Dev. Lit.

ACCOUNTING FOR CASELOAD: A SIMPLE MODEL*

Robert C. Luskin
Mary Lee Luskin

Indiana University
February, 1984

P.O.B. 1847
Bloomington, IN. 47402

82-IJ-CX-0041

Abstract

In this paper, we propose and estimate a simple distributed lag model to explain a court's pending caseload as a multiplicative function of the number of new filings and the average case processing time. Using data from Detroit's Recorder's Court, we estimate the model on a time-series of 24 monthly observations. As it turns out, the model as estimated fits the data very closely. Given its explanatory success, we suggest that it has both theoretical and practical utility--the first in specifying just how variations in new filings and case processing times combine to affect caseload, and the second in helping courts (and others) predict the magnitude and timing of the effects of such changes as may occur or be expected to occur in either new filings or mean processing time or both.

Among the most salient of the variables commonly used to characterize and explain the operations of courts is the number of cases pending, or caseload. Descriptively, it is a measure of the burden of work the court has to shoulder (in which role, it is often normed against the resources available--principally, the number of judges hearing cases--or weighted by complexity, as defined by types of cases). Other uses are evaluative. Caseload is a frequent criterion of court performance. When courts seek to demonstrate their accomplishments (as, for example, to legislatures), they commonly point to graphs showing decreases in the "backlog" of cases, which is to say, caseload. If, on the other hand, the backlog is on the rise, it may be cited as evidence of the inadequacy of present resources relative to demand, and hence of the need for more resources.

Still other uses are explanatory. Caseload has long been presumed to have a major impact on various other aspects of court performance and behavior, including the attractiveness of plea bargains offered, the proportion of cases ending in pleas (and on the other side of the same coin the proportion of cases going to trial), and the length of time it takes the court to process a case. The actual evidence is mixed,[1] and much of it methodologically flawed. Many of the relevant studies are entirely or essentially bivariate, many of the samples are hopelessly small, the dependent variable is not always the number of cases actually pending, and so on. But enough of the results are positive, and positive enough, to suggest that caseload does have some, at least, of the effects traditionally ascribed to it--if perhaps not all of them, or not under all conditions, or not to the degree once generally thought.

Explaining Caseload

That is as much as we need say about caseload's effects on other variables[2]--enough merely to establish that it is a variable worth looking at, and its own antecedents worth looking into. Here, our concern is with the explanation and prediction of caseload, not the other variables it may help to explain and predict in turn.

In a rather distant way, caseload is of course a function of the legal system and the social, economic, and technological environment in which it operates, since they affect both the rates at which people engage in various behaviors and the behaviors that get defined as criminal or actionable. In short, these factors affect the potential inflow of cases, and thus, other things being equal, the number of cases on the docket. Previous attempts to account for caseload have drawn upon variables of this general sort (Goldman et al., 1976, and Casper and Posner, 1976). But here we shall focus on more proximate causes. We take the number of new cases arriving as given, ignoring the prior variables that may influence caseload through it.

The Model

At this level of proximity, there are exactly two variables on which caseload (abbreviated hereafter as C) depends. One, as we have already indicated, is the number of cases arriving (call it A). The other is the rate at which the court has managed to dispose of the cases arriving (call it P, for processing time). The effect of each should be spread over the recent past. The greater the number of recent arrivals, and the less rapidly they are handled, the larger the accumulation of cases pending should be. Moreover, these effects should be nonadditive.[3] The smaller the number of cases arriving, the less it should matter how fast they are processed, and the faster they are processed, the less it should matter how numerous they are. At the extremes, if no new cases arrive, the rate at which they are (or would be) processed is irrelevant, and, if all new cases are processed instantaneously, the number arriving is irrelevant.

This is more than merely plausible. The relationship is almost an accounting one, giving the "Accounting" of the title, which is used there in the extended and more usual sense of explanation, something of its literal meaning as well. Let us denote the caseload at the end of some given time-period t--for consistency with the analysis below, let us say a given month, although it could as easily be a day, a week, or a quarter--by C_t. We may suppose that we have data on series of T time-points in all, so that t=1,2,...,T. Similarly, we may denote the number cases arriving during any previous month, say, the i^{th} one before, by A_{t-i}, where i=0,1,...,t. Finally, let us suppose--unrealistically, of course--that every case arriving during the $(t-i)^{th}$ time-period takes the same time to be processed. Let us denote that time as P_{t-i}.

