FJC Staff Paper

APPELLATE COURT CASEWEIGHTS PROJECT

Federal Judicial Center

FJC-SP-77-3

A STAFF PAPER IS THE PRODUCT OF A SHORT-TERM RESEARCH EFFORT BY CENTER STAFF. GENERALLY UNDERTAKEN IN RESPONSE TO QUERIES FROM A JUDICIAL CONFERENCE COMMITTEE, MEMBERS OF THE JUDICIARY, OR FROM THE CENTER BOARD OR DIRECTOR, A STAFF PAPER NORMALLY INVOLVES LESS EXHAUSTIVE RESEARCH METHODS THAN A CENTER REPORT. TOGETHER, STAFF PAPERS AND REPORTS ARE INTENDED TO GIVE AN OVERALL VIEW OF CENTER RESEARCH ACTIVITIES.

Appellate Court Caseweights Project

Research Division
Federal Judicial Center
June, 1977*

Please note: There is a typographical error in this document. The number VI was accidentally skipped when putting together the sequence of tables, but no text or table is missing from the report.

FJC-SP-77-3*

*This staff paper was prepared as an informal memorandum in November, 1976.

Table of Contents

Summary 1

I Description of the Project 7

II Results 10

III Analysis of Results 17

Conclusion 35

SUMMARY

The Appellate Court Caseweights Project was an attempt to develop an accurate and objective measure of caseloads in the United States Courts of Appeals. The utility of such a measure is that it would serve as a basis for equitable allocation of judicial resources to courts, or of cases to individual judges.

To the extent that caseloads of the United States Courts of Appeals have been measured in the past, such measurements have been based on what may be called "gross case processing volume." The number of cases filed per year or the number terminated per judgeship are common caseload measures based on gross case processing volume. It is well recognized, however, that such measures may offer inaccurate comparisons of the actual workloads of judges or of courts. While ten anti-trust cases typically would require far more work than would ten criminal appeals, caseload measures based on gross case processing volume do not recognize such differences. One obvious solution to this problem is to "weight" cases according to their difficulty (by determining, for instance, that an anti-trust case is the workload equivalent of four criminal appeals).

In 1974, the Federal Judicial Center initiated an Appellate Court Caseweights Project to develop a method for weighting appellate court cases and thus to produce weighted caseload measurements for the courts of appeals. This paper presents an analysis of the results of that project.

The Center's previous efforts at developing case weights were conducted in district courts and required detailed timekeeping by judges.

From the time records, the Center computed the total amount of judge time expended for given types of cases, and the total number of cases of each such "case-type," and thus determined relative weights for each "case-type." This timekeeping method imposed a substantial burden on participant judges, but the results did not seem to justify the imposition. The time-per-case spent on cases of the various case-types varied substantially between districts and between circuits, suggesting that the weights were not as accurate as might have been desired.

In the project here reported, the Center used a more direct method: it simply asked judges (from three courts of appeals) for their estimates of the relative workload, or burden, associated with each of 23 case-types.[1] Those estimates, together with the judges' estimates of the total time they spend working on cases in a year, are the data upon which this analysis is based.

Since the project obtained a set of burden estimates[2] from each of the three participating courts, the first step in the analysis was to compare, for each case-type, the three burden estimates obtained. This comparison showed rough agreement among the courts, but enough

---

[1] This set of case-types was the product of the combined efforts of an ad hoc panel of judges, personnel of the Center, and the project contractor. It was designed to: 1) include all cases brought before the courts of appeals, 2) be not so extensive as to make the estimation process a severe burden on judicial time, 3) have category labels which clearly indicated the types of cases included in each category, 4) assure that each category contained cases of similar burden, and 5) be susceptible to unambiguous translation into the categories in which the Adminstrative Office labels appellate cases. The case-types are listed in Table I (pp. 11-13).

[2] A set of burden estimates contains 23 numbers, each representing the burden, or case weight, ascribed to one of the 23 case-types.

disagreement to suggest that the accuracy of estimates was, at best, only slightly better than that obtained in the district court timekeeping study.

The second level of analysis was the comparison of the judges' estimates of their case-related workyear[3] with workyear hours computed from the burden estimates. The burden estimates yield figures for the number of judge-hours required to dispose of a typical case of each case-type. Given the number of cases of each case-type in the caseload of a particular court in a year, along with the burden estimates, it is a straightforward matter to compute the total number of hours such a caseload would require of each judge. The comparison of computed and estimated hours-per-judgeship-per-year showed only a rough level of agreement. Computed values for the typical workyear were generally higher (by as much as 98%) than the judges' estimates, which suggests that the judges over-estimated the time actually devoted to some or all of the case-types. While there is a strong argument that the explanation for this discrepancy lies not so much with the judges' inability to estimate as with the particular methodology employed, the conclusion is clear that the burden estimates are not valid measures of actual judge-time required by cases of the 23 case-types.

A third, more sophisticated level of analysis was then employed. The burden estimates were taken only as relative weights, with their

---

[3] That is, the total time they spend on case-related work in a year.

unit of measurement (hours), being ignored. Thus per-judge caseload measurements were computed in much the same manner as that mentioned above, but were used only for relative comparisons of workloads. For example, if a given caseload measurement computation yielded values of 3000 "units"[4] per judgeship for court A, and 2000 units per judgeship for court B, this was taken only as an indication that the workload in court A was 50% greater than that in B. By computing per-judge caseload measurements for each of the eleven courts of appeal, and then computing the average of those eleven values, standardized measurements could be obtained by expressing each value as a percentage of the average. Under this scheme, a court with a caseload measure of 100 was considered average in its per-judge workload, while a court with a measure of 150 had 50% more work than the "average" court. Since three sets of burden estimates were available (one set from each participating circuit), three different standardized caseload values were computed for each of the eleven courts.

