approach, and (3) predicting through algebraic deduction from processing rates (R_p) and arrival rates (R_a), which is the queueing theory approach. The Essay concludes with a discussion of how the methods can be applied to delay reduction, a comparative evaluation of the methods, and a call for further research.

II Predicting Through Inductive Statistical Analysis

There are various ways of classifying predictive schemes. A particularly useful way in light of the above mentioned purposes is in terms of statistical prediction and deductive prediction. Statistical prediction involves obtaining data on numerous cases and determining the relations between the amount of time they consume and their other characteristics. For example, if we want to classify cases or rank-order them in terms of predicted time-consumption so that we could hear the shorter cases first, a simple way of doing so for criminal cases might be to determine the average processing-time for each set of cases involving the same criminal charge. Thus, we might find that the average trial for a shoplifting case takes an eighth of a working day, or approximately one hour, and the average trial for a burglary case takes half a working day, or about four hours. We could do the same thing with personal injury cases, which are the most common civil cases that go to trial. We might find that rear-end collisions average a day of trial time, and intersection collisions with automobiles coming at right angles average two days to try.[10]

A more sophisticated form of statistical prediction would involve working with many predictor variables, not just the nature of the crime. Some of the variables may be capable of interval measurement, such as the maximum sentence possible given the charges, rather than just being positioned in nominal categories as is done when the cases are classified by the crime charged. Good predictors of time consumption in criminal cases might include (1) the severity of the crime as measured by the maximum sentence, (2) whether the defendant has court-appointed counsel or hired counsel, and (3) the number of witnesses that each side has indicated are likely to testify. To develop a statistical-predictive equation, one can obtain information on a large set of past cases showing how much processing time each case consumed (Y) and how

[10] For data which can be used to determine the average length of federal civil and criminal cases by nature of suit or offense for the 12-month period ending June 30, 1980, see Administrative Office of the United States Courts, 1980 Annual Report of the Director.

each case scored on each of those three predictor variables (X_1, X_2, and X_3). That information can then be entered into a computer along with a prepackaged statistical prediction program. The computer will then produce an equation of the form:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3. \qquad (0)$$

The numerical values of b_1, b_2, and b_3 show the relative predictive power of each variable. More specifically, b_1 shows how much time consumption goes up in portions of a day when X_1 or the maximum sentence goes up one unit or one month. The numerical value of "a" shows how much time would be consumed if each variable were scored zero.[11]

To use Equation 0 to predict time consumption for a given case, one (1) inserts into the equation the X_1, X_2, and X_3 score of the case being predicted, (2) does the multiplication and addition, and (3) reads off the value of Y, the predicted time-consumption. As with almost any prediction, the results will not be perfect. More specifically, this prediction method minimizes the squared deviations between the predicted time scores and the actual time scores, but it does not eliminate the deviations. Adding more predictor variables and considering more sophisticated non-linear relations may further reduce the deviations. In this analysis, Y or the variable being predicted was considered to be processing time (T_p). We could have used the same approach in order to predict waiting time (T_w) or total time (T_t).

This inductive statistical analysis approach may also be used to predict the number of cases at some stage of the judicial process. For example, one can obtain data from a large number of court systems or the same court system at many points in time showing the number of personal injury cases waiting to be tried as of the beginning of the year, symbolized N_r. One can also obtain data on (1) the population of the city, (2) the percent of people who ride mass transit to work rather than automobiles, and (3) the number of judges available to process cases. If that information were entered into a computer, the computer could produce an equation in the same form as the one previously described for predicting time consumption. The Y variable being predicted here, though, would be N_r, rather than T_p. The variable $\bar{T}_p$, the average processing time, could be used as a fourth predictor variable for predicting the January 1 backlogs, and likewise the backlogs could be used as a variable to predict time consumption. A similar analysis could be used to predict the number of cases arriving or processed between January 1 and any given date (N_a or N_p), or the backlog waiting to be processed as of any given date other than January 1 (N_b).

One especially useful aspect of statistical analysis is that it can be used to test and improve upon deductive models that are designed to capture the essence of reality. For example, one assumption of queueing theory is that the arrival rate and the processing rate for a given set of cases are independent of

[11] On prediction through the use of statistical equations,.see H. Blalock, Social Statistics 361-472 (1972); and N. Nie, C. Hull, J. Jenkins, K. Steinbrenner & D. Bent, SPSS: Statistical Package for the Social Sciences 320-367 (2d ed. 1975).

each other, as the basic exogenous or outside variables determining total time and backlogs. In reality, they may be strongly related to each other, meaning that when the arrival rate (R_a) goes up, the judges (or other processors) tend to increase the processing rate, so as to keep the backlog from getting too large. Likewise, when the arrival rate goes down, the judges may tend to relax the processing rate. This phenomenon is closely related to Parkinson's theory that time consumed tends to be adjusted to fit the amount of work to be done. An alternative set of assumptions might be that, when the processing rate (R_p) goes up, then the arrival rate also goes up, because people who otherwise might not sue due to slow processing are now more willing to do so. Likewise, when the processing rate goes down, the arrival rate may also go down, because people do not want to wait so long in line. Thus time savings that would otherwise occur from increasing the processing rate may be dissipated by an increase in the arrival rate, and increased time consumption that would otherwise occur from reducing the processing rate may not occur because of a corresponding reduction in the arrival rate. There are variations on the basic statistical Equation 0 given above in which (1) R_p can be the variable to predict to, with R_a being the variable to predict from, and/or (2) R_a can be the variable to predict to, with R_p being the variable to predict from. Those variations include such statistical procedures as cross-lagged analysis and two-stage least squares, which are discussed in statistics textbooks dealing with causal analysis.[12]

Along related lines, statistical analysis can also test for feedback loops between the alleged effects and the alleged causes. For example, another assumption of queueing theory is that time consumption and backlogs are caused by the nature of the arrival and processing rates, not the other way around. In reality, there may be an interesting reciprocal causation. Thus, when the arrival rate (R_a) goes up, backlogs and total time increase, but that causes the arrival rate to go down until an equilibrium may be reached. If the arrival rate goes down, backlogs and total time decrease, but that could cause the arrival rate to go back up. Likewise, when the processing rate (R_p) goes up, backlogs and total time decrease, but that causes the processing rate to go down until an equilibrium may also be reached there. If the processing rate goes down, backlogs and total time increase, but that could cause the processing rate to go back up. The same statistical methods mentioned above for determining the reciprocal relations between the arrival rate and the processing rate can be applied to the relations between total time and backlogs on the one hand, and the arrival and processing rates on the other.

III. Predicting Through Simple Logic

Deductive time-prediction involves predicting time consumption from

[12] On analyzing reciprocal causation see Miller, Logic of Causal Analysis: From Experimental to Non-Experimental Designs, in Causal Models in the Social Sciences (H. Blalock ed. 1971); and Nagel & Neef, Causal Analysis and the Legal Process, in Research in Law and Sociology 201 (R. Simon ed. 1978).

one or more premises that are arrived at either by definition, through inductive statistical analysis, or intuitively. One can classify deductive time-prediction methods into two approaches. One approach might be called a common-sense or simple-logic approach. It involves working with premises in the form of equations that have obvious face validity. All the equation premises that are part of a common-sense reasoning require no mathematical proof in order to convince someone that they fit reality, given certain reasonable and understandable assumptions. A second approach might be called a technical or mathematical modeling approach to deductive time prediction. It involves working with equation premises, when their validity is not intuitively obvious. This is true of the deductive system known as queueing theory. Some of its equation premises may not only lack face validity, but may even appear to be contrary to generally known relations between the time consumption and other variables, at least in the court case context.[13]

A. Predicting Time Consumption, Especially Waiting Time

Predicting through common-sense deduction from processing time and backlogs draws its premises from three different sources. Some are known by definition, others through inductive statistical analysis of data, and still others are known intuitively given the nature of time consumption in the processing of court cases and other governmental cases. For example, we know by definition that total time is equal to waiting time plus processing time:

$$T_t = T_w + T_p. \qquad (1)$$

We can know processing time for a given case through inductive statistical analysis as discussed above. One useful way to predict waiting time is through the equation:

$$T_w = (N_b)(\bar{T}_p). \qquad (2)$$

According to this equation, average waiting time is equal to the number of cases in the backlog multiplied by the average time it takes to process a case. The equation can be accepted intuitively at least for now. Later, additional support will be offered, and this basic equation will be modified to consider such matters as the case settlement rate, the number of judges, and the fact that cases go through successive lines for different kinds of processing.

One reason this equation is useful is because it emphasizes the importance of reducing backlogs and processing time in order to reduce waiting time and thus total time. The equation reflects reality in the sense that waiting time does vary positively in *direction* with both the size of the backlog and the amount of processing time. The equation, however, will not be able to accurately predict the *magnitude* of the waiting time from the size of the backlog and the average processing time unless (1) as soon as one case is finished the

[13] On deductive models in general, see M. Greenberger, M. Crenson & B. Crissey, Models in the Policy Process: Public Decision Making in the Computer Era (1976); S. Nagel & M. Neef, Policy Analysis: In Social Science Research 177-198 (1979); and The Process of Model-Building in the Behavioral Sciences (R. Stogdill ed. 1970).

next case begins with no breaks between cases or within cases, and (2) the cases in the backlog have an average processing time equal to $\bar{T}_p$. If both of those assumptions are met, then Equation 2 reflects reality with regard to both the direction and the magnitude of the relations between T_w on the one hand and N_b and $\bar{T}_p$ on the other. Even if its assumptions are not met, the equation may be useful for judging the relative influence of the variables on the right side of the equation, and for making decisions on the relative value of concentrating resources on one predictor or causal variable rather than another.

To illustrate how those three premises can be put together to deduce a prediction, we need some data. Suppose the backlog as of a given point in time is 20 cases, and the average processing time is a half day. Therefore, the waiting time for the next case in line should be 10 working days. If this case has an average processing time, then it would be disposed of in 10½ working days. In other words, we have gone through the following deductive prediction:

$$\begin{array}{rl} T_w & = (N_b)(\bar{T}_p). \\ N_b & = 20 \text{ cases.} \\ \bar{T}_p & = .5 \text{ days.} \\ \hline \therefore T_w & = (20)(.5) = 10 \text{ days.} \\ T_t & = T_w + T_p. \\ T_p & = \bar{T}_p. \\ \hline \therefore T_t & = 10 + .5 = 10.5 \text{ days.} \end{array}$$

Instead of predicting waiting time for a given case by knowing the size of the backlog and the average processing time, we could predict more precisely by knowing the processing time for each case in the backlog. That would involve an equation like:

$$T_w = T_{p1} + T_{p2} + \ldots + T_{pn}. \tag{3}$$

Equation 3, like Equation 2, assumes that as soon as one case is finished, the next case begins, but it does not assume the average processing time of cases in the backlog is equal to the average processing time of cases in general. Another variation on Equation 1 involves predicting waiting time for an *average* new case, rather than the waiting time for a specific new case. A simple way to predict the average waiting time would be from the average backlog and average processing time using an equation like:

$$\bar{T}_w = (\bar{N}_b)(\bar{T}_p). \tag{4}$$

Often in time-prediction analysis, one is predicting an average time, rather than a specific time, although one may not explicitly say so. Likewise, the prediction is often made from an average time, backlog, or rate without explicitly saying so, but one can usually tell from the context.

Equations 2, 3, and 4 all assume that the cases are processed in box-car fashion, back-to-back. That, however, is not an unreasonable assumption for a court system. There are virtually no court systems that experience times when there are no cases to process, unlike a toll booth where the toll taker may have time periods when there are no cars to be processed. All major court systems have continuous backlogs of varying length consisting of court cases waiting to be processed. There are, of course, interruptions in the processing of cases for weekends, holidays, and night time, but that problem is alleviated by specifying that by days, we mean working days of 8 hours per day, 40 hours per week, and about 250 working days in a 365-day year. That drops about 52 Saturdays, 52 Sundays, and 10 holidays. Likewise, it is not an unreasonable assumption to assume that the average processing time for the cases in the backlog is equal to the average processing time for cases in general, since the backlog is usually a big enough sampling of cases that its average is fairly typical.