Now consider what would happen to the caseload C_t if the number of cases arriving during the i^{th} prior month A_{t-i} were increased by some number--call it ΔA. If $P_{t-i} < i$ (where i is the time difference between t and t-i), all the cases arriving in month t-i will have passed out of the system by month t regardless, so that the addition of ΔA or any other number of cases makes no difference to C_t. On the other hand, if $P_{t-i} \geq i$, the cases introduced in month t-i will still be on the docket at the beginning of month t. In that event ΔA additional cases arriving in month t-i will increase C_t by ΔA cases. In short, there is either a zero or a one-to-one increase in C_t, depending on the processing time P_{t-i}.

Next consider the effect of a change in P_{t-i}, say, by ΔP months. If $P_{t-i} < i \leq P_{t-i} + \Delta P$, so that cases introduced in month t-i would have passed from the docket by the beginning of month t before the increase but would still be on it after, the increase would augment C_t by A_{t-i}, the number of cases involved. But if $P_{t-i} \geq i$, so that the cases introduced in month t-i would still be on the docket at the beginning of month t either with or without the increase in P_{t-i}, or if $i \leq P_{t-i} + \Delta P$ (which implies $i < P_{t-i}$), so that cases introduced in month t-i would be over and done with either with or without the increase, the effect of the increase on C_t is obviously 0. In short, the effect of variations in prior processing times is either zero or the number of cases involved (A_{t-i}), depending on what the processing time is before it is varied (P_{t-i}) and on the size of the variation (ΔP).

These relationships are instructive, but not really useful. Even if the assumption of uniform P_{t-1} were accurate, they would only be useful postdictively, as a means of assessing the hypothetical impact of variations in past arrivals or processing times. Since we cannot know how long cases presently arriving will take, we cannot generate predictions as to what would happen to the caseload if they were more or less numerous or took longer or shorter to complete. Another problem is that for a given lag i the effects of variations in P_{t-i} and A_{t-i} on C_t will be different for each t, which makes for unparsimonius and unwieldy explanation. Finally, cases introduced in a given month do <u>not</u> all have the same processing time. To allow processing times to vary would make the accounting unwieldier still. But not to allow it makes the accounting inaccurate.

Perhaps the best solution, if we are interested in prediction and more concise explanation, is to substitute the <u>average</u> processing time in the $(t\text{-}i)^{th}$ period (call it $\overline{P}_{t-i}$) for P_{t-i} and to develop a model that essentially averages the effects of both $\overline{P}_{t-i}$ and A_{t-i} over a number of time-periods, i.e., over a number of values of t. How much, on the average, can C_t be expected to increase when A_{t-i} increases by X cases, or when $\overline{P}_{t-i}$ increases by Y days? A model that can tell us that can enable us to predict and understand variations in caseload in the near future.

This argument suggests a "distributed lag" model. Clearly, the effects of both A and $\overline{P}$ must occur either "contemporaneously" (given that C_t is defined as of the end of the month) or with some lag. Clearly, too, the effects must be spread or distributed over a number of lagged observations. The number of arrivals will matter for several previous months, and so will speed with which they are processed.

Together, these considerations lead us to a distributed lag model in which the variables A and P are combined multiplicatively. Specifically, the model is

$$(1)\quad C_t = \alpha + \beta_0 A_t P_t + \beta_1 A_{t-1} P_{t-1} + \ldots + \beta_M A_{t-M} P_{t-M} + u_t \ ,$$

where α and the β's are unknown parameters to be estimated, M is the lag before which A and $\overline{P}$ have no effect, and u is an unmeasured disturbance summarizing the causative factors of which the model takes no explicit account.[4] (We assume that they are uncorrelated with the $A_{t-i}P_{t-i}$, so that their having been omitted is not an obstacle to estimation.)[5]

Under this simple model, the effect of the number of arraignments in the $(t\text{-}i)^{th}$ month is

$$(2)\quad \beta_i P_{t-i}$$

and the effect of the mean processing time in the same month is

(3) $\beta_i A_{t-i}$,

where, in both cases, $i=0,1,\ldots,M$. Provided that $\beta_i > 0$, each quantity is both positive and an increasing function of the value of the other variable in the same month.[6]