The first result of this third level of analysis was that, for all but one of the eleven courts, the three standardized caseload values were in very close agreement. This meant that, despite the apparent differences among the three sets of burden estimates, they yielded very similar caseload value (hereinafter termed "weighted caseload values," since they employ the burden estimates as case weights).

---

[4] The meaning of those units (e.g., hours) being ignored.

The second, and more important result of this analysis was that these weighted caseload values also agreed very closely with unweighted standardized caseload values (based on gross case processing data, such as terminations per judgeship per year). The conclusion was clear: weighting had very little effect on caseload measurements.

The reason for the ineffectiveness of weighting -- as evidenced by the similarity of caseload measurements based on weighted and unweighted case terminations -- was not difficult to infer. If the caseload of a given court is broken down into categories based on case difficulty (i.e., by the amount of judge-time required by the case), that caseload can be identified with a "case difficulty distribution," such as 5% hard cases, 60% moderate cases, 35% easy cases. The reason for the ineffectiveness of case weighting thus appears to be that courts of appeals have very similar "case difficulty distributions."[5] If this is indeed true, then it can be seen in retrospect that case weighting could not be expected to have any significant effect. Weighting is useful in relative caseload measurement only if the caseloads to be measured have different "case difficulty distributions." If those distributions are all the same, then weighted and unweighted caseload measures are equally valid.[6]

The conclusion to be suggested is that, since the courts of appeals have such similar caseloads, caseload weighting cannot be expected to

---

[5] The D.C. Circuit is an exception.

[6] The crux of this thesis is that the concern is with relative caseload measurements. A simple analogy: if basket A contains 10 apples and 4 oranges (cases) and basket B contains 5 apples and 2 oranges, then no matter what the prices (weights) of apples and oranges may be, basket A will cost twice as much as B.

have a substantial effect on relative caseload measurements. Unfortunately, this conclusion can only be suggested, because the analysis has led to another, perhaps more important, conclusion: that the inconsistencies in appellate court statistical reporting are of sufficient magnitude that they render impossible any statistical analysis of the precision necessary to fully evaluate such matters as caseweights.

Appellate court caseload measurements must be founded on data relating to the volume of cases handled in those courts. If, for instance, such measurements are to be based on case filings (be they weighted or unweighted), then of course it is necessary to know how many filings each court has. But the definition of "filing" must remain constant. If the same caseload would in one court be counted as 1000 filings, but in another as 1500, then obviously any caseload measure based on filings would be quite misleading. Unfortunately, it appears that non-uniformity of that sort may exist in much of the data reported by the appellate courts. An intimate look at the data produced in the computer analysis for this project reveals several anomalies that can most optimistically be described as "curious."[7] A less than optimistic view of the data suggests that much of the appellate court statistical data (from Form JS-34 reporting) is unreliable.

The clearest and most important conclusion of this analysis is that appellate court statistical reporting must be reviewed and probably

---

[7] The reader is referred to the main text for a more thorough discussion (see pp. 32-35).

revised to assure that each reported case event (e.g. a filing) represents the same actual event in each circuit. Any use of appellate court caseload data must rest on shaky ground until the uniformity of that reporting is assured. This conclusion does not, however, lay to waste the efforts of the judges who participated in this project. At the least, their efforts have suggested a most significant result regarding case weighting in the courts of appeals; at best, their efforts may prompt a major improvement in the data base upon which much of the analysis of developments in appellate court administration must depend.

The main text of this paper provides a somewhat more detailed description of the project itself, and a more thorough discussion of the various analyses mentioned above.

I. Description of the Project

The purpose of this project was to test a new method for determining the judge time required by various types of cases. The utility of knowing such time requirements, or case "burdens," is that they might provide a means to evaluate more precisely the burden of a given court's or judge's caseload, and thus provide objective standards for allocation of cases to panels and for allocation of judicial resources to courts. While several previous studies have attempted to measure case burdens in terms of time, they generally have not provided results of satisfying consistency. Moreover, these studies usually have employed the rather expensive and time-consuming method of having judges keep detailed time records. The new method tested in this project avoided time-keeping in

favor of simply asking judges to estimate how much of their time a given type of case typically consumes.

The particular estimation method employed was a three-stage iterative plan. In the first stage, the judges from three courts of appeals completed a questionnaire in which they evaluated a taxonomy of 23 case-types,[8] and then estimated the relative time burden of each case-type. This time burden was estimated relative to a "base case," with the "typical direct criminal appeal[9] serving as this base. Thus burden was given in base case units. The base case had by definition a weight of 1. A case estimated to require twice as much time as the base case would thus have a burden (or "weight") of 2, while a case taking only three-fourths the time of the base case would have a weight of 0.75. In order to provide a conversion factor for translating base case units to actual time burdens, the judges also estimated the time, in hours, required for the base case. In the second stage, the judges of each circuit met with project personnel to re-examine the questionnaire results. At

---

[8] Defining this taxonomy was in itself a substantial task. In order to be useful to the aim of this project, the taxonomy had to: 1) include all cases brought before the Courts of Appeal, 2) be not so extensive as to make the estimation process a severe burden on judicial time, 3) have category labels that clearly indicated the types of cases included in each category, 4) assure that each category contained cases of similar burden, and 5) be susceptible to unambiguous translation into the taxonomy by which the Administrative Office labels appellate cases (so that we could determine how many cases in each of our categories were handled by each court). Seven existing taxonomies were examined for their conformity to these criteria, but none was found suitable. The combined efforts of an ad hoc panel of judges, personnel of the Federal Judicial Center, and the project contractor were necessary to devise the taxonomy that was submitted to the participating circuits.