As part of the deductive reasoning process, it is sometimes useful to make predictions by transposing terms in some of the above equations. If, for example, we know that total time (T_t) equals waiting time plus processing time, then we logically know that waiting time (T_w) is total time minus processing time, and processing time (T_p) is total time minus waiting time. Likewise, if we know that waiting time (T_w) is backlog multiplied by processing time, then we know that backlog (N_b) is waiting time divided by processing time, and that processing time (T_p) is waiting time divided by backlog. Those two deductions from Equation 2 seem less obvious than the ones deduced from transposing Equation 1. If, however, we substitute the above hypothetical data, they make more sense. Thus, if the waiting time is 10 days, and the backlog is 20 days, that means 10 days per 20 cases, or a processing time of a half day for one case. Likewise, if the waiting time is 10 days, and the processing time is .5 days, that means 10/.5, or a backlog of 20 cases.

It is also sometimes useful to deduce predictions by substituting portions of one equation for portions of another. For example, one could substitute the expression for waiting time (T_w) from Equation 2 in place of the waiting time variable in Equation 1. Likewise, one could substitute the transposed definition of processing time (T_p) from Equation 1 for the processing time variable in Equation 2. There are many such substitutions which could be made. Their usual purpose is to be able to make predictions from certain variables for which information is available, when information may not be readily available for all the variables. As an illustration, if we wanted to predict total time (T_t) from processing time and backlog, instead of from processing time and waiting time, we could use the formula that: $T_t = T_p + (N_b)(\bar{T}_p)$. Likewise, waiting time (T_w) could be predicted by knowing the backlog and the total time. That involves a litte more algebra and the resulting equation is not so obvious. One can, however, show that: $T_w = (N_b)(T_t)/(N_b + 1)$. That equation is, however, not worth separating out, since one would normally want to predict total time from waiting time, not the reverse.

B. Predicting Backlogs

In stating the basic purposes of this article, we mentioned predicting backlogs as well as time consumption. Predicting the size of the backlog is often useful in predicting time consumption, as indicated in Equations 1 and 3 above. Time consumption, though, is what we really want to predict and reduce. We want to reduce backlogs, but only because longer backlogs cause increased time consumption if everything else is held constant.

A simple way to predict how many cases will be in the backlog of cases awaiting processing (as of a given point in time) is by finding out the number of cases remaining at the beginning of the time period (N_r), plus the number of cases that have arrived during the time period up to the present time (N_a), minus the number of cases processed during the time period up to the present time (N_p). That statement can be expressed by the equation:

$$N_b = N_r + N_a - N_p. \qquad (5)$$

All three variables on the right side of the equation are fairly easy to obtain within a court system. This equation is true by definition of backlog, just as saying total time equals waiting time plus processing time is true by definition of total time. If there were 5 cases remaining at the beginning of the period, 20 cases arriving up to the present time, and 10 cases processed, we would logically expect 15 cases to be in the backlog waiting to be processed. The number of cases remaining at the beginning of the time period or year is the same as the backlog at that time. Thus, N_b equals N_r if the number of arrivals is the same as the number of cases processed. For those two figures to be equal, however, would be an unlikely coincidence, although they may be approximately equal.

Averages for N_b, N_r, N_a, and N_p (as of a given date) could be determined by looking to data compiled over a number of years, the way weather analysts determine what the average temperature is for a given day of the year. A more common way to determine an average backlog would be to determine what the backlog was for a lot of recent points in time, sum those backlogs, and divide by the number of points in time. One might also predict the average backlog by using Equation 4 above and determining the ratio between average waiting time and average processing time ($\bar{T}_w/\bar{T}_p$). We are, however, discussing the prediction of backlogs independent of time consumption information, so that we can use the backlog figures to predict time consumption.

C. *Relating Time Consumption to Backlogs*

A series of equations for predicting the variables on the right side of Equation 5 may be arrived at independently from equations that involve transposing the terms in Equation 5, or that involve substituting time consumption terms for the backlog term by using forms of Equation 2 or 4. Thus the number of cases processed could be expressed in terms of the working days which have passed in the time period and the processing rate, or:

$$N_p = (D)(R_p). \qquad (6)$$

For example, if 8 working days have gone by and cases are processed at a rate of 2 per day, then one would expect 16 cases to have been processed. Likewise, the processing rate can be expressed as the number of cases processed divided by the number of working days that have gone by. By substituting $(D)(R_p)$ for N_p in Equation 5, it can be seen more clearly how the backlog varies inversely or negatively with the processing rate. Thus, the more cases that are processed per day, the lower the backlog gets. The processing rate is the reciprocal of the processing time in the simplified situation when there is only one judge. In other words, if the processing rate (R_p) is 2 cases per 1 day, then the processing time (T_p) is 1 day per 2 cases, or .5 days per one case. Thus the backlog varies negatively with the processing rate, but varies positively with the processing time, as is indicated by transposed versions of Equations 2 and 4.

The number of cases arriving can also be expressed in terms of the number of working days which have passed and the arrival rate, or:

$$N_a = (D)(R_a). \qquad (7)$$

For example, if 8 working days have gone by and cases arrive at a rate of 1.5 per 1 day, then one would expect 12 cases to have arrived in 8 days. Likewise, the arrival rate can be expressed as the number of cases that have arrived divided by the number of working days that have gone by. By substituting $(D)(R_a)$ for N_a in Equation 5, we can also see more clearly how the backlog varies positively with the arrival rate. If we obtain the reciprocal of the arrival rate of 1.5 cases per 1 day, that will tell us that 1/1.5 or .67 days are required for one case to arrive, meaning that every two-thirds of a day, one case arrives. The .67 days is the arrival time (T_a) or how often on the average a new case arrives. T_a equals $1/R_a$, just as T_p equals $1/R_p$.

By substituting an expression involving the arrival rate for cases arrived, and substituting an expression involving the processing rate for cases processed, we can express backlog in terms of those two rates, as follows:

$$N_b = N_r + D(R_a - R_p). \qquad (8)$$

We can also express waiting time in terms of arrival and processing rates, as follows:

$$T_w = (N_b)(1/R_p). \qquad (9)$$

Equation 9 would more fully express waiting time as a function of the two rates if the Equation 8 definition of backlog were substituted for the backlog variable in Equation 9. These two equations help clarify one way the common-sense deductive-prediction differs from queueing-theory deductive-prediction. Queueing theory predicts backlogs and waiting time only from arrival rates and processing rates. It does not refer to the additional variables we have used in Equation 5, namely, the number of cases remaining at the beginnning of the time period and the days that have passed up to the present time.

D. Dropouts, Multiple Channels, Multiple Stages, and Sequencing

At least four additional variables need to be considered in order to make the basic time prediction formulas more realistic when applied to court cases and other government case processing. First is the fact that many cases that enter into the line waiting to be processed or subjected to a trial never get there, because (1) the plaintiff decides not to sue or prosecute, (2) the defendant decides not to defend but to lose by default, or (3) the plaintiff and defendant reach a settlement. That requires changing the basic Equation 2 to introduce a variable that might be called the settlement or dropout percentage (S) expressed as a decimal, so that Equation 2 now becomes:

$$T_w = (1-S)(N_b)(\bar{T}_p). \qquad (10)$$

For example, suppose there are 20 cases in the backlog, and they take .5 days apiece to process when they go to trial, and .75 drop out or are settled. Then we would expect the waiting time to be .25 of 10 days. This is so because for all practical purposes, there are really only 5 cases in the backlog if only one-fourth of the 20 are going to go to trial. The cases that are not going to go to trial do not generate waiting time, since waiting time depends on trial time or processing time. The perseverance rate (P) is the complement of the settlement rate (S). This means that if .25 of the cases persevere to trial, then .75 represents dropouts or settled cases, although it is more customary to talk about a settlement rate than a perseverence rate.[14]

Another additional consideration which might improve our predictability is the number of judges hearing the cases. We have been implicitly assuming only one judge or processor. If, however, we increase the number of judges and supportive personnel to two units (J) and they each work at about the same rate, then we would expect the amount of waiting time to be cut in half. This changes the basic equation to read:

$$T_w = (1-S)(N_b)(\bar{T}_p)/(J). \qquad (11)$$

For example, if the above analysis showed the waiting time to be 2.5 days, and two judges divide the cases, then the new waiting time should be 1.25 days. This, however, assumes that the 5 cases that go to trial can be equally divided. That may be true when many cases are involved, as in a typical urban court system. With only 5 cases, though, one judge is probably going to receive 3, and the other 2. If the cases take approximately equal time, then the waiting time will be equal to that of the two-case line, since that is the line that our present case will move into, being the shorter of the two lines. In a large, reasonably well-organized court system, however, the lines should be about equally long before each judge. To be more realistic, we should not divide by J, or the number of judges, since that presumes judges spend their working days doing nothing but trying cases. In reality, an average judge may only spend half a day doing so. The other half of the eight hours may be spent processing cases at stages other than the trial stage, reading the latest appellate court reports to keep abreast of the law, and fulfilling general administrative duties. The J variable should thus be multiplied by .5, or whatever the percentage of case-trying time actually is for the average judge in the system.

[14] On the importance of out-of-court settlements in the criminal and civil justice process, see Franklin, Chanin & Mark, Accidents, Money and the Law: A Study of the Economics of Personal Injury Litigation, in Dollars, Delay and the Automobile Victim: Studies in Reparation for Highway Injuries and Related Court Problems 27, 39-40 (1968); and Newman, Pleading Guilty for Considerations: A Study of Bargain Justice, 46 J. Crim. L.C. & P.S. 780 (1956).

That .5 could be called the case-trying-time coefficient and symbolized (C).[15]

Now that the concept of judge-time (J multiplied by C) has been clarified the relation between processing time (T_p) and processing rate (R_p) can be considered in the realistic situation where there are multiple judges who spend only a percentage of their time trying cases. For example, if a court system with 2 judges processes an average of 8 cases per 1 day over a substantial period of time, that does not mean cases on the average are processed 1 day per 8 cases, or 1/8 days per 1 case, or 1 hour per 1 case at an 8-hour day. If both judges are devoting all their time to trying cases, then 1 day is the equivalent of 2 judge days. Therefore, to determine the processing time from the processing rate, we would not simply flip the 8 cases/1 day to get 1 day/8 cases, but instead we would flip 8 cases/2 judge-days to get 2 judge-days/8 cases, or .25 judge-days/1 case, or 2 hours/1 case at an 8-hour judge-day. In other words, T_p should equal J/R_p, not $1/R_p$; and R_p should equal J/T_p, not $1/T_p$. Being even more realistic involves taking into consideration that the judges may only devote 15 percent of their time to trying cases. Therefore, the consideration is not about 8 cases/2 judge-days, but rather 8 cases/(.15)(2 judge-days), or .30 days/8 cases, or .0375 days/1 case, or .30 hours/1 case at an 8-hour day. In other words, with a multiple judge court, relations between T_p and R_p should be: $T_p = (CJ)/R_p$, and $R_p = (CJ)/T_p$.

A further matter to consider is that moving from the filing of the initial complaint to the final disposition may involve waiting in many lines, stages, or phases, not just a line of cases awaiting trial. There may be lines awaiting preliminary hearing, grand jury indictment, plea bargaining with the prosecutor, the picking of a jury, and finally a line awaiting trial. That means total time should be considered equal to the sum of the waiting times and the processing times which occur in each line or stage that a given case is likely to go through. In other words, variations on Equation 11 should be used for each separate line to calculate separate waiting times. Separate processing times can be determined through the inductive statistical analysis discussed earlier. Total time then involves summing these separate waiting times and processing times by virtue of the definition of total time.[16]

A still further consideration relevant to the time consumption of court cases is the order or sequencing in which the cases are heard. For example, if there are two cases and one has a processing time of 10 days, and the other has a processing time of 5 days, then the waiting time of each case will depend on the order in which they are heard. If the 10-day case is heard first, then it has a waiting time of 0 days, processing time of 10, and total time of 10. Under those circumstances, the 5-day case has waiting time of 10 days, processing time of 5, and a total time of 15. If, on the other hand, the 5-day case is heard first, then it has a T_w of 0, T_p of 5, and T_t of 5. The 10-day case would then

[15] On the importance of judge-time and the lack of it, see Zeisel, supra note 5, at 222.