The Data

The data here come from Detroit Recorder's Court--the municipal criminal court of Detroit--and span the two-year period from April, 1976 through March, 1978. Among other things, the data include the number of initial arraignments ("on the arrest warrant") each month and the number of cases (of defendents, actually) on the docket as of the end of the month, both taken directly from pretabulated court records. For our purposes, however, the data from the records are not sufficient. The caseload, as Recorder's Court reckons it, includes only those cases which have made it past the preliminary examination. This means that the "arrivals" variable is most appropriately the number not of the initial arraignments "on the arrest warrant" but of the post-preliminary examination arraignments "on the information." But one deficiency of the pretabulated records is that they afford no count of the latter. Another is that they do not include anything in the way of case processing time.

Nonetheless, A and $\overline{P}$ can be estimated. The data also comprise a (random) sample, stratified by month, of all of the cases begun within the two-year period, and two of the case-wise variables recorded are the length of time to completion and whether or not the case is disposed of before or at the preliminary examination. Since the sample size is adequately large--about 85 cases each month--we may readily compute estimates of the monthly mean processing time and of the number of arraignments each month, the former directly, and the latter as the number of arraignments on the arrest warrant times the sample proportion surviving the preliminary examination.

We end up with a monthly time-series of 24 months. The variables' minima, means, and maxima over this period (and over the somewhat shorter period on which we actually estimate the model--see below) are given by Table 1. In passing, we may note that this period saw the introduction of a number of structural innovations that were specifically designed to, and did, reduce processing times, with the result that both processing times and caseloads were as a rule substantially lower toward the end of the period than toward the beginning. Thus, since the sample on which we estimate the model consists of the last 17 observations, the means and minima of C and $\overline{P}$ are lower there than in the sample as a whole.

Table 1 About Here

Estimation

The estimation of equations such as (1) is typically hindered by the presence of extreme collinearity among the lagged values of the explanatory variable: in this model, the product term $Z_{t-i} \equiv A_{t-i}P_{t-i}$, which, for purposes of estimation, is most conveniently treated as a single variable at each i. The variables Z_t, Z_{t-1}, ... Z_{t-M} are in general so many and so similar that it is impossible to distinguish their individual effects very well. But by adopting some simplifying assumptions about the β's, we can reduce the number of regressors and the collinearity among them. Here, we assume that the β's can be sufficiently approximated by a polynomial in the lags (in i) of some pre-specified degree. This assumption, which results in what is known as a "polynomial" or "Almon" lag scheme (after

Almon, 1965; see also, e.g., Johnston, 1972 or Pindyck and Rubinfeld, 1981), has the advantage of being relatively flexible, in that it permits various patterns of effects over varying lags.

Admittedly, the results may to some extent depend on several additional and mostly nontheoretical specifications: the length of the maximum effective lag M, the degree--call it F--of the polynomial that approximates the β's, and the further restrictions, if any, that are placed on the latter. Of necessity, the choice of M and F is generally made on empirical grounds. First, M is generally set so as to maximize the proportion of variance explained, but without excessive cost in degrees of freedom, collinearity, or plausibility of results. Here, the proportion of variance explained is maximized at M=6, but is only trivially (.004) lower at M=7, where the collinearity is substantially lower and the precision of the estimates correspondingly greater. Thus we set M=7. Once given M, the choice of F is a matter of whether the necessarily higher proportion of variance explained with successively higher F's is statistically worthwhile, a criterion that leads here to F=2. We may note that F=2 compels the pattern of effects over i to be either flat or more or less U- or inverted-U-shaped, with at most one turning point.