[9] Except in the 6th Circuit, where the judges felt that Diversity Motor Vehicle Personal Injury cases would be a more stable reference point.

these meetings, a Consensor (an electronic voting device) was used to facilitate presentation of the judges' patterns of "voting" on case burdens and to aid in moving them toward "consensus." The final stage of the estimation method was a follow-up questionnaire, which presented the meeting results and asked for a final re-evaluation of the burden estimates. At each of the estimation stages, the judges indicated not only their estimate of case-type burdens, but also a numerical indication of the confidence they had in their estimate. The group judgments of burdens (in base case units) and hours required per base case were then computed as confidence-weighted averages, with the effect that the vote of a judge expressing a confidence of 10 would count twice as much as a vote with confidence of 5.

In addition to estimating case-type burdens, the judges also provided estimates for two ancillary data sets. In order to provide a basis for evaluating reasonableness of the burden estimates, the judges were asked to estimate their total time expenditures in a year. With an eye toward possible future elaboration or simplification of the case-type taxonomy, the judges also were asked to evaluate the "adequacy" of certain "indicators" of case burden. The indicators are descriptors of case characteristics (e.g., number of parties before the court; procedural stage at termination of appeal); and the judges used a numerical scale to evaluate the adequacy of each indicator as a correlate to case time burden.

II. Results

The final data results of the project[10] are presented in Tables I, II, and III. The Tables include the case-type burden-weights and converted burden-hours, the average judge work-year time breakdown, and the indicator adequacy data. These data are presented for each of the three courts, with averages developed across the three courts presented in a fourth column.

---

[10] Note that the "final" case burden data for the D.C. Circuit is in reality the product of the meeting (second stage of estimation). Because the D.C. follow-up questionnaire asked for re-estimation of burdens for groups of case-types, and only 3 of these questionnaires provided thorough responses, we feel that the final D.C. results are not as reliable as the meeting results.

TABLE I

Case-Type Burdens

Case Type	D.C. Circuit		6th Circuit		8th Circuit		Average	
	Burden Weight	Hours	Burden Weight	Hours	Burden Weight	Hours	Burden Weight	Hours
1. Tax Court of the U.S. Cases	2.8	8.1	2.7	16.7	1.9	14.1	2.5	13.0
2. NLRB Cases	2.8	8.1	1.4	8.7	1.0	7.4	1.7	8.1
3. Power, Transportation, and Communication	10.1	29.3	3.3	20.5	3.2	23.7	5.5	24.5
4. Healty, Safty, and Environment	9.9	28.7	3.9	24.1	1.4	10.4	5.1	21.1
5. Other Regulatory Agency Cases	9.3	27.0	3.0	13.7	2.6	19.2	5.0	20.0
6. Original Proceedings	2.7	7.8	0.3	1.9	0.7	5.2	1.2	5.0
7. Civil Rights	4.5	13.1	3.7	22.9	3.2	23.7	3.8	19.9
8. Prisoner Actions Other Than Collaterial Attack	2.8	8.1	0.3	1.9	0.7	5.2	1.3	5.1
9. Labor	3.8	11.0	2.8	17.3	1.7	12.6	2.8	13.6
10. Anti-Trust	9.6	27.8	5.1	31.6	5.8	42.9	6.8	34.1
11. Patents	4.0	11.6	5.1	31.6	5.0	37.0	4.7	26.7
12. Copyright, Trademark, and Unfair Trade Practices	4.0	11.6	3.0	18.6	3.2	23.7	3.4	18.0
13. Bankruptcy	3.3	9.6	1.6	9.9	1.6	11.8	2.2	10.4

TABLE I

Case-Type Burdens

	Case Type	D.C. Circuit Burden Weight	D.C. Circuit Burden Hours	6th Circuit Burden Weight	6th Circuit Burden Hours	8th Circuit Burden Weight	8th Circuit Burden Hours	Average Burden Weight	Average Burden Hours
14.	Tax Suits	4.0	11.6	2.0	12.4	1.9	14.1	2.6	12.7
15.	Securities, Commodities, Exchanges, and Stock-holder Actions	4.5	13.1	3.0	18.6	4.9	36.3	4.1	22.7
16.	Injury Actions by Marine & Railway Employees	3.6	10.4	2.0	12.4	1.9	14.1	2.5	12.3
17.	Other Marine Actions	4.0	11.6	2.2	13.6	2.2	16.3	2.8	13.8
18.	Suits Challenging Validity of Action or Inaction of Federal Agencies or Officials	9.6	27.8	1.4	8.7	3.1	22.9	4.7	19.8
19.	Other Civil Actions Based on Federal Statutes	4.0	11.6	1.4	8.7	1.7	12.6	2.4	11.0
20.	Other Civil Actions with U.S. as Plaintiff	3.5	10.2	1.5	9.3	1.7	12.6	2.2	10.7
21.	Diversity Actions	2.8	8.1	1.9	11.8	2.2	16.3	2.3	12.1
22.	Direct Criminal Appeals	1.0	2.9	1.0	6.2	1.0	7.4	1.0	5.5
23.	Collateral Attacks	1.2	3.5	0.4	2.5	0.8	5.9	0.7	4.0

(con't)

TABLE I

Case-Type Burdens

	Case-Type	D.C. Circuit Burden Weight	D.C. Circuit Burden Hours	6th Circuit Burden Weight	6th Circuit Burden Hours	8th Circuit burden Weight	8th Circuit burden Hours	Average Burden Weight	Average Burden Hours
19A.*	Freedom of Information Act	6.4	18.6						
7A.*	School Desegregation			6.6	40.9				
16A.*	Social Security			0.9	5.6				
4A.*	Environmental Protection Agency Cases					3.6	39.2		

*These case-types represent special additions to the Taxonomy made during the meetings. Each of these was recognized as a separate case-type by only one court. Thus, for instance, while the 6th Circuit assigned separate burdens to school desegregation cases (Type 7A), and civil rights cases [other than school desegregation] (Type 7), the other circuits assigned burdens only to the more general category, civil rights cases (Type 7).