[16] On the many stages through which criminal and civil cases go, see J. Chaiken, T. Crabill, L. Holliday, D. Jacquette, M. Lawless & E. Quade, Criminal Justice Models: An Overview 90-106 (1976) (hereinafter cited as Chaiken).

have a T_w of 5, T_p of 10, and T_t of 15. Thus, by hearing the shorter case first, the average total time is reduced from $(10 + 15)/2$ down to $(5 + 15)/2$. The common sense approach to taking order into consideration is to calculate total time for each case as the sum of the waiting time and the processing time, using alternative orderings such as first-come, first-served, or shortest-cases-first subject to a maximum time constraint. When many cases are involved, a computer is helpful to calculate the waiting time and processing time for each case under each alternative order, using a prediction-equation like Equation 3. Order, however, does not affect the total time of the last case to enter the line. Its total time is always the sum of the processing times (in accordance with Equation 3) plus its own processing time, no matter how the cases are ordered. When we say the waiting time is influenced by order, we mean the average waiting time of a set of cases (which could be symbolized by $\bar{T}'_w$, or $\bar{T}_w$ primed), as contrasted to predicting the waiting time of either a specific case (T_w) or of an average case ($\bar{T}_w$) entering the back of the waiting line.[17]

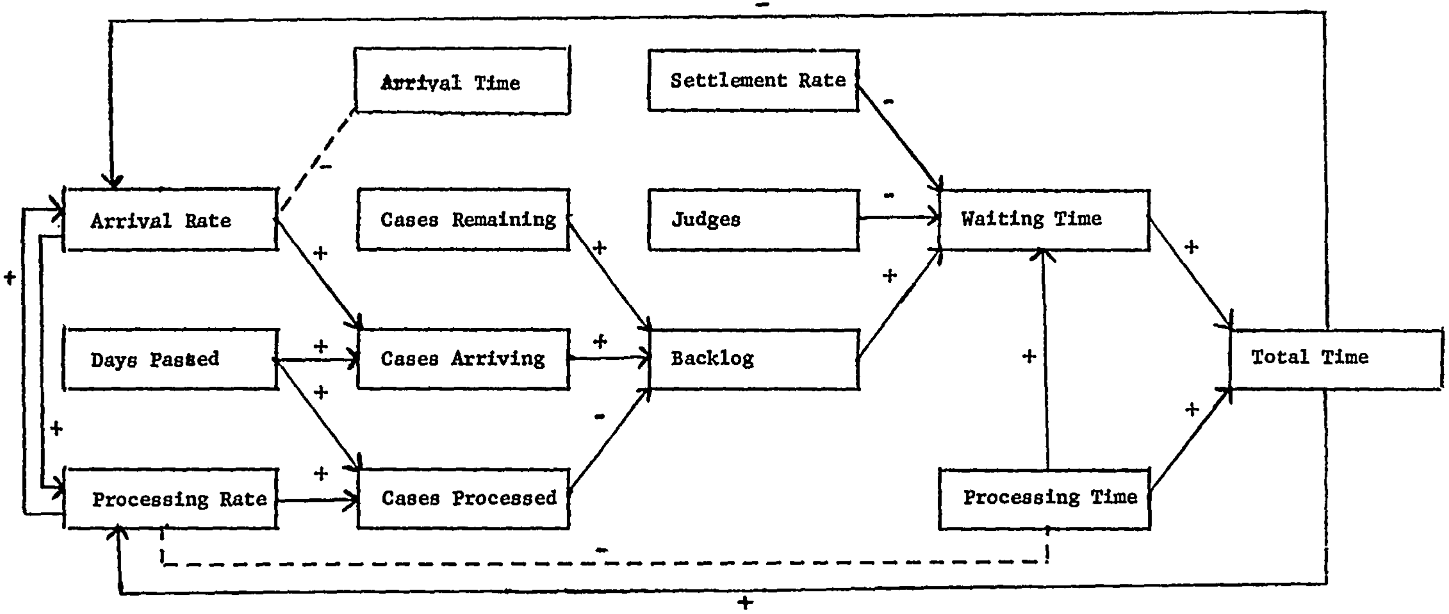
E. Summarizing the Relations

Figure 1 summarizes the relations among the variables that determine total time consumption. The key variable being predicted is total time on the right side of the figure. It is mainly predicted from waiting time and processing time, both of which vary positively with total time, as indicated by the plus signs. The direction of the arrows indicates which variables are causes and which are effects. Waiting time is an effect of processing time. It is also an effect of the settlement rate, the number of judges, and the size of the backlog, with the first two variables having a negative relation with waiting time and backlog having a positive relation. The size of the backlog is in turn determined by the cases remaining, the cases arriving, and the cases processed, with the first two variables having a positive relation and the third variable having a negative relation. The number of cases processed is a positive effect of both days passed and the processing rate, just as the number of cases arriving is a positive effect of both days passed and the arrival rate. The only relations in the figure that are not causal relations are those shown by broken lines rather than by arrows. They indicate that arrival time is the reciprocal of the arrival rate, and that processing time is the reciprocal of the processing rate, although mediated through or multipled by judge-time on multiple-judge courts.

In addition to those basic relations, Figure 1 also shows some reciprocal or feedback relations, where the arrows go in both directions, or go back toward a variable from which causal arrows have previously come. For example, there does seem to be a circular relation between the arrival rate and the processing rate whereby increases in the arrival rate produce increases in the processing rate, and increases in the processing rate stimulate further increases in the arrival rate. How often the arrows occur is a matter for statistical empirical analysis, rather than algebraic deductive analysis, but the common-

[17] On the relevance of case order, see Zeisel, supra note 5, at 200-205; and Nagel & Neef, supra note 1, at 474.

FIGURE 1. SUMMARIZING THE RELATIONS AMONG THE VARIABLES THAT DETERMINE TOTAL TIME CONSUMPTION
(According to the Common Sense Approach)



NOTES:

1. The variable at the tail of each arrow is a cause, and the variable at the head is an effect.
2. Broken lines indicate non-causal definitional relations, i.e., $R_p = 1/T_p$ and $R_a = 1/T_a$.
3. Pluses indicate positive relations, and minuses indicate negative relations.
4. Arrows from right to left indicate feedback relations.

sense approach recognizes that those two variables do affect each other. There are also important feedback relations from total time back to the arrival rate, whereby increases in total time reduce the arrival rate (i.e., cause potential litigants not to sue), and whereby increases in total time tend to speed up the processing rate (i.e., cause judges to work faster to get the delay down). The causal arrows going from left to right are probably stronger in most situations than the feedback arrows moving in the other direction. If the feedback arrows were stronger, that would defeat the gains that could be made by reducing the variables that influence total time. The extent to which one relation is stronger than the other in different situations is also a matter for statistical empirical analysis.[18]

With this figure, it can more clearly be seen how the variables relate to each other, including the relations that involve intermediate variables. For example, processing time has a positive effect on backlog, since processing time through the processing rate has a negative effect on cases processed, which in turn has a negative effect on backlog, and two negatives make a positive. One can likewise trace the relation between any variable and any other variable by following the arrows and bearing in mind that an odd number of negative relations makes for a negative overall effect, but an even number of negative relations makes for a positive overall effect. Tracing the relations between two variables can also include the circular and feedback relations. For example, the reciprocal arrows between arrival rate and processing rate in Figure 1 imply that once the arrival rate increases, the processing rate increases, thus increasing the arrival rate, and so on infinitely upward, or infinitely downward if the arrival rate decreases. That circular causation, however, dampens down because when the arrival rate increases, the backlog increases, and total time increases, which in turn causes the arrival rate to

[18] A feedback arrow could also be drawn from total time to the settlement rate, but it is unclear whether that feedback relation is positive or negative. When the total time goes up, plaintiffs in personal injury cases may prefer to settle rather than to wait so long for damages to be awarded. But an increase in total time may also cause defendant insurance companies to be less willing to settle so that they can have the money to invest for a longer period of time. Those two forces may balance each other. Likewise in criminal cases, when total time goes up, defendants in jail may be more willing to settle by pleading guilty, especially if offered probation or a sentence equal to time already served. But an increase in total time may encourage defendants out of jail to forego a guilty plea in order to prolong their freedom. Finally, a time increase may cause prosecutors to be more willing to settle than defendants, since such an increase allows more time for witnesses to disappear, die or forget, and the prosecution is generally more dependent than the defense on witnesses. Willingness to settle in this context mainly means raising or lowering initial bids in a civil or criminal case.

An increase in total time can also influence *judge-time* by causing the system to hire more judges, and by causing judges to devote a high percentage of their time to trying cases. An increase in total time may, however, have a demoralizing effect on some judges, causing them to slow down even more. In the long run, an increase in total time may become accepted, making the action taken against the increase less vigorous.

Exactly how lawyers and judges respond to changes in time consumption is an empirical statistical question answerable by analyzing behavioral data, rather than by deduction from intuitively accepted premises. For a model helpful in understanding lawyer decision-making in out-of-court settlements, a model into which the empirical data could be inserted, see Nagel & Neef, Plea Bargaining, Decision Theory, and Equilibrium Models, 51 Ind. L.J. 987 (1976) (Part I), 52 Ind. L.J. 1 (1976) (Part II). For a model helpful in understanding judicial-personnel decision-making in the context of time reduction, see Nagel & Neef, supra note 1, at 498-500.

decrease. Likewise when the arrival rate decreases, the backlog decreases, and total time decreases, which causes the arrival rate to increase. Figure 1 can thus be helpful in seeing the direction of the relations in a systems context, as contrasted to the individual equations. The equations, even when realistic, tend to take reality out of context by assuming or by implying (1) constancy among the other variables, (2) independence among the variables on the right side of the equation, and (3) no feedback from the left to the right side of the equation, unless a system of equations is used.

IV. Predicting Through Queueing Theory

A substantially different deductive model for predicting time consumption and backlog has been developed by mathematicians and operations researchers under the general concept of queueing theory. Instead of predicting from processing time and backlogs, queueing theory relies on processing rates and arrival rates.[19] The basic concepts of the two approaches are closely related, but yet their predictive formulas are substantially different. With a one-judge court, the processing rate (as previously mentioned) is simply the reciprocal of the processing time, and the arrival rate is the reciprocal of the arrival time. Thus, if we are only dealing with 2 typical days, and 2 cases arrive on the first day and 4 cases arrive on the second day, then the average arrival rate is 6 cases per 2 days, or 3 cases per 1 day. The arrival time is .33 days per 1 case. Likewise, if over those 2 days, 6 cases are disposed of the first day, and 4 cases the second day, then the average processing rate for the 2 days is 10 cases per 2 days, or 5 cases per 1 day. The processing time is .20 days per 1 case. An especially useful variable in queueing theory is the ratio between the arrival rate (R_a) and the processing rate (R_p). That ratio can be symbolized R_R, or R_a divided by R_p. In the above example, it would be equal to 3 cases per day divided by 5 cases per day, or an R_R of .60.

A. *Predicting Time Consumption, Especially Waiting Time*

In queueing theory, as contrasted to the previous common-sense deductive approach, just the arrival rate and the processing rate are used to predict total time and waiting time. This is unlike Equations 8 and 9 above, which use those rates along with days passed and the number of cases remaining at the beginning of the time period. In order to predict total time, the standard queueing formula is:

$$T_t = 1/(R_p - R_a). \qquad (12)$$

Proving the equation is complicated and the proof is meaningful only if the number of arrivals per day (R_a) distribute themselves in a certain hill-shaped pattern that peaks to the left, and if the number of dispositions per day (R_p) distribute themselves in a certain convex negative curve.[20] Those and other

[19] On queueing theory in general, see note 2 supra.

[20] For proofs of the basic queueing theory equations, see S. Richmond, supra note 2, at 405-38.

assumptions may not be so well met by the way court cases are processed in a typical urban court system, as contrasted to the ways cars are processed at a tollgate where there are no dropouts or multiple stages. In tollgates, there are also times with no cars, as contrasted to the typical urban court which always has cases waiting to be processed, as if the waiting time were infinitely long. Although Equation 12 may involve a complicated proof with some questionable assumptions, the direction of the relations among the variables does make sense. The equation shows that total time varies negatively with the processing rate, since R_p is a positive variable in the denominator. That means if cases are processed faster (more per day), the total time goes down. Likewise, the equation shows that total time varies positively with the arrival rate, since R_a is a negative variable in the denominator. That means if cases arrive faster (more per day), the total time per case goes up.

The standard queueing theory formula for predicting waiting time is:

$$T_w = R_R/(R_p - R_a). \qquad (13)$$

In other words, in queueing theory, waiting time equals total time multiplied by the ratio of the rates. That is the first of many queueing relations that seem to defy common sense. Queueing theory recognizes that waiting time also equals total time minus processing time, as indicated back in Equation 1. There does, however, seem to be an inconsistency for waiting time to also equal total time multiplied by the ratio between processing time and arrival time. The logical inconsistency is, however, resolved by going back to the assumptions of queueing theory, although those assumptions may be empirically inconsistent with the realities of court-case processing.