That leaves the question of whether to constrain the β's, and, if so, how. A common practice is to set β_{-1} and/or β_{M+1} equal to 0--which in the present model would be to assert that the values of the explanatory variables in the future (at times t+1 and after) and/or their values more than M months in the past (at times t-(M+1) and before) have no effect on caseload. These have a certain intuitive appeal, but are also capable of exerting a heavy--critics say excessive--influence on the estimates of the other, nonzero effects. With F=2, the pattern of effects is forced to be symmetrical, with $\beta_i = \beta_{M-i}$ and the peak (or nadir) precisely in the middle. Thus, in the interest of letting the data speak more nearly for themselves, we leave the model unfettered. The one constraint we do adopt is less confining, and is in fact suggested by the data themselves. Without constraints, the estimate of β_0 is anomalously but insigificantly negative, the likeliest inference from which is that β_0 is some small positive number. Thus strictly as a means of tidying up the results, we do impose the one restriction that β_0=0. Since the unconstrained estimate is insignificantly different from 0 anyway, the effect on the rest of the estimates is slight. Indeed, no reasonable specification--of M, of F (for which F=2 is the only reasonable choice), or of the values of the β's (not even $\beta_{-1} = \beta_{M+1} = 0$)--produces results too greatly different.

The final choice to be made is of estimator. This is not an open-and--shut matter either. As always with time-series data, one cannot but suspect the disturbance of being autoregressive and should usually make statistical allowances if it is. Here, however, it is not clear whether disturbance is autoregressive, or if so in what way. The relatively small number of observations--the lagging of Z up to i=7 reduces the effective N from 24 to 17--makes such determinations difficult. The evidence of autoregression is weak and murky. The Durbin-Watson test for the first-order variety is inconclusive (DW=1.34). And although the correlations between (the ordinary least squares-generated estimates of) u_t and u_{t-i} are not tiny (averaging a bit below .3) and seem to display the damped sinusoidal pattern characteristic of second- or higher-order autoregression, they are neither individually nor collectively significant. Similarly, some of the partial autocorrelations are not really small, but none is significant either.[7]

Even if we were to conclude on this tenuous evidence that the disturbance is autoregressive, it would still be unclear what, if anything, to do about it. The usual remedy is to use generalized instead of ordinary least squares to estimate the equation (GLS instead of OLS). But the advantages of GLS are only asymptotic, emerging only as the sample size becomes infinitely large. Whether they would emerge here, given the limited number of observations and the apparent mildness of whatever autoregression there may be, is uncertain (see Rao and Griliches, 1969). Furthermore, the partial autocorrelation function rapidly runs out of degrees of freedom, making the specification of the order of the autoregression involved a more than usually risky business. If it is not accurately specified, the move from OLS to GLS may do more statistical harm than good. The results, in this instance--we have in fact seen them for both estimators--do not differ too dramatically.[8] But, given, as we have noted, that the data do not exactly cry out for a correction for autoregression, we are likely, we think, to do best by opting for OLS.

Results

The estimates we thus obtain are displayed in Table 2, along with their estimated standard errors[9] and the R^2.

Table 2 About Here

The first thing to observe is the size of the R^2. At .975, it could hardly be larger. The model as estimated explains the variation in caseload almost perfectly. True, the smallness of the sample does make explanation easier, but even the "adjusted R^2" (where the adjustment is in essence for the smallness of the sample in relation to the number of independent parameters)[10] is fully .972. This is high even for time-series data, though given the near-accounting relationships behind the model not entirely surprising.

Consider next the estimates of the coefficients.[11] In accordance with (2) and (3), these reveal the impact of any given change in the value of either of the explanatory variables, for any given value of the other, in each and any of the months preceding. If, for example, the number of arraignments is a constant 1000, the current caseload can be expected to increase by 45 cases for each day's increase in the mean processing time of the cases begun in the month before, by 78 cases for each day's increase in the mean processing time of the cases begun two months before, and so on. By the same token, the coefficients also reveal the effect of a change in the value of either of the explanatory variables (again for a fixed value of the other) in a given month on the caseload in the same or any subsequent month. Thus, if the number of arraignments is again a constant 1000, and the mean processing time in a given month were to increase by one day, the caseload could be expected to increase by 45 cases at the end of the next month, by 78 cases at the end of the month after that, and so on. Notice by the way, that the coefficients trace out an essentially inverted-U-shaped pattern as they vary over the length of the lag involved. The effect of each of the explanatory variables, given a constant value of the other, is smallest in both the most immediate and most distant past, reaching its peak roughly mid-way between.