TABLE II

Average Judge Work-Year Data

	D.C. Circuit	6th Circuit	8th Circuit	Average
Gross Work-Year (hours)	2740	2850	2500	2697
Percent of Work-Year Spent on Non-Case Related Work	22%	26%	24%	24%
Time Spent on Motions Not Related to Cases (hours)	72	179	105	119
Total Time Devoted to Submitted Cases (Computed from Above; hours)	2065	1930	1795	1930
Percent of Case Time Spent on Extreme Cases*	33%	40%	26%	33%

*This figure was elicited because the judges were instructed to disregard "extreme" cases in arriving at their burden estimates. The extreme cases were defined as the 10% most burdensome and 10% least burdensome cases. The use of this figure is discussed in the section on analysis.

TABLE III

Average Adequacy Values* For Indicators

	Indicator	D.C. Circuit	6th Circuit	8th Circuit	Average	Average Rank
1.	Number of Parties Before Appellate Court	5.2	2.7	1.8	3.2	15
2A.	Federal Government Present as Appellant	7.7			7.7	
2B.	Federal Government Present as a Party	5.2	2.4	1.2	2.9	19
3.	Number of Cross Appeals	5.9	3.9	3.3	4.4	11
4.	Number of Issues Presented in Briefs	6.2	2.7	3.8	4.2	10
5.	Presence of an Opinion from the District Court	5.3	4.1	3.8	4.4	9
6.	Type of Counsel for Parties (e.g., Retained, Appointed, House Counsel, Pro Se, etc.)	5.4	2.3	1.2	3.0	18
7.	Nature of Relief Sought in Trial Court (e.g., Money Damages, Injunction)	4.2	1.7	2.5	2.8	20
8.	Length of Appendices from District Court	4.5	3.2	4.5	4.1	12
9.	Number of Amicus Curiae Briefs Filed	5.2	3.1	3.8	4.0	13
10.	Aggregate Length of All Briefs Filed	6.3	4.3	5.4	5.3	7
11.	Number of Motions Disposed of with Hearing	6.0	1.0	1.6	2.9	16
12.	Time Used in Oral Argument	5.9	4.8	4.9	5.2	8

*Adequacy was valued on a scale from 0 to 10, with 0 meaning the indicator would be of no value in indicating case burden, and 10 meaning the indicator would correlate nearly perfectly with case time burden.

(con't)

TABLE III

Average Adequacy Values* For Indicators

	Indicator	D.C. Circuit	6th Circuit	8th Circuit	Average	Average Rank
13.	Procedural Stage at Termination of Appeal	8.0	7.8	4.9	6.9	4
14.	Length of Disposition (e.g., Number of Pages, with Oral Dispositions Translated to Page Equivalent)	6.8	5.8	5.4	6.0	6
15.	Type of Disposition (e.g., Signed or Per Curiam, etc.)	6.9	6.9	6.3	6.7	2
16.	Presence of Dissenting or Concurring Opinions	6.4	6.1	5.6	6.0	5
17.	Aggregate Length of All Dissenting and/or Concurring Opinions	6.6	6.2	6.2	6.3	3
18A.	Petition Granted for En Banc Review	8.8	8.8	6.3	8.0	1
18B.	Petition for En Banc Review	1.7		1.0	1.4	21
19.	Number of Citations in All Opinions (Including Repetitions)	3.0	3.0	3.6	3.2	16
20.	Number of Citations in All Opinions (Excluding Repetitions)	3.6	3.0	3.6	3.4	14

III. Analysis of Results

a. Approach

The principal question for analysis of the burden estimates is whether these estimates appear to be useful in measuring caseload burdens. Their utility would be clear, of course, if they proved to be accurate measures of the actual time burdens imposed by the "typical" cases of the various case-types. However, the problem here is that there are no standards against which to compare the estimated burdens; we do not know how much time a given type of case does take. The only alternative short of measuring that time is to examine both the internal consistency (the extent to which the three circuits agree on the burden-hours) and external consistency (the extent to which measurements of total court caseload based on the burden estimates for individual case-types agree with each other or with other caseload estimates).

b. Internal Consistency

Internal consistency may be analyzed in a relatively subjective manner. By referring to Table I, and comparing the burden-hours estimates across circuits, one will note instances where case-types exhibit both close agreement and strong disagreement. Two examples of this phenomenon are case-type 13 (Bankruptcy), where there is rather close agreement among the circuits on burden-hours, and case-type 5 (Other Regulatory Agency Cases), on which the circuits disagree rather strongly. Since cross-circuit disagreement on time-burdens tends to detract from the potential utility of a standard set of case-type burdens, it is most

desirable to identify the reasons why certain case-types exhibit such disagreement.

A variety of possible reasons for cross-circuit variation in case-type time burdens may be suggested. One likely explanation is that there are actual variations in the burden of a case-type due to variations in applicable law or other regional characteristice (e.g., the D.C. Circuit may tend to get more difficult regulatory agency cases; diversity actions may vary in difficulty according to applicable state law). Another possible reason for the disagreement is that certain case-types in the taxonomy are not sufficiently narrow or unambiguous to assure that the judges of each circuit were really considering the same sorts of cases when they made their burden estimates. A more basic and perhaps more compelling explanation is simply that, no matter how specific the case-type may be, the experience of judges or of courts with that case-type will usually be quite varied. In other words, it may be that even within a very narrow class of cases, the time required by individual cases will vary across a broad range without tending to cluster about a "typical" time that is susceptible of accurate estimation. Thus, even though a court has experienced the full range of difficulty of cases within a given case-type, it may not be able to distill an average or "typical" case time. Moreover, a court that has experience mostly with the "easier" cases of a given case-type would, of course, estimate a lower burden time than would a court that has experienced mostly "harder" cases.