Applying Equation 12 and 13 to the above hypothetical data shows that total predicted time should be the reciprocal of 3 days minus 5 days, or 1 divided by 2, or .50 days per case. Likewise, predicted waiting time would be .60 divided by 3 minus 5, or .30. That is consistent with the idea that total time equals waiting time plus processing time, since the hypothetical data previously showed that processing time was .20 days per case. The calculations, however, quickly defy common sense as the arrival rate begins to approach or equal the processing rate as often happens in court systems over short or long periods of time, such as in the Washington D.C. court data for the 365 days of 1974. If R_a equals R_p, one can see that both total time and waiting time will involve division by zero, so that both total and waiting time will become infinitely long. Even worse, if R_a exceeds R_p, then both equations will involve dividing by a negative number, which would mean that both total and waiting time will become negative. Conceiving negative time is even more difficult than conceiving infinite time.

B. Predicting Backlogs

To predict the number of cases in the average backlog waiting to be processed, queueing theory uses the equation:

$$N_b = R_R^2/(1 - R_R). \qquad (14)$$

Although justifying the exact nature of this equation is quite complex, the direction of the relation between backlog and the ratio of the rates does make sense. The direction is positive since the ratio constitutes both the numerator and the negative variable in the denominator. That means if the ratio of the arrival rate to the processing rate goes up, the backlog will increase. If, however, the ratio approaches 1.00, then Equation 14 involves dividing by zero and the backlog will become infinitely long. Common sense mandates that when the arrival rate equals the processing rate, or arrival time equals processing time, then the system will be in a kind of equilibrium whereby the courts will be processing cases about as fast as they arrive, as contrasted with the explosive situation where the backlog becomes infinitely long. Again, the situation becomes even more incomprehensible if R_a exceeds R_p, since dividing by a negative number in Equation 14 leads to a prediction of a meaningless negative backlog. In reality, the arrival rate over short and long periods might be greater than the processing rate, thereby building a bigger backlog in view of Equations 5 and 8. The backlogs never become infinite and certainly never negative because (1) there are also times when the processing rate is greater than the arrival rate, (2) generating an infinite backlog would require an infinite number of days with more cases arriving than are being processed, (3) many arrivals drop out even after certifying they are ready for trial, and (4) more judges get added.

Queueing theory also shows a concern for what might be called the backlog primed, which consists of the sum of the cases waiting to be processed plus those currently being processed. That equation is:

$$N'_b = R_R/(1 - R_R). \qquad (15)$$

In analyzing court delay, backlog primed should not be a subject of much concern, as contrasted to the backlog of cases awaiting trial, because once a case goes to trial it is practically completed, and there are relatively few cases in trial, as compared to all the cases awaiting trial. A good approximation of a backlog primed formula simply would be:

$$N'_b = N_b + J. \qquad (16)$$

That equation suggests the addition to the regular backlog of a number of cases equal to the number of judges. That is based on the reasonable assumption that each judge is currently hearing one case, since in most trial court systems, judges complete each trial before starting another one. That assumption is not so true in quasi-judicial proceedings, where administrative judges hear cases in pieces and write opinions later, rather than handing down their decisions at the close of the evidence and the arguments.

A related queueing theory formula demonstrates that the backlog primed (or total backlog) equals the regular backlog (or waiting backlog) divided by the ratio of the rates:

$$N'_b = N_b/R_R. \qquad (17)$$

This follows algebraically from Equations 15 and 14. Comparing Equations 17 and 16, however, implies that dividing the waiting backlog by the ratio of the rates is the equivalent of adding the number of judges to the waiting backlog in order to obtain the total backlog. That does defy common sense, and it shows there is no logical relation between common-sense time-prediction and queueing-theory time-prediction. Equation 17 also seems to run contrary to the accepted idea that the total backlog (N'_b) equals the waiting backlog (N_b) plus the processing backlog (i.e., the quantity of cases currently being processed). Equation 17 in effect says that adding the processing backlog to N_b is the equivalent of dividing N_b by R_R. That equivalence is as difficult to grasp as saying that adding processing time (T_p) to waiting time (T_w) is the equivalent of dividing T_w by R_R to get total time (T_t), as is done in relating Equations 13 and 12. To add to the strangeness of queueing theory, one can algebraically prove that N_b/R_R is equal to $N_b + R_R$ since (1) $N_b/R_R = N'_b$ according to Equation 17, and (2) $N_b + R_R = N'_b$ if Equation 14 is substituted for N_b and if Equation 15 is substituted for N'_b.

There are a number of queueing theory equations (based on Equations 12, 13, 14, 15, and the basic assumptions) that theoretically enable the prediction of the probability that there will be zero cases backlogged in the system, just 1 case, 2 cases, any number of cases, any number more than a certain number, or any number less than a certain number. For the purpose of processing court cases, it is necessary to predict and control time consumption so that the the backlogs may be predicted. If the predicted backlog as of a given point in time is 2,000 cases, it is not important to know the chance probability that there might suddenly be 2,100 cases as a result of a chance upsurge in cases arriving, or 1,900 cases as a result of chance upsurge in cases processed. Any of those reasonable temporary deviations from the predicted backlog would still mean that any given judge in the system would have a case ready for processing as soon as he finishes the case or stage he is now working on.

It would, however, be important to know how backlogs might be increased or decreased as a result of introducing new case-diversion procedures (like no-fault insurance) or new case-settlement procedures (like court-appointed mediators). Neither queueing theory nor common-sense deduction will answer questions like those. They require a statistical analysis, preferably one that involves an experimental or quasi-experimental group compared with a control group. An experimental group of cases in this context is one that has been randomly assigned to a delay-reduction method while quasi-experimental group would receive a delay reduction procedure through self-selection or other non-random assignment, but with an attempt to determine and statistically control for or equalize the characteristics of the quasi-experimental group and the control group. A control group is a group of cases that is not subjected to the delay-reduction procedure whose effect on time consumption it is sought to determine.

C. Relating Time Consumption to Backlogs

Queueing theory, like the common-sense approach, also can make predictions of waiting time from processing time and backlogs, but the equations look quite different, since they are based on Equations 12 through 15. More specifically, to predict waiting time from average backlog and processing time queueing theory uses the equation:

$$T_w = N'_b/R_p. \tag{18}$$

Equation 18 logically follows from Equations 13 and 15 by simply substituting the formula for N'_b from Equation 15 into Equation 18 and then simplifying. Doing so gives Equation 13. The expression on the right side of Equation 18 is algebraically equal to $(N'_b)(T_p)$, which makes Equation 18 look like Equations 2 or 4 with two exceptions.

The first difference, which is not important, is that queueing theory usually uses the backlog primed rather than the normal backlog to predict waiting time. That technically means that if there are 10 cases in line and one being processed, then the next case to enter the line will have to wait for 11 cases to be processed before it comes to the head of the line. If, however, there is 1 case being processed, it may be more than half done and can be considered as not so important in producing waiting time. This is especially so since a judge normally processes only one case at a time, unlike a telephone switching system which processes many phone calls simultaneously. Thus, there is never more than one case being processed or heard in a given trial-court line before a given judge, and thus using N'_b or N_b does not represent an important difference.

The second difference between Equation 18 and Equations 2 and 4, which is important, is that the backlog variable in Equation 18 is predicted from Equations 14 or 15, rather than from Equations 5 or 8. Equations 14 and 15 rely on a completely different kind of reasoning than Equations 5 and 8. The queueing theory equations rely on mathematical models, which involve assumptions that are difficult to comprehend intuitively, and that may run contrary to empirical reality in the context of court cases and other governmental cases more so than our common-sense approach to time prediction.

An alternative way to relate time consumption to backlogs would be to express waiting time in terms of the waiting backlog rather than total backlog. Doing so involves the equation:

$$T_w = N_b/R_a. \tag{19}$$

That equation algebraically follows from Equations 13 and 14 by substituting the formula for N_b from Equation 14 into Equation 19 and then simplifying. Doing so gives the same equation as Equation 13. Equation 19 means that if there are 10 cases in the backlog and the arrival time is .67 days (meaning 1 case every ⅔ of a day), then the predicted waiting time would somehow be ⅔ of 10 days. This is not a meaningful result. At first glance, Equation 19 looks even worse than it really is, since R_a being the denominator implies that there is a negative relation between the arrival rate and waiting time, meaning if R_a increases T_w decreases. That, however, is offset by the fact that R_a has a triple positive influence on T_w by virtue of its being squared to calculate N_b and also being a negative variable in the denominator of N_b. In other words, the direction of the relations among the variables in queueing theory fit common sense, but not necessarily the magnitudes that are predicted by those relations, especially magnitudes that are infinite, negative, or based on variables that do not seem to belong.

D. Dropouts, Multiple Channels, Multiple Stages and Sequences

The dropout phenomenon is not so common in the kinds of waiting lines that queueing theorists usually deal with. A typical queueing theory situation is a tollgate. It is difficult to conceive of a car pulling up to a tollgate line, and then deciding to turn around and go back the wrong way on the highway, rather than go through the tollgate. Sometimes, people do not have the money to pay at the tollgate, but their cases are generally resolved through an I.O.U., a waiver, or an arrest. The subject of dropouts is discussed in advanced queueing treatises under the concept of reneging or Markovian impatience as an especially complex problem in view of the assumptions associated with queueing theory. Even in advanced queueing theory treatises, exact formulas are not given, but instead the reader is referred to more specialized treatments. In court processing, however, about three-fourths or more of all the criminal and civil cases drop out without ever coming to the head of the line, as a result of one side or the other withdrawing, or both sides reaching a mutual agreement. Taking into consideration a settlement or perseverance probability as is done in Equation 10 seems to make more sense than ignoring or complicating a phenomenon that is so important and so basically simple in understanding and reducing time consumption and backlogs in court-case processing.[21]

The phenomenon of multiple judges or channels through which cases can go also seems simple from a common-sense perspective, but becomes quite complicated in queueing theory, although not as complicated as the phenomenon of cases being settled or withdrawn before trial. Queueing theorists have developed an exact equation for predicting time consumption and backlogs where one of the variables is the number of judges or channels. That equation is as follows:

$$N'_b = [(R_a)(R_p)(R_R)^J/(J-1)!(JR_p-R_a)^2](I) + (R_R). \qquad (20)$$

The expression $(J-1)!$ means subtract 1 from the number of judges and then multiply that figure by $J-2$, $J-3$, and so on down to $J-J+1$. The symbol "I" stands for the probability that there will be idle time, meaning a zero backlog. For the purpose of court case processing, the I probability can be set at about zero. Doing so would simplify the above equation since everything to the left side of plus sign would then become zero and the total backlog would become equal to R_R, regardless of how many judges there are. That result is about as meaningless as an infinite or negative backlog. To calculate an exact I probability rather than set it to zero requires an equation even more compli-

[21] On the handling of dropouts or reneging in queueing theory, see Gross, supra note 2, at 134-40.

cated than Equation 20.[22]

One would think that queueing theory would predict time consumption for multiple stages by simply summing the predicted total times for each separate processing stage. That summation approach only applies if many assumptions are met. The more common queueing situation with multiple stages or phases is referred to as queueing in tandem or series and requires the solving of many equations simultaneously. The equations are especially difficult to solve because they are difference-differential equations, which require the use of complicated equation-solving methods and advanced calculus.[23]

On the matter of taking into consideration the order or sequencing of the cases to determine the average total time per case, queueing theory does show some concern for case order under the concept of queueing discipline, including the hearing of cases on a priority basis rather than a first-come, first-served basis. Under queueing theory, to predict time consumption or backlogs when cases are heard on a priority basis requires solving sets of difference-differential equation, as in the multi-stage situation. The priority discipline situation, though, is worse in the sense that queueing theory only seems capable of talking in terms of priority categories, like category 1, category 2, and (with difficulty) more than two categories. The kind of priority system that would be meaningful for court cases, however, might involve hearing the cases in the order of their predicted processing times. Each case would thus have a different processing time, rather than be positioned in two crude categories, like short cases and long cases.[24] In other words, the problems of dropouts, multiple judges, multiple stages, and deviations from first-come, first-served, which are so much a part of court case processing cannot be easily handled by queueing theory. Queueing theory equations may, however, be quite helpful in understanding and reducing time consumption at tollgates, telephone exchanges, manufacturing processes, and other types of waiting lines.[25]

E. Summarizing the Relations

Figure 2 summarizes the relations among the key variables that determine total time consumption according to queueing theory. The most important variable is total time, which is shown at both ends of the figure in order to make the figure easier to read. The upper part of the figure shows that both the waiting and total backlogs have a positive relation with total time, accord-

[22] On the handling of multiple judges or channels in queueing theory, see S. Richmond, supra note 2, at 428-430.