In a sense each variable's effects can be summarized by caseload's "long-run response" to it. Adapting the usual definition to this multiplicative model, this is simply the amount of change that can be expected to occur in response to a <u>constant</u> unit change, at every effectual lag, in the value of the one explanatory variable, for a given, constant value of the other. Thus the long-run response to the number of arraignments is

$$(4) \qquad (\sum_{i=0}^{7} \beta_i)\, P,$$

while that to the mean case processing time is

$$(5) \qquad (\sum_{i=0}^{7} \beta_i)\, A,$$

where $\overline{P}$ and A (note the absence of a subscript) are constant values of the "other" variable. For a given value of the other variable, the long-run responses can be estimated by simply substituting estimated for actual β's in (4) and (5). This gives, as the long-run response to arraignments, (.0607) $\overline{P}$, and, as the long-run response to mean processing time, (.0607) A.

Let us consider some plausible numbers. Suppose, for example, that the number of arraignments increased by 100 in every one of the preceding months. This is only a 12 percent increase over the average monthly level in this court during this period, and is thus by no means so big as to be at all unlikely. If the average case processing time was, say, 44 days (which was the shortest we observed in the period we studied), the long-run response would be (100) (.0607) (44) = 267. In other words, we ought to expect an increase of 267 cases in the caseload as a result. This, from a practical point of view, is in the nature of a lower bound. Mean processing times much shorter than 44 days are possible but not likely. If, on the other hand, the court averaged as much as 158 days per case (the highest monthly average we observed), the additional 100 cases per month in the preceding seven months would result in an addition of 959 cases to the caseload.

Or, again, consider the long-run response to mean processing time. Suppose the mean processing time increased by 10 days per case in each of the seven months before. This is again a shift of roughly 12 (actually, 13) percent of the the mean over the entire period, and not unlikely to occur. If the number of arraignments each month were a constant 543 (the lowest number in this period), the result would be another 330 cases on the current docket. If, at the other end of the range of likely responses, the number of new cases each month were at its observed maximum of 1133, the result would be instead another 688 cases.

Still other estimates can be formed for other combinations of changes in the one variable and values of the other. But the main points to be made are, first, that the model enables one to form such estimates by simply plugging in the appropriate values of A and $\overline{P}$, and, second, that even very modest changes in A and $\overline{P}$ have a very substantial impact on caseload.

Summary and Conclusions

To sum up, then: we have developed and estimated a simple, theoretically appealing, and empirically successful model of caseload as a multiplicative, lagged function of the number of cases arriving and the speed with which they are processed. The effect of each is (a) substantial, though spread over a number of lags; (b) dependent, at a given lag, on the value of the other at the same lag; and (c) at its peak in the middle temporal distance.

The major practical use of this model is to generate predictions of caseloads. Given estimates of the parameters, one need only plug in the actual or anticipated (or feared or hoped-for) mean processing times and numbers of new cases in each of a series of seven consecutive months in order to project the caseload at the end of the eighth.[12] In similar fashion, the increments or decrements that would occur in response to changes in the number of arraignments or mean processing time can also be estimated. This is true both of transient changes occurring in only a single month and of long-term changes occurring over any subset of the preceding months. For any such change(s), it is possible to trace the time-path of the resulting changes in caseload, and thus to see when it peaks and what it is at the peak. Such estimates should enable a court to anticipate its workload more accurately. And to control it, to the extent that it can control the number of incoming cases or (more likely) the time it takes to dispose of them.

Of course, we have estimated the model for one possibly atypical court only. In other courts, the number of lags over which both arraignments and processing times have their effects can be expected to differ. Similarly, the magnitudes of the effects will doubtless vary from court to court and even, perhaps, from period to period within this court. One would not want to use these data to make predictions for other courts, or even for this court too far in the future. But judging from the R^2, the model seems to approximate an averaging out of the underlying near-accounting relationships very nicely, which suggests that it should be predictively useful wherever appropriate data are available.

Notes

*Most of the data for this study were collected under the auspices of the American Judicature Society under a grant (No. 78-NI-AX-0076) from the National Institute of Law Enforcement and Criminal Justice. Additional data collection and the analysis of this article have been funded by a grant (No. NIJ-82-IJ-CX-0041) to the authors from the National Institute of Justice's Performance Measurement Program.

We want to thank David Kessler for providing computer assistance and Matthew Morey and Esfandiar Massoumi for discussing the model with the senior author and contributing helpful comments. The analysis and conclusions are of course our own and do not necessariy represent the views of the United States Department of Justice, or of any of the individuals aforementioned.