The suggestion that cross-circuit variation in case-type time burdens is due to inherent variation in the cases with a case-type and/or variation in the experience of the circuits with that case-type, finds some support from the Center's 1969-1970 Federal District Court Time Study. That study obtained detailed time records from over 60 per cent of all district judges, from which were derived case-type time burdens for a very detailed taxonomy of trial cases. While the District Court Time Study is not directly comparable to the present study, it is worth noting that there the variations among case-type time burdens were generally similar to, but slightly larger than, the cross-circuit variations obtained in the present study.[11] This result suggests that the variations observed in the present study may stem from variations in judge experience, not from the method used.

While we have suggested that there is some inherent variability associated with the process of assigning time burdens to case-types within a taxonomy, it is far from clear that all of the variability experienced in this project was unavoidable. We must still consider

---

[11] This result was obtained by comparing normalized variations in burden-time. For a given case-type, the average of all burden-time estimates was divided into the maximum deviation of all estimates from that average. The resulting normalized variation thus is an expression of the maximum deviation from the average as a percentage of the average (e.g., for three burden-time estimates of 2, 6, and 7 hours, the average is 5 and the maximum deviation from 5 is 3 hours; 3 is 60 percent of 5, thus the normalized variation is 60). In the present study, the average normalized cross-circuit variation for all case-types was 36, while in the District Court Study, the average "expected" variation was about 46.

whether refinements in the taxonomy could reduce cross-circuit disagreement on time burdens. The possibility remains that some of this disagreement was caused by ambiguous or over-broad case-type descriptions. Resolving this possibility, however, requires the considered advice of the judiciary. Only the judges can know whether a given case-type description represents a true family of similar cases. In the end, of course, we must recognize that there can be no perfect taxonomy -- every case is unique to some extent, hence every taxonomy of cases is imperfect. It is a matter of degree, and the degree is impossible of precise measurement.

c. External Consistency -- Absolute Caseload Measures

As shown in Table II, the circuit judges were asked to make estimates on various facets of their work year. These data provide one device for gauging external consistency of the burden estimates. The Table II data provide an estimate of the total judge-time devoted to relevant cases.[12] The burden estimates, along with Administrative Office data on cases handled in each circuit in 1975, provide the means to compute the same quantity. By summing the product of burden hours and the number of relevant cases for each case type, a computed case time can be derived. Since the judges were instructed to direct their burden estimates at

---

[12] What the precise definition of "relevant cases" (i.e., those cases to which the burdens are applicable) should be is not clear. Relevant cases could mean all cases terminated, or only those terminated "with judicial action." As discussed infra, the most reliable definition available, though not necessarily the most logical, confined relevant cases to those terminated after submission or oral hearing, with brief(s) filed. Unless otherwise noted, that definition will be applied hereinafter.

the middle 80% of each case-type (i.e., non-extreme cases), the estimated and computed case-times must be adjusted accordingly. The comparison of these two adjusted values, as illustrated in Table IV, is a measure of external consistency.

TABLE IV

Circuit	Estimated Time per Judgeship on Relevant Cases (Hours)	Time per Judgeship on Same Cases, Computed on Basis of Burden Estimates (Hours), FY75	Computed Time as a Percentage of Estimated Time
D.C.	2065	2108	102%
6th	1930	3828	198%
8th	1975	2776	155%

While it is clear that there is substantial difference between the estimated ad computed values for the 6th and 8th Circuits, the result may nevertheless be seen as encouraging. The values computed on the basis of the case-type time burdens are not so unrealistic as to suggest that the individual burden estimates are grossly inaccurate. Indeed, since it is the estimated time burden of the base case (direct criminal appeal) that determines the time burdens of all case-types (the other case-types were assigned burdens relative to the base case), misestimation of the single figure "base case hours" directly affects the total computed time per judgeship. For instance, if the 6th Circuit had estimated that the typical direct criminal appeal took 4 hours (instead of the 6.2 hours actually estimated) then its computed time per judgeship would have been 2470 hours, instead of 3828. Since wide variability in the

difficulty of direct criminal appeals might make a 50% over-estimation of average time consumption quite understandable, it might well be suggested that discrepancy between the estimated and computed time-per-judgeship figures is not so much the fault of the judge's inability to estimate as it is of the methodology that caused the computed figures to rely so heavily on a single estimate among many. While some tendency toward over-estimation of burdens is suggested, it nonetheless appears fair to say that there is some promise in this method of estimating caseload burden.

In order to determine whether variations of this estimating method might have produced computed caseload estimates more consistent with the judges' work-year estimates, several methods of computing caseloads from the burden estimates were developed and tested. Each of these variations is discussed briefly below, and a comparative chart of the results is presented in Table V.

Variations of Caseload Computation

(a) Precise computation. This is the same computation as that used for Table IV.

(b) Non-adjusted computation. This method ignores the adjustments for the 20% "extreme" cases, and computes caseload as the sum, over all case-types, of the products of time burden and number of cases.

(c) Three-case-type taxonomy. This method simplifies the taxonomy into three general case-types selected in a fairly subjective manner. The three general case-types are:

(1) High burden, consisting of case-types 3, 4, 10, and 11 of the original taxonomy.

(2) Low burden, consisting of case-types 2, 6, 22, and 23.

(3) Medium burden, including all other cases.

The time burden assigned to each of these three case-types was the average of the burden-hours of each of the original taxonomy case-types included within them (e.g., the burden-hours for the high burden case-type is the average of the burden-hours of case-types 3, 4, 10 and 11). The computation of caseload was analogous to that of variation (b), above.

(d) District of Columbia Three-type Taxonomy. This computation applied the burdens for the D.C. Circuit three-case-type taxonomy to the caseloads of the circuits.