[23] On the handling of multiple stages or series in queueing theory, see Gross, supra note 2, at 198-213.

[24] On the handling of priority ordering in queueing theory, see id. at 178-97.

[25] Queueing theory also provides equations for predicting waiting time from arrival rates and processing rates in situations that do not apply so well to the courts: truncating, whereby complaints are turned away in civil or criminal cases because there are too many; jockeying, whereby complainants move from one judge to another because the second judge now has a shorter line; or recycling, whereby cases that reach a certain stage have to start over again at an earlier stage. These processes may occur in other kinds of waiting lines, but not so much in judicial waiting lines.

ing to Equations 18 and 19. The ratio of the arrival rate to the processing rate has a positive relation with both the waiting and total backlogs, according to Equations 14 and 15. The ratio also has a positive relation with waiting and total time, according to Equations 12 and 13. The arrival rate has a positive relation with both backlogs, since it is the numerator of the ratio. Likewise, the processing rate has a negative relation with both backlogs, since it is the denominator of the ratio. The processing rate has a negative relation with both waiting and total time, according to Equations 12 and 13, and it has a negative relation with the ratio by definition. The arrival rate also has a positive relation with waiting time, according to Equations 12 and 13, and it has a positive relation with the ratio by definition.

In addition to those basic queueing theory relations, it should also be noted that queueing theory assumes a zero relation between the arrival rate and the processing rate. That may be contrary to common knowledge about how government and non-government bureaucrats respond to increased and decreased work, and to common knowledge about how citizens and customers respond to fast and slow bureaucrats or salespeople. Queueing theory allows for no feedback relations from total time and backlogs (at the upper part of Figure 2) down to arrival and processing rates (in the lower part). That may also be contrary to related common knowledge. Figure 2 could be supplemented by inserting variables that relate to the number of judges or channels, and the settlement or reneging rate. Doing so would make Figure 2 too complicated, even though those variables can be accommodated in Figure 1, which is based on the common-sense approach to time prediction. Likewise, the negative reciprocal relations between the processing rate and processing time, and between the arrival rate and arrival time could be shown, but that is not necessary to queueing theory.

To summarize further the relations among the queueing theory variables, Equations 12 through 19 can be applied to the hypothetical data given at the beginning of the queueing theory section. Doing so yields the following numerical results:

$R_a = 3$ cases. $R_p = 5$ cases. $R_R = .60$.

12. $T_t = 1/(5-3) = .50$ days.

13. $T_w = .60/(5-3) = .30$ days. (Also $.50 - .20 = .30$ days.)

14. $N_b = .60^2/(1-.60) = .90$ cases.

15. $N'_b = .60/(1-.60) = 1.5$ cases.

17. $N'_b = .90/.60 = 1.5$ cases. (Also $.90 + .60 = 1.5$ cases.)

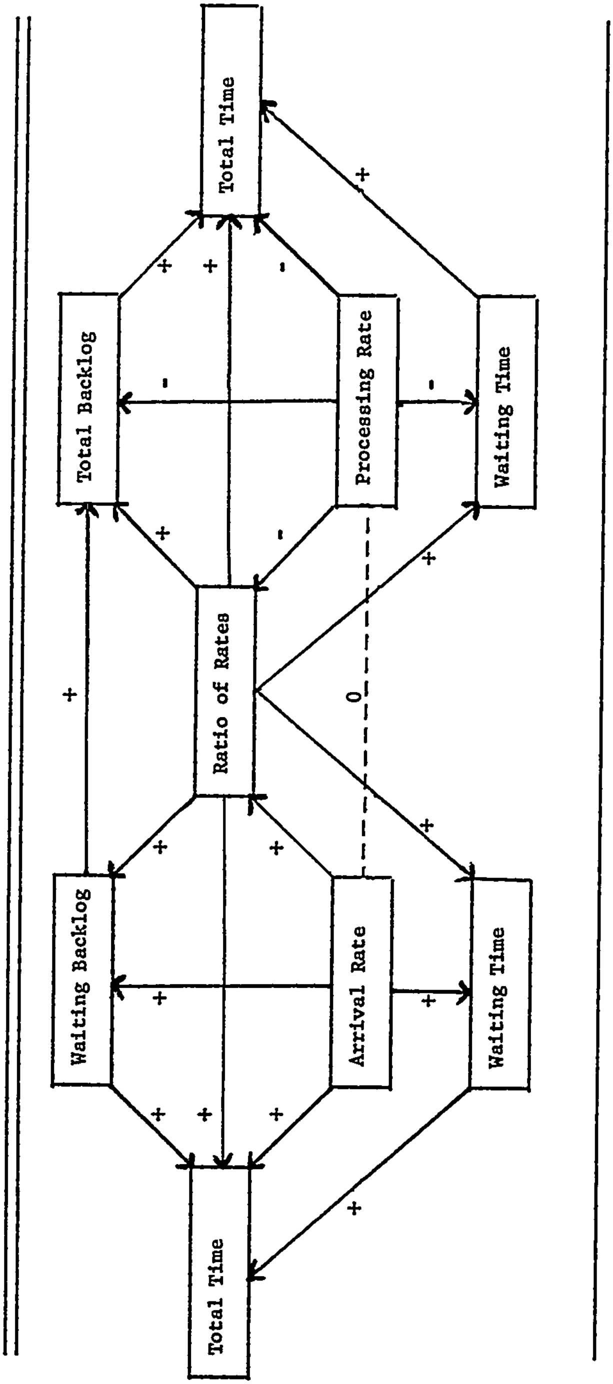
18. $T_w = 1.5/5 = .30$ days.

19. $T_w = .90/3 = .30$ days.

That summary does show there is internal consistency among the queueing theory equations, although not necessarily external consistency with reality in the court case context.

To summarize still further the relations between the queueing theory equations and the other set of equations, the first set of equations can be ap-

FIGURE 2. SUMMARIZING THE RELATIONS AMONG THE VARIABLES THAT DETERMINE TOTAL TIME CONSUMPTION
(According to the Queueing Theory Approach)



plied to the same data. To do so, however, it is necessary to know the size of the backlog when the cases began. The backlog must have been bigger than 4 cases since the hypothetical data says that 6 cases were processed on the first day, but only 2 cases arrived on the first day. We could assume a hypothetical backlog of 12 cases. With that item and the above data, the common-sense equations yield the following numerical results, with an assumption of a one-judge court through Equation 10:

$T_p = .20$ days. $N_b = 12$ cases.

1. $T_t = 2.40 + .20 = 2.60$ days (Equation 2 has to be solved first.)
2. $T_w = (12)(.20) = 2.40$ days. (Predicted T_w at day 0.)
3. We do not have data on the individual cases, only the averages.
4. Same as 2 above, if averages are involved.
5. $N_b = 12 + 6 - 10 = 8$ cases. (N_b after 2 days.)
6. $N_p = (2)(5) = 10$ cases.
7. $N_a = (2)(3) = 6$ cases.
8. $N_b = 12 + 2(3 - 5) = 8$ cases.
9. $T_w = (8)(1/5) = 1.60$ days. (Predicted T_w after day 2.)
10. $T_w = (1-.75)(8)(.20) = .40$ days. (Assuming a settlement rate of .75.)
11. $T_w = (1-.75)(8)(.20)/(.50)(4) = .20$ days. (Assuming 4 judges in trial half-time.)
16. $N'_b = 8 + 4 = 12$ cases.

This summary shows that there is also internal consistency among the common-sense equations. The two sets of equations are equally easy to apply with a calculator, with the exception of the calculations of reneging, multiple judges, multiple stages, and sequencing, where only the common-sense approach involves simple equations. Unfortunately, hypothetical data does not provide a meaningful test of how accurately the respective approaches can predict time consumption and backlogs. The queueing theory approach does seem to lack accuracy when R_a approaches or exceeds R_p. One purpose of this Essay is to stimulate the testing of these respective approaches with actual court data, which the author is currently obtaining from the Administrative Office of the U.S. Courts and the Federal Judicial Center. That data may also lend itself to applying statistical inductive analysis to predicting processing times for different kinds of cases.

V. Applications, Evaluations, and Research

In the context of predicting and reducing court-case time, the most important equations are those that relate to waiting time, rather than total time or backlogs. This is so because (1) waiting time constitutes nearly all of total time, (2) the portion of total time that is trial time or processing time is not as subject to prediction by deductive equations, and is not as subject to reduction without violating constitutional rights, and (3) the waiting time equations explicitly or implicitly include the backlog equations since the size of the backlog is a key determinant of waiting time. The essence of the 12 common-sense equations just summarized is thus:

$$T_w = (1-S)(N_b)(T_p)/(CJ). \tag{21}$$

Likewise, the essence of the 7 queueing-theory equations just summarized is:

$$T_w = (R_a/R_p)/(R_p-R_a). \tag{22}$$

Equation 21 may look more complicated than Equation 22 because it has more variables (5 versus 2), but that is only because it more realistically considers the settlement rate (S), the backlog (N_b), processing time (T_p), and judge-time (CJ), and not just the arrival rate and processing rate for a one-judge court with no reneging or settlements. Equation 21 is also simpler in having intuitive meaningfulness, although its meaningfulness is clearer when one discusses each part separately, as was previously done. In this context of Equation 21, it is not useful to express N_b in terms of R_a and R_p. That was done in Equation 8 mainly to try to relate the common-sense approach to the queueing-theory approach through those shared variables. R_p is in effect included in Equation 21 since it is equal to CJ/T_p. R_a is also in effect included in Equation 21 since it is a key determinant of N_b.

A. *Applying the Common-Sense Equations to Making Policy Decisions*

These equations can be helpful in making policy decisions relevant to reducing waiting time. That type of policy analysis first requires deciding what the optimum or desired waiting time is, which is virtually the same as deciding the optimum total time since waiting time is so much of total time. Various approaches can be used to determine optimum or desired waiting time, such as looking to other courts, previous time periods, the time that could be saved if above average cases were reduced to the average, minimizing the sum of the delay costs and speed-up costs, or somewhat arbitrarily choosing a waiting time that is 10 percent below whatever the current waiting time is.[26] The optimum or desired waiting time can be symbolized $T^\star_w$.

The next step is to solve for the corresponding optimum values of $S^\star$, $N^\star_b$, $T^\star_p$, $C^\star$, $J^\star$, $R^\star_a$, or $R^\star_p$, when all the other variables are held constant. For example, suppose the desire is to reduce predicted waiting time from .20 days to .10 days, using data from the previous Section IV-E. How much must the settlement rate be increased in order to arrive at a waiting time of .10 days? Answering that question involves solving the equation:

$$S^\star = 1.00 - [(T^\star_w)(CJ)/(N_b)(T_p)]. \tag{23}$$

The equation is arrived at by simply transposing the terms from Equation 21. The $T^\star_w$ in the numerator of the fraction in brackets is not cancelled out by $(N_b)(T_p)$ in the denominator since $T^\star_w$ is not necessarily equal to T_w. Applying this equation to the numerical data yields the following results:

$$S^\star = 1.00 - [(.10)(.50)(4)/(8)(.20)] = .87.$$

In other words, to reduce waiting time from .20 days to .10 days by just chang-

[26] On finding an optimum time to consume, see authorities cited in note 9 supra.

ing the settlement rate, it is necessary to raise the settlement rate from .75 to .87. The deductive validity of that statement can be shown by inserting .87 in place of .75 in Equation 11 of Section IV-E, and observing that the new predicted waiting time is then .10 days rather than .20 days.