[1] See, on caseload's relation to plea bargaining and the relative frequency of trials, Feeley (1979), Heumann (1975, 1978, and 1979), Nardulli (1979), Rhodes (1978), and Hausner and Seidel (1979); on its relation to the decision to prosecute, Rhodes (1976); on its relation to bail-setting policy, Fleming (1979) and Feeley (1979); on its relation to sentencing, Feeley (1979) and Nardulli (1979); on its relation to court "productivity," Gillespie (1976); and on its relation to case processing time, Gillespie (1977), Church *et al*. (1978), Martin and Prescott (1981), and Luskin and Luskin (1983).

[2] Except for n. 5 below.

[3] Roughly speaking, the *effect* that one variable has on another is the amount of change that can be expected to occur in the one as a result of each unit of change in the other, other things being equal--i.e.,

again roughly, the partial derivative of the one with respect to the other (see Luskin, 1983). Accordingly, the effect of arraignments on caseload is the number of additional cases we can expect there to be on the docket as a result of each additional arraignment, and the effect of case processing time is the number of additional cases we can expect there to be as a result of each additional day the court takes, on the average, to process its cases.

[4]This model bears a strong resemblance to the queuing theory equation known as "Little's formula." Specifically, Little's formula is $L=\lambda W$, where L in this context denotes the expected caseload, λ the expected number of arraignments, and W the expected processing time. (See Hillier and Lieberman, 1980, eg.) Plainly, L corresponds to C, λ to A, and W to $\overline{P}$. But there are differences too. L, λ, and W are in the nature of expected values, whereas C, A, and $\overline{P}$ are not, and partly because of that (1) is stochastic (i.e., includes a disturbance), whereas Little's formula is not. Further, Little's formula is derived--indeed L, λ, and W are defined--on the assumption of a "steady-" or equilibrium state. In (1), in contract, the time-invariant quantities L, λ, and W are replaced by the time-subscripted C, A, and $\overline{P}$, with the effects of the second two on the first allowed to vary with the time elapsed, and apportioned over a set of lags.

[5]We are of course aware that mean processing time may depend on the caseload as well as vice versa (see the discussion in Luskin and Luskin, 1983). Given that C_t indicates the caseload at the end of the t^{th} month, P_t may be a function, among a good many other things, of C_{t-1}. Nevertheless, we consciously ignore this additional equation, and thus avoid the complications that the combination of multiple equations with

lagged endogenous variables and autoregressive disturbances, not to mention nonadditivities in the endogenous variables, would bring. We feel justified and tolerably safe in doing so because the dependence of processing times on caseloads seems to be slight (Luskin and Luskin, 1983), with the result that the system consisting of both equations is practically recursive.

[6]Admittedly, these effects and the equation that implies them do not quite capture all the subtleties of the near-accounting relationships described above. But to judge from the R^2 below, the approximation of an averaged-out version of them must be close.

[7]The .3 figure and the tests of significance are based on the auto-correlations between u_t and u_{t-1} through u_{t-4} only, in keeping with the rule-of-thumb of considering only the first N/4 elements of the series (Hibbs, 1974). For larger i, the pattern is much the same, however. For more on these sorts of diagnostics, see Hibbs again.

[8]The GLS R^2 (defined as the squared correlation between the predicted and actual values of caseload) differs by only .007, and the GLS and OLS estimates of the coefficients show broadly similar profiles. The biggest differences are that the GLS estimates rise and then fall a bit more sharply with increasing lags and that the GLS estimate for $_7$ is substantially smaller (in fact, insignificant at the .05 level). The long-run response is a trifle larger, at .0639.

[9]Estimated as in Johnston (1972).

[10]For this model, F+1 less the number of additional constraints on the β's: here, since we constrain $\beta_0=0$, $(F+1)-1=F=2$.

[11]Despite its size, the estimate of the "constant term" (about which we shall not bother to comment apart from this note) is a small fraction of its standard error, and thus statistically indistinguishable

from zero. This, too, is an attractive result, since, under the purely hypothetical scenario in which the court either received no additional cases or processed all the cases it received instantaneously in each of the previous months, we should expect the caseload, which in that event would simply be α, to be 0. Not that it would tell very much against the model if $\alpha \neq 0$. That would merely mean that the actual regression hypersurface bent toward the origin as it approached it, and hence away from the regression hyperplane of our linear model--i.e., that the model did not apply so far outside the range of values we actually observe.