TABLE V

Comparison of Various Computations of Caseload:

Hours per Judgeship per Year Spent on Submitted Cases, with Average for 3 Circuits, and Caseload as a Percentage of Estimated Caseload (in parentheses)

	D.C. Circuit	6th Circuit	8th Circuit	Average
Judges' Estimated Caseload:	2065 (100)	1930 (100)	1795 (100)	1930 (100)
Computed Caseload Based on:				
(a) Precise Computation	2108 (102)	3828 (198)	2776 (155)	2904 (152)
(b) Non-Adjusted Computation	1766 (86)	2871 (149)	2568 (143)	2401 (126)
(c) Three Case-Type Taxonomy	1911 (93)	2602 (135)	2536 (141)	2350 (123)
(d) D.C. Three-Type Taxonomy	1911 (93)	2703 (140)	2064 (115)	2226 (116)

Table V indicates that the two computations based on the three case-type taxonomy (c, d) achieve greater consistency with the judges' estimated caseload time than do those based on the larger, 23 case-type taxonomy. While this tends to indicate that a less extensive taxonomy may serve our purposes as well as the taxonomy used in this project, this result is one that should be taken with a substantial grain of salt. Several problems are apparent. First, the three-case-type taxonomy is merely the most consistent of several computational approaches that were tried. It is practically certain that, after significant effort, some method of manipulating the results could be found that would achieve near perfect consistency with the judges' estimates; but that does not mean that such a manipulation would achieve success in a subsequent application. Secondly, the time burdens associated with the three case-types were ones derived by the Center and are not necessarily representative of what the judges' estimates might have been had they been asked for burden estimates for this simplified taxonomy. Finally, as mentioned previously, the various caseload computation devices are judged here by their degree of consistency with the judges' estimates of caseload. Since there is no way of knowing how accurate those estimates are, there is no way of knowing whether consistency correlates with accuracy. It may be that the caseload computations based on the 23 case-type taxonomy are in fact the more accurate ones.

d. External Consistency -- Relative Caseload Measures

The most revealing analysis of the case-type burden estimates is that which views them as a device for relative (as opposed to absolute)

caseload measurement. Here the burden estimates are used merely to compare among several caseloads and not to determine the actual caseload burdens in hours. Thus, for instance, the D.C. Circuit estimates gave case-type 11 (patents) a weight of 4, and case-type 22 (direct criminal appeals) a weight of 1. A caseload of ten patent cases (40 units) is thus twice as burdensome as a caseload of two patent cases and twelve direct criminal appeals (20 units).

One way of testing consistency of the burden estimates in relative measurements of caseloads is to apply the three sets of estimates (from the D.C., 6th and 8th Circuits) to the 1975 caseloads of each of the eleven circuits, and then compare the relative caseload burdens of each of the eleven courts across the three different sets of estimates. In order to make such comparisons, however, an adjustment must be made for the fact that a "unit" of burden under one of the three sets of estimates is not necessarily the same as a unit from another set. In other words, since the concern here is with relative caseload measures, it matters not that according to one set of burden estimates, Circuits A and B have average caseloads measured at 3000 and 2000 units (respectively), but are measured at 1500 and 1000 units according to another set of estimates. In each instance, Circuit A is 50% more "burdened" than Circuit B. Alternatively, it may be observed that in either case, Circuit A has a caseload burden equal to 120% of the average burden (e.g., the average of 3000 and 2000 is 2500, and 3000 is 120% of 2500). Such adjustments have been employed to compare the three sets of burden

estimates as estimators of relative caseload burdens; for each set of burden estimates, the caseload measure of each of the eleven circuits is computed, the average of those eleven values is calculated, and then each circuit's measure is expressed as a percentage of that average.

A comparison of these relative caseload values is presented in Table VII (p. 29). That table provides, for each circuit: per judgeship weighted caseloads (relative to the average) as computed from each of the three sets of case-type burden estimates (columns 1 through 3); and a number of relative caseload measures based on gross (unweighted) per judgeship case processing data (columns 4 through 8). The three most striking features of Table VII are: the surprising similarity of the three sets of weighted caseload values (similarity of columns 1, 2, and 3); the equally strong similarity between the measures based on unweighted relevant cases (column 4) and the weighted measures (columns 1-3); and the inconsistency between these four measures and any of the other caseload measures (columns 5-8). Each of these features suggests significant implications on the utility of weighting and/or of caseload measurements in general.

The consistency of the three weighted caseload measures is apparent from the fact that, except for the D.C. Circuit, the maximum difference among the three measures of a given circuit is 11 percentage points (the Second Circuit was rated 99 -- just below average -- according to the D.C. court's burden estimates, and 110 -- 10% above average -- by the 8th Circuit's estimates). What makes this consistency so surprising

is that the three sets of burden estimates (on which the caseload measurements are based) were so inconsistent.

This anomaly is probably the result of several factors. First, the taxonomy of 23 case-types was so detailed that only 6 of the 23 represented more than 5% of the relevant cases; about 7 of the case-types represented less than 1% of the cases. Thus, even an extreme difference in the burdens ascribed to a given case-type would tend to produce a minimal difference in weighted caseload measures. Second, to the extent that inconsistencies in burden estimates vary in their "direction," their effects would tend to "wash out:" if, for instance, the Sixth Circuit gave case-type A twice the burden given by the Eighth Circuit, but gave case-type B only 1/2 the burden given by the Eighth, then, assuming equal numbers of cases of the two types, the inconsistent burdens would nullify each other's effect on weighted caseload values. Finally, note that if two courts have identical proportions of cases of each case-type, then any set of burden estimates will produce the same relative weighted caseload measures (a simple analogy: if basket A contains 10 apples and 4 oranges (cases) and basket B contains 5 apples and 2 oranges, then no matter what the prices (weights) of apples and oranges may be, basket A will cost twice as much as B; it is the relative cost that is important, not the absolute cost). Probably the strongest reason for the consistency of the three weighted caseload measures is simply that the 11 circuits have roughly the same proportions of cases at each level of burden.