Equation 21 not only enables a calculation of an optimum value for each of the five variables on the right side of the equation, but it also provides a better idea as to how much leverage one gets (in reducing waiting time) from a one-unit change on each variable. For example, by examining Equation 21, it can be seen that a one-unit increase in S will result in a decrease in T_w that is equal to the value of $(N_b)(T_p)/(CJ)$. That is pretty good leverage, because N_b tends to be a large number. Using the hypothetical data, this means that if the settlement rate goes up one unit from an S of 0 to an S of 1.00, then the waiting time will go down (8)(.20)/(.50)(4), which equals .80 days. Thus, if the settlement rate goes up .01 units, the waiting time should go down .80/100 days or .008 days. This means if the settlement rate goes up .12 units from .75 to .87, waiting time should be reduced 12 times .008 days. That is a reduction of −.10 days from .20 to .10, which checks with the previous calculations.[27]

The optimizing aspects of the other four variables in Equation 21 and the two variables in Equation 22 can also be analyzed. The amount the backlog (N_b) must be reduced in order to get the waiting time down from .20 to .10, can found through the equation:

$$N^{\star}_b = (T^{\star}_w)(CJ)/(1-S)(T_p). \tag{24}$$

As with Equation 23, this equation is arrived at by transposing the terms from Equation 21, after determining an optimum or desired value for T_w. The optimum reduction of backlog as shown by this equation is: $N^{\star}_b = (.10)(.50)(4)/(1-.75)(.20)$, or 4 cases. That means the backlog must be reduced from 8 cases to 4 cases, if the waiting time is to be reduced from .20 to .10. Inserting a 4 in place of the 8 in Equation 11 of Section IV-E is proof of that result. It should also be noted that a one-unit reduction in the backlog does not have much impact on reducing waiting time, since the slope or marginal rate of return of waiting time to backlog is $(1-S)(T_p)/(CJ)$, as derived from Equation 21. That slope does not have any big numbers in it, since it does not include the backlog. Processing time tends to be a small number, and it is further discounted by the perseverance rate, which is the complement of the settlement rate. The number of judges tends to be relatively small compared to the size of the backlog, and it is normally substantially discounted by the coefficient of case-trying time.

To reduce waiting time from .20 to .10 by reducing processing time involves solving for $T^{\star}_p$ in the equation:

$$T^{\star}_p = (T^{\star}_w)(CJ)/(1-S)(N_b). \tag{25}$$

Applying that equation to the hypothetical data yields: $T^{\star}_p = (.10)(.50)(4)/(1-.75)(8)$, which equals .10. Therefore, to reduce the waiting

[27] On how one determines the slope or marginal rate of return of a change in Y relative to a change in X, when one knows how Y relates to X, see M. Brennan, Preface to Econometrics: An Introduction to Quantitative Methods in Economics 111-129 (1973).

time from .20 to .10, it is necessary to reduce T_p from .20 to .10. The slope of waiting time to processing time is $(1-S)(N_b)/(CJ)$. This means that in a one-judge court where all cases in the backlog are tried, a one-unit reduction in processing time will be multiplied by the size of the backlog, to produce a big effect on waiting time. For example, if there are 500 cases in the backlog, and processing time is 2 days, then a one-day reduction in processing time will reduce waiting time from 1,000 days to 500 days, which is a good return or leverage on a one-unit change in processing time. That is worth emphasizing because court reformers often consider processing time reductions to be unimportant, since processing time is already generally low. That, however, ignores the multiplier effect of N_b in the slope, leverage, or MRR factor. The leverage, though, is reduced to the extent that the backlog is (1) discounted by the perseverance rate and (2) diluted by the amount of judge-time. The leverage of processing time is also worth emphasizing because policy analysts sometimes consider two variables that are multiplied together to be of equal importance since they both have an exponent of one, such as T_p and N_b. That, however, ignores the fact that the average T_p may be much smaller (or larger) than the average N_b, and that the spread around the averages may also be quite different where spread is a relevant consideration.

The optimizing aspects of judge-time in the denominator of Equation 21 are somewhat different than those of the variables in the numerator. For example, if the desire is to get waiting time down from .20 to .10, by increasing the percentage of time that judges spend in trial, the relevant equation would be:

$$C^\star = (1-S)(N_b)(T_p)/(T^\star_w)(J). \qquad (26)$$

With the hypothetical data, $C^\star = (.25)(8)(.20)/(.10)(4)$, which exactly equals 1.00. That means the judges would have to be spending all their time in trial to produce that desired waiting-time reduction. What this illustrates is that optimum values may sometimes exceed feasibility constraints. It is simply not feasible to have judges spend all their time trying cases and no time in pretrial or other activities. With other hypothetical data, $C^\star$ could have even come out greater than 100 percent, which would have been contrary to measurement feasibility, as well as pragmatic feasibility. Another interesting aspect of the relation between T_w and C is that a one-unit increase in C produces a decrease in T_w equal to $(1-S)(N_b)(T_p)/(J)(C^2)$. This slope (with C^2 in it) stems from the rule of slopes that says, if $Y = 1/X$, then the slope of Y relative to X is $-1/X^2$. In more common-sense terms, what is happening is that for every increase in C, there is a decrease in T_w, but at a diminishing rate. That means T_w goes down more when C goes up from 0 percent to 10 percent, than when C goes up from 90 percent to 100 percent. In other words, the amount of leverage C has on T_w depends on C itself. On the other hand, the variables in the numerator of Equation 21 bear a constant relation with T_w, regardless of the numerical value of those delay-reduction variables.

There is a similar relation between T_w and the number of judges (J), since J and C are closely connected in the concept of judge-time. To determine how many additional judges are needed for waiting time to move from .20 to .10, the relevant equation is:

$$J^{\star} = (1-S)(N_b)(T_p)/(T^{\star}_w)(C). \qquad (27)$$

Equation 27 is the same as Equation 26, except that J and C change places. Inserting the hypothetical data, the result is: $J^{\star} = (.25)(8)(.20)/(.10)(.50)$, which equals 8 judges so that the court system would have to double the number of judges from 4 to 8. Like the slope of T_w relative to C, the slope of T_w relative to J also involves diminishing returns, meaning that an increase from 0 judges to 1 judge has a bigger effect in diluting the backlog than an increase from 10 judges to 11 judges. More specifically, the slope of T_w to J is $(1-S)(N_b)(T_p)/(C)(J^2)$. The chosen value of J has to be inserted into that slope in order to determine how much leverage it will have. This is unlike the backlog variable, since a backlog increase from 0 to 1 case has just as much effect on waiting time as a backlog increase from 10 to 11 cases.

After deciding on an optimum or desired waiting time, and after determining how much S, N_b, T_p, C, and J would have to change in order to achieve that desired waiting time, the next logical step in applying Equation 21 is to decide how to allocate a court budget or other resources to those variables in order to achieve the desired waiting-time. The optimum waiting time $(T^{\star}_w)$ may be found by either (1) increasing the settlement rate 12 percentage points, (2) decreasing the backlog by 4 cases, (3) decreasing the processing time by .10 days, or .80 hours at an 8-hour day, or 48 minutes at a 60-minute hour, (4) increasing the percentage of time judges spend in trial by 50 percentage points, or (5) increasing the number of judges by 4 judges. If the cost of each alternative is known, then the least expensive option is the obvious choice.

That approach has a least three drawbacks that need to be taken into consideration. First, some of the alternatives may not be feasible, such as increasing C by 50 percentage points. The budget may not be big enough to provide for hiring 4 more jùdges. The easiest approach to reducing processing time may be not allowing oral testimony or cross-examination, but that would be unconstitutional at least in criminal cases. Thus, the measurement, economic, and political constraints must all be taken into consideration. Second, some of the costs for each of the five alternatives may be non-monetary in nature. For example, it may be constitutional to divert 4 civil cases to an informal non-adversarial proceeding, but so doing may increase the likelihood of inaccurate decisions. Some non-monetary costs may have to be given a monetary value to compare the alternative delay reduction methods, or else the various goals (like saving money and having accurate decisions) will have to be given relative importance weights, as part of a broader benefit-cost analysis. Third, there are likely to be diminishing returns between dollars spent and achieving improvements in S, N_b, T_p, C, and J. This means that for $1,000 spent to hire special mediators to encourage settlements, the settlement rate would move from .75 to .80, but to move it from .80 to .85 might cost $10,000 in mediator time. In other words, the relation between S and dollars-spent ($) may not be a linear relation of the form: $S = a + b(\$)$, but rather a diminishing returns relation of the form: $S = a(\$)^b$. If there were (1) a constant or linear relation between dollars spent and each of the five delay-reduction variables and (2) a constant relation between each of those variables and waiting time, then it would be appropriate to spend all one's time-reduction budget on the one best variable after meeting minimum constraints on the other variables, rather than allocating one's time-reduction budget among all five variables or a subset greater than one. There are a number of recent books and articles on the subject of optimally allocating a budget mix in order to take into consideration constraints, non-monetary goals, non-linear relations, and other matters. Like predicting court-case time, arriving at an optimum mix can also be an exercise in simple logic.[28]

B. Applying the Queueing Theory Equations to Making Policy Decisions

Equation 22 like Equation 21 can also be applied to making policy decisions. The three step process involves (1) determining the desired waiting time, (2) determining how much the variables have to change in order to achieve that waiting time, and (3) determining how to allocate one's budget resources among those variables in light of steps 2 and 1. A big disadvantage of Equation 22 is that it only applies to a one-judge court, but Equation 20 can also be used for multiple-judge courts. The even more complicated reneging equations of queueing theory can be used to consider the dropout rate. Using Equation 22, however, sufficiently illustrates the complexity and some other problems involved in applying queueing theory to the above kind of policy analysis.

There are only two variables subject to manipulation by way of Equation 22, namely, the arrival rate and the processing rate. Operating in accordance with the data shown in Equations 12 through 19 in Section IV-E, the optimizing problem can be stated as one of reducing waiting time from .30 days down to .10 days. Queueing theory Equation 19 predicts a waiting time of .30 days, whereas common-sense Equation 11 predicts a waiting time of .20 days as our starting point. Those differences are partly due to the nature of the equations, but also to the fact that the common-sense example uses four half-time trial judges in the common-sense example rather than one full-time trial judge as in the queueing theory example. The amount that the arrival rate must be reduced in order to get T_w down from .30 to .10, is found by solving for $R^{\star}_{a}$ in the equation:

$$R^{\star}_{a} = (T^{\star}_{w})(R_p^2)/[(T^{\star}_{w}R_p)+1]. \qquad (28)$$

That equation is derived by transposing the terms in Equation 22 and inserting $T^{\star}_{w}$ in place of T_w, as was done in working with Equation 21. Applying that equation to the hypothetical data yields: $R^{\star}_{a} = (.10)(25)/[(.10)(5)+1] = 1.67$ cases. That means the current arrival rate of 3 cases per day would have to be reduced to 1.67 cases per day in order to get waiting time down from .30 to .10,

[28] On optimally allocating a budget or arriving at an optimum mix of alternative activities, see S. Nagel, Policy Evaluation: Making Optimum Decisions (1981); E. Stokey & R. Zeckhauser, A Primer for Policy Analysis 134-58, 177-200 (1978); and M. White, R. Clayton, R. Myrtle, G. Siegel & A. Rose, Managing Public Systems: Analytic Techniques for Public Adminstration 205-23, 319 (1980).

given the processing rate of 5 cases per day and the queueing theory assumptions. The consistency of the 1.67 figure can be tested by inserting it into Equation 22 along with an R_p of 5, and observing that T_w comes out .10. The 1.67 is an average arrival rate, meaning that for every day in which there is an arrival of one case, there are two days in which the two cases arrive.

The method for determining the slope of waiting time to the arrival rate using Equation 22 is more complicated than the method for determining the slope of waiting time to S, N_b, T_p, C, or J using Equation 21. The method results in a slope equal to $1/(R_p{-}R_a)^2$. That expression shows that the slope of T_w to R_a depends on the numerical value of R_a, meaning the relation is a curved line rather than a straight one. A more complicated analysis of that expression and of Equation 22 indicates the curved line relating T_w to R_a curves upward at an increasing rate. This means that as R_a increases (when R_p is held constant), T_w increases even more rapidly that R_a does. When R_a approaches the value of R_p, then T_w (according to queueing theory) approaches infinity. A more useful and common-sense approach to the notion of slope in the context of queueing theory would be to say that with a starting waiting time of .30 days, by going down from an arrival rate of 3 cases per day to an arrival rate of 1.67 cases per day, waiting time was reduced to .10. That means a change in T_w of .20 and a change in R_a of 1.33. Thus, the slope over that portion of the curve is .20/1.33, or .15/1, meaning a decrease of 1 case per day in the arrival rate would produce a decrease of about .15 days in waiting time, and a 1.33 decrease in R_a would produce a .20 decrease in T_w.