References

Almon, Shirley (1965). "The Distributed Lag Between Capital Appropriations and Expenditures," 33 Econometrica 178.

Casper, Gerhard and Richard A. Posner (1976). The Workload of the Supreme Court. Chicago: American Bar Foundation.

Church, Thomas W. Jr., et al. (1978). Justice Delayed: The Pace of Litigation in Urban Trial Courts. Williamsburg, VA: National Center for State Courts.

Feeley, Malcolm (1979). The Process is the Punishment: Handling Cases in a Lower Criminal Court. New York: Russell Sage Foundation.

Fleming, Roy B. (1979). "Punishment Before Trial: A Political Choice Model of Policy Changes in Pretrial Sanctioning," in Peter F. Nordulli (ed.), The Study of Criminal Courts: Political Perspectives. Cambridge, Mass.: Ballinger.

Gillespie, Robert W. (1976). "The Production of Court Services: An Analysis of the Scale Effects and Other Factors," 5 Journal of Legal Studies 243.

__________ (1977). Judicial Productivity and Court Delay: An Exploratory Analysis of the Federal District Courts. Washington, D.C.: National Institute of Law Enforcement and Criminal Justice.

Goldman, Jerry, et al. (1976). "Caseload Forecasting Models for Federal District Courts," 5 Journal of Legal Studies 201.

Hausner, Jack and Michael Siedel (1979). An Analysis of Case Processing Time in the District of Columbia Superior Court. Washington, D.C.: Institute for Law and Social Research.

Heumann, Milton (1978). "A Note on Plea Bargaining and Case Pressure," 9 Law and Society Review 515.

_________ (1978). Plea Bargaining: The Experience of Prosecutors, Judges, and Defense Attorneys. Chicago: University of Chicago Press.

_________ (1979). "Thinking About Plea Bargaining," in Peter F. Nardulli (ed.), The Study of Criminal Courts: Political Perspectives. Cambridge, Mass.: Ballinger.

Hibbs, Douglas A., Jr., (1974). "Problems of Statistical Estimation and Causal Inference in Time-Series Regression Models," in Herbert L. Costner (ed.) Sociological Methodology, 1973-1974. San Francisco: Jossey-Bass.

Hillier, Fredrick S. and Gerald J. Lieberman (1980). Introduction to Operations Research, 3rd ed. San Francisco: Holden-Day.

Johnston, J. (1972). Econometric Methods, 2nd ed. New York: McGraw-Hill.

Levin, Martin A. (1975). "Delay in Five Criminal Courts," 4 Journal of Legal Studies 83.

Luskin, Mary Lee and Robert C. Luskin (1983). "Case Processing Time in Criminal Courts: Issues of Explanation and Reform," unpublished manuscript, Department of Forensic Studies, Indiana University, Bloomington, IN 47401.

Luskin, Robert (1983). "Effects in Structural Equation Models: Issues of Definition, Interpretation, and Specification," unpublished manuscript, Department of Forensic Studies, Indiana University, Bloomington, IN 47401.

Maddala, G. S. (1977). Econometrics. New York: McGraw-Hill.

Martin, John A. and Elizabeth A. Prescott (1981). Appellate Court Delay: Structural Responses to the Problems of Volume and Delay, Michael Hudson, ed. Williamsburg, VA: National Center for State Courts.

Nardulli, Peter F. (1979). "The Caseload Controversy and the Study of Criminal Courts," 70 The Journal of Criminal Law and Criminology 89.

Pindyck, Robert S. and Daniel Rubinfeld (1981). Econometric Models and Econometric Forecasts (2nd ed). New York: McGraw-Hill.

Rao, Potluri and Zvi Griliches (1969). "Some Small Sample Properties of Several Two-Stage Regression Methods in the Context Autocorrelated Errors," 64 Journal of the American Statistical Association 253.

Rhodes, William (1976). "The Economics of Criminal Courts: A Theoretical and Empirical Investigation," 5 Journal of Legal Studies 311.

__________ (1978). Plea Bargaining: Who Gains? Who Loses: Washington, D.C.: Institute for Law and Social Research.