This propostion, that the circuits have similar distributions of high, low, and medium burden cases, finds support in a number of ways. It can be seen directly by calculating for each circuit the percentage of cases falling into each case-type of the three-case-type taxonomy discussed at page 22, supra. These percentages are displayed in Table VIII.[13] Additional support for this propostion is found in the close agreement between weighted and unweighted caseload measures, which was identified earlier as the second major feature of Table VII.

---

[13] The table shows that the D.C. Circuit is clearly different from the rest in that 17% of its caseload is in the high burden category, while 6% is the largest of such proportions among the other circuits. This difference is due largely to the fact that the D.C. Circuit had 66 cases of type 3 (power, transportation, and communication cases), while no other circuit had more than six such cases. This disproportion, along with the substantially higher relative burden ascribed to that case-type by the D.C. Circuit, accounts for much of the 18 point disagreement among the three circuits on the relative caseload measure of the D.C. Circuit (see Table VII, first row of columns 1 through 3).

TABLE VII

Fiscal 1975 Relative* Caseload Per Judgeship, and Rank (), for the 11 Circuits, Based on:

	Case-Type Burden Estimates of:			Unweighted Case-Processing Data:				
Circuit	D.C. Cir	6th Cir	8th Cir	Relevant Cases	Filings	Terminations	Pending Cases	Signed Opinions
D.C.	81 (8)	64 (11)	63 (11)	60 (11)	74 (11)	74 (11)	116 (4)	57 (11)
1st	103 (4)	97 (6)	96 (6)	93 (7)	96 (5)	88 (8)	65 (10)	130 (3)
2nd	99 (5)	104 (4)	110 (4)	100 (5)	116 (3)	126 (2)	81 (9)	110 (5)
3rd	94 (7)	96 (7)	95 (7)	94 (6)	93 (7)	91 (7)	82 (8)	67 (8)
4th	74 (10)	80 (8)	78 (8)	86 (8)	113 (4)	113 (4)	130 (3)	63 (10)
5th	161 (1)	160 (1)	162 (1)	173 (1)	132 (1)	135 (1)	138 (2)	147 (1)
6th	98 (6)	102 (5)	98 (5)	101 (4)	96 (5)	93 (6)	87 (6)	66 (9)
7th	123 (2)	130 (2)	130 (2)	123 (2)	88 (8)	98 (5)	85 (7)	139 (2)
8th	75 (9)	75 (9)	74 (9)	79 (9)	76 (10)	78 (10)	54 (11)	107 (6)
9th	116 (3)	117 (3)	118 (3)	121 (3)	127 (2)	119 (3)	175 (1)	92 (7)
10th	72 (11)	70 (10)	71 (10)	66 (10)	84 (9)	75 (11)	88 (5)	123 (4)
Column #	1	2	3	4	5	6	7	8

*Each caseload value is expressed as a percentage of the average of the eleven circuits. Thus, in any column, a value of 100 means the circuit is of average caseload according to the measurement technique given in the column heading. A value of 150 would mean 50% more burdened than average.

TABLE VIII

Percentage of Cases in Case-Types of the 3-Case-Type Taxonomy, by Circuit

Case-Type	Circuits: D.C.	1	2	3	4	5	6	7	8	9	10	Circuit Average
Low Burden	48	41	48	47	61	55	50	52	54	53	41	50
Medium Burden	35	56	49	49	37	41	47	43	44	44	53	45
High Burden	17	2	3	4	2	3	3	4	2	3	6	4

This close agreement suggests, of course, that weighted measurements of court[14] caseloads are not much different from unweighted measurements; it suggests that weighting has little effect. This in turn means that inasmuch as the present concern is with measuring the caseloads of courts, weighting by case-types may be only minimally useful. Unfortunately, weighting efforts cannot yet be abandoned outright, because two confounding factors lend uncertainty to these results. The first is that differences that have been observed between weighted and unweighted caseloads are, though slight, nevertheless too large to be dismissed as trivial. The

[14] It should be noted that this result applies only to relative measurements of court caseloads. The use of weights in determining the relative caseloads of individual judges probably would make a difference, since it is not likely that the proportions of cases of similar burden assigned to individual judges would be consistent among the judges of a given court. Indeed, assignment of cases to judges by a case-type weighting scheme would probably result in a more consistently even workload distribution than would random or rotational assignment. It is not clear, however, that a very detailed weighting scheme (i.e., taxonomy) would achieve any greater consistency than would a simple, 3-weight scheme: the scheme might be useful without being complicated. Such devices are in fact used in some courts, largely based on intuitive weighting.

second factor is the rather disturbing suggestion that the unweighted caseload measures may themselves be based on quite inconsistent data (i.e., that the circuits label cases in differing fashions).

The non-triviality of the difference between weighted and unweighted measures can be seen rather simply from that fact that the average difference between these measures was more than 7 percentage points (average for 10 circuits, with the D.C. Circuit omitted as an anomaly). Moreover, the difference was at least 10 points for four of those circuits. Since most circuits have 9 judgeships, an increase or decrease of 1 judgeship would change the typical circuit's relative caseload measure by about 11 percentage points. While an average discrepancy of 7 percentage points between weighted and unweighted measures appears rather small, it is equivalent to a difference of more than half of a judgeship. Thus, if relative caseload measures were used to dictate the allocation of judgeships among circuits, different results would obtain depending on whether weighted or unweighted measures were used. Hence it cannot be said that weighting has an insignificant effect.[15]

---

[15] It is also important to note that weighting appears to have a directionally consistent effect. Our analysis included the computation of both weighted and unweighted caseloads based on a variety of definitions of "relevant" cases for FY 73 and FY 75. We noted a strong correlation to the effect that if the weighted measure of a given circuit was higher (or lower) than the unweighted measure for a given definition of "relevant" cases and a given fiscal year (e.g., terminations in FY 73), then there was a strong probability that the weighted measure would be higher (or lower) for all other combinations of definition and year (e.g., filings in FY 75).