In light of queueing-theory Equation 22, how much would the processing rate have to go up in order to get waiting time down from .30 to .10? The answer involves solving for R_p in the equation:

$$(T^{\star}_{w})(R_p^2) - (T^{\star}_{w})(R_a)(R_p) - (R_a) = 0.$$

That equation algebraically follows from Equation 22. Solving for R_p in the above equation involves using the quadratic equation-solving formula, so that the optimum or desired value of R_p is:

$$R^{\star}_{p} = \{-(-T^{\star}_{w})(R_a) \pm [(-T^{\star}_{w})^2(R_a)^2 - (4)(T^{\star}_{w})(-R_a)]^{.5}\} \ / \ (2)(T^{\star}_{w}). \quad (29)$$

Inserting the hypothetical values of .10 for $T^{\star}_{w}$ and 3 for R_a, $R^{\star}_{p}$ equals either 7.18 or −4.20. Only 7.18 makes sense since there cannot be a negative processing rate. The 7.18 value especially makes sense if it is inserted back into Equation 22 along with an R_a of 3. The result using these values is a waiting time of .10.[29]

The slope of waiting time to the processing rate involves the same complicated method used to determine the slope of waiting time to arrival rate. It results in a slope equal to $[R_a(R_a - 2R_p)] \ / \ [R_p^2(R_p - R_a)^2]$. That expression indicates that the slope of T_w to R_p is a curved line since the slope depends on the value of R_p. Further analysis of the slope and of Equation 22 also shows that the curve relating T_w to R_p curves downward at a diminishing rate. This means that as R_p increases (when R_a is held constant), T_w decreases but more

[29] On the nature of quadratic equation solving and for a good refresher of high school algebra, see H. Sommers, Living Mathematics Reviewed 29-38, 55-62, 105-12 (1943).

slowly than R_p does. The above complicated slope can be contrasted with the simple and informative slope of the relation between T_w and T_p using Equation 21. That common-sense slope tells us that as T_p decreases one unit, T_w decreases by the size of the backlog where there is one full-time trial judge and no dropouts. To further relate Equation 22 to Equation 21, $1/T_p$ may be substituted for R_p, and $1/T_a$ for R_a in Equation 22. The result after simplifying would be:

$$T_w = T_p^2/(T_a-T_p). \qquad (30)$$

That equation follows the queueing theory assumptions but expresses waiting time in terms of processing and arrival times, rather than processing and arrival rates. The slope of waiting time to processing time using that equation is $[2(T_a)(T_p) - (T_p^2)]/(T_a-T_p)^2$, which is still more complex to understand or apply than simply saying the slope of T_w to T_p is N_b, as is done in the common-sense approach. The additional complexity might be worthwhile if it meant additional accuracy in fitting reality, but queueing theory as applied to the courts may be both more complex and less empirically valid.

With the above queueing theory analysis, waiting time can be reduced from .30 days to .10 days either by reducing the arrival rate from 3 cases per day to 1.67 cases per day, or by increasing the processing rate from 5 cases per day to 7.18 cases per day. As with the common-sense analysis, one could attempt to determine the lowest possible cost for each of those two alternatives, and then pick the least expensive alternative. A more appropriate solution that partly reduces the arrival rate and partly increases the processing rate might involve simultaneously solving for R_a and R_p in equations like the following:

$$.10 = (R_a/R_p)/(R_p-R_a).$$

$$\$10/(R_p-R_a)^2 = [\$20(R_a)(R_a-2R_p)]/[R_p^2(R_p-R_a)^2].$$

The first equation can be solved for numerical values for R_a and R_p that will yield a waiting time of .10 days. The second equation assures numerical values of such a nature that there will be the same marginal rate of return from dollars allocated to decreasing the arrival rate as from dollars allocated to increasing the processing rate. Equation 2 takes into consideration that 10 monetary units are needed to obtain the benefits of a one-unit decrease in the arrival rate, and 20 monetary units are needed to obtain the benefits of a one-unit increase in the processing rate. Equation 2 also takes into consideration the slope or MRR of T_w to R_a, and T_w to R_p. The second equation can be made more realistic by recognizing that the cost of decreasing R_a probably goes up as more of an increase is demanded. It would then be necessary to determine statistically or by accounting the relation between R_a and $\$R_a$ which may be of the form: $R_a = a(\$R_a)^b$. Substituting that expression wherever R_a or R_p appears in the above pair of equations would provide a result for $\$R_a$ and $\$R_p$, rather than R_a and R_p.

A similar analysis could be applied with the common-sense Equation 21, but there, since there are five unknowns, five equations must be solved simultaneously. Those five equations might be like the following:

$$.10 = (1-S)(N_b)(T_p)/(CJ).$$
$$\$5(N_b)(T_p)/(CJ) = \$8(1-S)(T_p)/(CJ).$$
$$\$8(1-S)(T_p)/(CJ) = \$4(1-S)(N_b)/(CJ).$$
$$\$4(1-S)(N_b)/(CJ) = \$7(1-S)(N_b)(T_p)/(J)(C^2).$$
$$\$7(1-S)(N_b)(T_p)/(J)(C^2) = \$9(1-S)(N_b)(T_p)/(C)(J^2).$$

The first equation seeks numerical values for S, N_b, T_p, C, and J that will yield a waiting time of .10 days. The next four equations say to set the MRRs of S, N_b, T_p, C, and J equal to each other if possible so that nothing will be gained by shifting dollars from one activity to another. In other words, a solution is sought that will achieve a desired T_w goal level while minimizing expenditures across the relevant activities. An alternative would be to develop a set of equations that will minimize waiting time while exactly spending a given budget. The first equation in that set would have a form like: $\$R_a + \$R_p = \$1{,}000$. That perspective may apply in many allocation problems, but not so well here because court budgets do not have as their categories S, N_b, T_p, and C, but rather categories like secretaries, supplies, rent, heat, etc., which are difficult to relate to waiting time.[30]

C. *Evaluating the Alternative Approaches*

We have discussed three methodological approaches to predicting court-case time-consumption, namely, prediction through (1) inductive statistical analysis, (2) simple deductive logic from processing times and backlogs, and prediction through (3) deductive queueing theory from processing rates and arrival rates. There are advantages and disadvantages between statistical and deductive approaches, and between the two deductive approaches of queueing theory and the common-sense approach.[31]

The statistical approach has the main advantage of being more data-based than a deductive approach. Both approaches are data-based in the sense that deduction involves some empirical experience with the real world from which one draws out some simplifying assumptions or premises from which various conclusions can be deduced. Deductions may also involve statistical premises such as the numerical predicted-value of T_p. The statistical approach, however, draws its conclusions directly from data, rather than indirectly by way of data-based premises. Being data-based means the statistical approach can more easily improve upon its predictions by constantly trying to reduce the deviations between actual and predicted scores. Deductive approaches tend to emphasize improving on internal consistency and comprehensiveness by considering other variables and relations, rather than external consistency with empirical data. The statistical approach also has the advantage of making fewer assumptions, especially about extraneous variables being

[30] On using a computer to solve simultaneous equations which optimally allocate scarce resources, see C. McMillan, Mathematical Programming: An Introduction to the Design and Application of Optimal Decision Machines (1970).

[31] On the relative merits of statistical and deductive prediction and related issues, see A. Kaplan, the Conduct of Inquiry: Methodology for Behavioral Science (1964); D. McGaw & G. Watson, Political and Social Inquiry (1976).

held constant. Both approaches, though, do make assumptions. The statistical approach, for example, tends to assume that good prediction minimizes the sum of the squared deviations between actual and predicted scores, rather than the absolute deviations. The statistical approach also tends to assume accuracy of measurement, representative samples of data, meaningfulness in statistically controlling for overlap among the predictor variables, and sometimes linear or at least one-directional relations.

Being data-based also generates the main disadvantage of the stastistical approach, namely that the predictions and relations are often distorted by variables that cannot be controlled for in the real world. For example, one may find in a large sample of courts or time periods that when a large amount of money is spent for time reduction ($), there is greater time consumption (T_t). An ordinary statistical analysis would not be so capable of distinguishing between the effect of $ on T_t, which should be negative, and the effect of T_t on $, which should be positive in the sense that an increase in a problem stimulates expenditures to deal with the problem. Likewise, the relation between $ and T_t may be heavily influenced by the fact that both variables are positively stimulated by urbanization, industrialization, and related variables that are often difficult to control statistically. Deductive analysis is able to deal with those problems by extracting from reality and making statements that explicitly assume other variables are held constant. Deductive analysis is, however, not inherently simpler than statistical analysis, since some deductive mathematical models (like queueing theory with multiple judges, reneging, and multiple stages) can be more complex than simple linear-bivariate statistical-equations.

Although the queueing theory and common sense approaches may be partly in conflict, the statistical and deductive approaches tend to work well together. Statistical reasoning is highly deductive. For example, if statisticians were attempting to determine whether the processing time of bench trials is shorter than the processing time of jury trials, they might reason as follows: If the difference can be attributed to chance more than 5 times out of 100 (given the size of the difference and the size of the samples), the assumption is that the difference is a chance difference due to the sample of bench trials and jury trials, rather than to a real difference, so that if a difference that could be attributed to chance with a probability of .15 is observed, the conclusion is that the difference is due to chance and not to bench trials really being shorter than jury trials. Deductive reasoning is not necessarily highly statistical, but it can almost always benefit from statistical analysis in such ways as:

1. Providing the data for the empirical premises of the deductive models, such as data on processing or arrival rates for the queueing theory model.
2. Testing the conclusions which the deductive models reach concerning predicted waiting times, and often provide insights concerning relevant variables that the models may have excluded or inadequately handled.
3. Providing opinion surveys to determine (a) what goals should be achieved, such as the desired waiting time, (b) the monetary

value of reducing delay, so as to aid in allocating budget resources, and (c) the relative weight of time saved to accuracy achieved or to other goals that cannot be easily expressed in dollars.

4. Testing the assumptions of deductive models, such as the queueing theory assumption that arrival and processing rates are independent of each other.
5. Testing for causation including reciprocal causation between time consumption and backlogs.

The main advantage of a queueing theory deductive approach over the common-sense approach is that the queueing theory approach enables predictions further back in time or in what may be called the predictive funnel. Figure 1 shows total time at the right end of what roughly looks like a funnel on its side. The common-sense approach bases its predictions on variables that are close to total time almost by definition, such as processing time and backlog, although the settlement rate and judge-time may also be involved in common-sense predictions. Queueing theory works with arrival rates and processing rates, which are further back in time. The situation is somewhat analogous to predicting voting behavior by asking voters before they enter the polling booth how they are going to vote, versus predicting from background characteristics like age, sex, and occupation, which occur prior in time to voting attitudes and behavior. Queueing theory also has a good track record for being helpful in reducing time consumption in situations like telephone congestion, tollgates, airline check-in counters, unloading of freight at platforms, supermarket cash registers, and various kinds of manufacturing or repair lines.[32]

The situations in which queueing theory has been especially helpful, however, help illustrate why queueing theory may not be so relevant to reducing time consumption in the courts. The waiting lines in the courts are different from those situations where queueing theory has worked well, unless queueing theory is defined so broadly as to include any kind of analysis of waiting lines. The main apparent disadvantages of queueing theory relative to the common-sense approach include such matters as:

1. Complexity.
2. Lack of intuitive validity.
3. The inconceivable ideas of infinite and negative waiting time and backlogs.
4. The idea that an arrival rate that exceeds the processing rate is just as bad as one that equals it.
5. The assumption of idle time, which may apply to a toll booth that has no cars waiting to be processed, but not to a judge who is not in trial.

[32] For success stories where queueing theory has helped in predicting waiting time, idle time, and other time consumption figures with varying arrival rates, service rates, and numbers of channels, see T. Saaty, supra note 2, at 302-58; A. Lee, supra note 2, at 93-206; and Gross, supra note 2, at 453-63. There does not seem to be any such success story in the queueing theory literature that involves the judicial process, although the judicial process has been beneifically studied from the perspective of related methods such as flow-chart simulation See Chaiken, supra note 16.

6. Assuming an infinite passage of time, rather than a starting point and days passed.
7. Assuming that the arrival rate does not exceed the processing rate, when it generally does in court cases.
8. Not adequately considering dropouts.
9. Not adequately considering multiple judges who are not 100 percent available for trial work or other case processing.
10. Not adequately considering multiple stages.
11. Not adequately considering case order. By "not adequately" it is meant that those court-important concepts are only considered in unrealistic or overly complex ways.

Less important disadvantages of queueing theory relative to the commonsense approach include in random order such matters as:

1. An overconcern for total backlog, rather than waiting backlog.
2. Relating total backlog to waiting backlog by adding or multiplying the ratio of the rates, rather than subtracting the cases in process or the number of judges.
3. An overconcern for the probabilities of backlogs of a certain size.
4. Relating total time to waiting time by multiplying by the ratio of the rates, rather than subtracting processing time.
5. Relating waiting time to backlog by dividing by the arrival rate, rather than by multiplying by the processing time.
6. Assuming arrival and processing distributions that may not fit what occurs in the courts.
7. An overconcern for the importance of arrival and processing distributions, rather than averages. By "distributions" is meant the spread around the average and especially the shape of the spread.[33]

D. Research on Actual Cases

It is possible to obtain from the Interuniversity Consortium for Political and Social Research at Ann Arbor, Michigan, data on approximately 15,000 criminal cases which were filed in 1974 in Washington, D.C. This is data that is part of the Prosecution Management Information System (PROMIS) compiled by the Institute of Law and Social Research with funding from the Law Enforcement Assistance Administration. Working with that data reveals that on the average day, there were substantially more arrivals than dispositions,

[33] Sue Johnson mentions several characteristics of the judicial process that make queueing theory less applicable. See note 3 supra. She, however, emphasizes that there is variation in service rates, arrival rates, number of judges, quality of judges, and cases—not such a serious problem for any deductive system that talks in terms of averages and recognizes that there is variation around the averages. Queueing theory does do that. People who object to quantitative methods often argue that the social world is unpredictable because there is so much variation in it. If there were no variation, however, there would be nothing to predict or to explain. Likewise, Steven Flanders objects to attempts to model the judicial process on the grounds that the process is too complex to be modeled in view of multiple paths from start to end, difficult-to-measure variables, multiple variables, high discretion, and conflicting interests. See note 3 supra. The more complex the system is, however, the more it might benefit (in terms of understanding) from attempts to capture its essence through either deductive or statistical methods.,

even though dispositions included cases that were settled through dismissals and guilty pleas, as well as bench and jury trials. The explanation is simply that the total number of arrivals during the year exceeded the total number of dispositions, and thus the backlog in Washington, D.C., at the beginning of 1975 was greater than the backlog at the beginning of 1974. Perhaps the Washington courts hired more judges in 1975 in order to reduce that growing backlog and the growing waiting-time, which a growing backlog produces. Queueing theory equations, however, cannot be meaningfully applied to such court data where the average arrival rate is greater than the average processing rate. If one tallies the arrival rates for the 365 days to see how they cluster, the most obvious pattern is one of many cases filed on Monday, a middling number of cases filed on Tuesday through Friday, and only a few cases filed on the weekend. There may actually be many cases arriving on the weekend, but they are not officially filed until Monday. If one tallies the processing rates for the 365 days, the most obvious pattern is one of getting cases disposed of at the end of the week (which fits the pattern of agencies or offices that send out a lot of mail on Fridays), fewer cases processed at the beginning of the week, and almost none on the weekends. Those distributions may be helpful in knowing the extent to which the courts should hire weekend personnel in order to even out the workload, but they may not be very helpful with regard to queueing theory, since they do not fit the distributions which are assumed by either the standard or the more exotic queueing theory equations.[34]

A big defect in the Washington, D.C., data set is that no record was made for each case that went to trial as to how long the trial took. That means there is no direct way of determining the key variable of processing time for testing various aspects of the common-sense approach to time prediction. If the District of Columbia courts had only one judge who worked full-time doing nothing but trying cases, the processing time or trial time could be determined by calculating the reciprocal of the processing rate. Since the District of Columbia has many judges, and they average a relatively small percentage of time actually trying cases, the more appropriate formula for determining trial time from the processing rate would be CJ/R_p. By estimating how many judges were working on an average day in 1974, and what percent of their time was spent in trying cases, values for C and J could be approximated and used to further estimate T_p from R_p. It would, however, be preferable to know T_p directly in view of its importance in the common-sense analysis, and in order

[34] If one tallies the number of days when there were 0-9 arrivals, 10-19, 20-29, and so on up to 100-109, we find that the number of days in each of those equal intervals was 51, 2, 17, 38, 63, 73, 62, 37, 18, 2, and 2. Of the 51 days in the 0-9 interval, 28 had 0 arrivals, and 23 had 1-9. That distribution produces a curve with two peaks. One is in the 0-9 region, and probably represents weekend days. The other larger peak is in the 50-59 interval. It is part of a symmetrical normal curve that is not the kind of Poisson arrival distribution assumed by most queueing theories. Likewise, if one tallies the number of days when there were 0-9 cases serviced, 10-19, 20-29, and so on up to 100-109, we find that the number of days in each of those equal intervals was 75, 29, 16, 16, 31, 59, 75, 50, 10, 3, and 1. Of the 75 days in the 0-9 category, 15 had 0 cases serviced, and 60 had 1-9. That distribution again produces a curve with two peaks in approximately, but not exactly, the same places. One peak is in the 0-9 interval, but the second equally large peak is in the 60-69 interval rather than the 50-59 interval. It is also a roughly symmetrical normal curve, not the kind of exponential servicing distribution assumed by most queueing theories.

to test how accurately the CJ/R_p formula predicts T_p. The fact that trial time is not recorded on so important a data set (or for that matter on any of the data sets of the National Center for State Courts) indicates that people concerned with judicial process research and efficiency may not adequately recognize the importance of that variable.

Fortunately, there is one data set that does show trial time for both criminal cases and civil cases. It is data on federal cases from across the country compiled by the Administrative Office of the U.S. Courts and made available for special research through the Federal Judicial Center. The data is complicated to use and may take a long time to analyze thoroughly. It is, however, potentially useful for a number of purposes relevant to the problems involved in predicting and reducing court case time. One especially important use involves developing statistical equations for predicting trial time and other time consumption variables from the characteristics of the cases. That would include characteristics and statistical prediction methods like those discussed in Section II, above.

Another important use is testing how well the common-sense equations can predict by using the first half of a year to obtain the basic parameters and then applying them to making predictions for the second half of the year in order to check the accuracy of the predictions. The basic parameters would include S, N_b, T_p, C, and J. The settlement rate is simply the number of cases dismissed, withdrawn, and pleaded guilty divided by the total number of cases. To determine the average backlog, it would only be necessary to calculate the backlog for each day and divide by the number of days. Knowing the backlog for each day requires knowing the initial backlog on day one, the number of cases arriving each day by their dates, and the number of cases processed or disposed of each day. It may be necessary to identify the judges in order to determine how many there were on an average day or week and what percent of the day or week they spent in trying cases. The queueing theory equations can also be tested by using the first half of the year to obtain their basic parameters of average R_a and R_p and then applying them to making predictions for the second half of the year.

Such a data set can also be useful for determining the extent to which the arrival rates and the processing rates rise and fall together. Through a statistical prediction analysis, the numerical values of a, b_1, b_2, A, B_1, and B_2 can be determined in the following equations:

$$(R_a)_t = a + b_1(R_p)_{t-1} + b_2(R_a)_{t-1}.$$
$$(R_p)_t = A + B_1(R_a)_{t-1} + B_2(R_p)_{t-1}.$$

The first equation determines the relation between the arrival rate at time t, and the processing rate of the previous week, when controlling statistically for the arrival rate of the previous week. In other words, one year of case data can be treated in terms of 52 weekly time-periods. The second equation determines the relation between the processing rate at time t, and the arrival rate of the previous week, when controlling for the previous week's processing rate. In effect this is saying that there is at least a week's lag with regard to the effect on the arrival rate of changes in the processing rate and vice versa, although a month lag might be more appropriate. By controlling for the previous week's arrival rate in the first equation, we are in effect controlling for all the variables that influence the arrival rate other than the processing rate. Likewise the second equation controls for all the variables that influence the processing rate other than the arrival rate. If the b_1 slope is substantially different from 1, the processing rate does influence the arrival rate. Likewise if B_1 is substantially different from 0, this shows the arrival rate also influences the processing rate. The relative sizes of b_1 and B_1 indicate the direction in which the influence is greater. A similar statistical analysis could be done with the relation between total time (T_t) and backlog (N_b).

Another important use to which a data set like that can be put is to test the common-sense optimizing equations 23-27 to see how much S, N_b, T_p, C, and J would have to change in order to reduce waiting time by 10%, 50%, or some other reduction figure. Likewise, the queueing theory optimizing Equations 28-29 can be tested to see how much R_a and R_p would have to change in order to reduce waiting time by the same percentage. The relative cost for those changes in S, N_b, T_p, C, and J, or the changes in R_a and R_p, would determine which variable or combination of variables would achieve the desired waiting-time reduction at a minimum cost. Performing that part of the analysis would be an article in itself, and possibly quite relevant to obtaining a number of insights into the applicability of optimization analysis to the judicial process.

The data set can also be used to test the effects of reordering the cases on a delay reduction. That use might involve the following steps:

1. Determine through a statistical inductive analysis the relation between processing time and the characteristics of the cases which should yield an equation of the form: $T_p = a + b_1X_1 + \ldots + b_nX_n$.
2. Apply that equation to obtain a predicted T_p score for every case.
3. For each of the 52 weeks, arrange the cases that arrived that week in the order of their predicted processing times.
4. Assign trial dates to the cases in the order of their processing times with the shorter cases first, rather than on a first-come, first-served basis. That will mean interchanging the trial dates of the cases that went to trial in order to fit the principle of the shorter cases first.
5. Recalculate the waiting times and the total times for the cases in light of the time-prediction equations. This will require a computer program that is capable of calculating predicted waiting times and total times through the use of equations like 1 and 3, but with an expanded concept of a processing time that includes all the processing time (T'_p, i.e., T_p primed), not just the trial processing. To calculate that figure for each case might involve multiplying T_p by a constant like 20, just as personal injury lawyers tend to multiply out-of-pocket medical costs by about 5 in order to predict damage awards.
6. After making the calculations of the waiting and total times, ob-

serve how much the recalculated times differ from the actual times. A fairer comparison might involve (1) determining the predicted waiting and total times with the original court dates, and then (2) determining the predicted waiting and total times with the interchanged court dates. That way the same time-prediction method is applied to both sequencing methods.

7. Vary the system to group the cases on a two-week basis, rather than a one-week basis, or a monthly basis.
8. Introduce as a constraint, maximum waiting time. The constraint is binding unless a postponement is granted to one of the parties. Vary that parameter.

This type of analysis should provide many insights into the implementation problems of optimum sequencing, its time-saving benefits, and its costs in terms of longer cases being made to wait longer for the good of the cases in general. The analysis can provide the contents of another potential article, but one that first requires doing the kind of comparison of time-prediction methods which the present Essay provides. The sequencing analysis is a good example of using simulation to test some aspects of a potential judicial reform. The computer simulation can provide insights without subjecting any criminal defendants or civil litigants to being guinea pigs as part of an actual experiment. After the simulation has aided in developing the system, it can then possibly be applied to actual cases that have not yet been decided, as contrasted to actual cases that have already been decided.[35]

All of the above testing could be applied to appropriate data sets for criminal and civil court cases at various levels of government, and to cases involving disputes that are resolved before a third party in administrative agencies. What may be needed now is the kind of data-based testing which has been suggested above to determine such matters as (1) the feasibility of statistically predicting processing time, (2) the ability of the common-sense equations and the queueing-theory equations to predict waiting time, (3) the nature of the causal relations among the variables relevant to time consumption, (4) the meaningfulness of the optimizing equations as applied to judicial delay reduction, and (5) the effects on delay reduction of rearranging the trial dates of cases to givė shorter cases more priority. That kind of testing may be of substantial benefit in making the courts more efficient. It may also improve time-prediction, causal analysis, optimization analysis, and simulation as tools in analyzing the effects of alternative legal policies. Those tools can help make the legal process more efficient, effective, and equitable than it has been.

[35] On optimum sequencing as applied to situations other than the courts, see K. Baker, Introduction to Sequencing and Scheduling (1974); and R. Conway, W. Maxwell & L. Miller, Theory of Scheduling (1967).