The most disturbing and most important factor that must restrain a final judgement against caseload weighting is that existing appellate court statistical reporting may not provide consistent measures of gross unweighted caseloads (i.e., numbers of relevant cases). Without such consistent measures, comparisons of weighted and unweighted caseload measures rest on very shaky grounds. A fairly well-known inconsistency in statistical reporting may serve as a dramatic example of the problem.

Until recently, the 4th and 10th Circuits routinely place prisoner petitions (which constitute case-types 8 and 23) on the general docket, while in other circuits most of such cases were placed on the miscellaneous docket.[16] As a result, most prisoner petitions in the 4th and 10th were recorded as case filings (and, subsequently, as terminations), while in the other circuits the majority were never counted as filings (or as terminations). This was not reflective of a difference in the attention accorded such petitions, but of a mere difference in labelling: identical cases receiving identical treatment would be labeled as filings in the 4th and 10th, but not in any other circuits. As a result, the 4th Circuit recorded 109 collateral attack filings per judgeship in Fy 73, while no other circuit recorded more than 42 per judgeship. Since the 4th Circuit's total of all filings per judgeship was only 225 in that year, it is apparent that the "over reporting" of prisoner

---

[16] Matters placed on the miscellaneous docket are not considered "filings" in the appellate court statistical reporting plan, hence they are never counted as "cases."

petitions severely distorted that circuit's relative caseload measures based on unweighted filings (that measure being 143). Since prisoner petitions were assigned relatively low weights, the distortion was moderated somewhat by the weighted measures; the largest of the three weighted measures was 115. The difference in the weighted and unweighted measures was very large, about 30 percentage points, but was caused merely by inconsistent reporting procedures, and not by a real difference in the caseload structure of the circuit.

Among the variety of definitions of "relevant" cases that were used to produce caseload measures, it was found that these "over-reported" prisoner petitions were within all but the most restrictive of the definitions. That is, they are usually counted as filings and, subsequently, as "terminations after submission or oral hearing" in the 4th and 10th Circuits (while not so counted in the other circuits). Yet briefs are not generally filed in these cases, and therefore they rarely are counted as "termination after submission or oral hearing, with briefs filed." Thus when relative caseload measures were computed based on this restrictive definition of the cases to be counted, the unweighted measure of the 4th Circuit became 87, while the weighted measures varied from 82 to 88. Weighting thus appears to have little effect, while the restrictive definition has vastly altered the circuit's caseload measure.[17]

---

[17] It should be noted that the practices of the 4th and 10th Circuits are similar, but not identical. Moreover, our concern is not with variations in practice, but with variations in the meaning of the statistical term, "filings," that result.

It appears that the inconsistencies in labelling of prisoner petitions can be eliminated by restricting the definition of relevant cases to those in which briefs were filed. However, it is not known what similar inconsistencies may yet exist that were not thus eliminated. If other such inconsistencies do exist, then we cannot be sure how much they may have distorted the comparison of weighted and unweighted caseload measures. They might have caused the two measures to appear more or less consistent than they should.

Moreover, the data produced in the course of the analysis tend to suggest that other case "labelling" discrepancies in fact do exist. This is seen by observing the proportions of cases in each circuit that failed to attain a given stage of "procedural labelling." For instance, in every circuit but the 4th and 10th (for FYs 73 and 75), briefs were filed in at least 88% of all cases terminated after submission or oral hearing. In the 4th and 10th, however, briefs were filed in no more than 67% of such cases. This anomaly is striking, and should alert the researcher to the possibility of inconsistent labelling (in this case, the anomaly is due to those circuits' practices with respect to prisoner petitions). Similarly, when examining the proportion of terminated cases that were terminated without ju icial action, an irregularity is found in the 1st Circuit. That circuit terminates about 30% of its cases without judicial action, while no other circuit so terminates more than 19% of its cases (7 of the others so terminate less than 13%). The reason for this irregularity is that the 1st Circuit dockets a

case upon the filing of a notice of appeal, before payment of the docket fee. If the docket fee is not paid, the case is dismissed. Unfortunately, there is no way to determine what proportion of "terminations without judicial action" is attributable to this unique 1st Circuit practice, and what proportion is comparable to the "terminations without judicial action" of other circuits. Finally, the proportion of those cases terminated with judicial action (excluding consolidations) that do not receive an oral hearing or consideration upon submission varies broadly among the circuits from 2% to 37%. While this, too, may relect real variation in practice, it may also reflect the situation where two cases receiving identical treatment in different circuits would in one circuit be labelled a submission, but in another be labelled as a termination without submission or oral hearing.

Conclusions

The primary aim of court caseload measurements (be they weighted or unweighted) is to obtain an objective assessment of judicial workload. If case processing volume (e.g., filings, terminations with judicial action) is to be the basis for such caseload measurement, then it is necessary to assure that the various courts assign procedural progress labels according to identical criteria. We cannot measure caseloads based on filings so long as a given caseload would in one circuit be counted as 100 filings, but in another as 200 filings. We must measure by the same yardstick. Analyzing the effects of caseload weighting,

or selecting a reliable basis for unweighted measurements, cannot be accomplished until that single yardstick is defined.

A clear message of the analysis in this project is that the circuit court docket report requirements (Form JS-34) must be evaluated and clarified. Only then can a reliable judgment on the utility of case weighting be made. However, the apparent inability to draw definite conclusions about the merit of appellate case weighting hardly suggests that the efforts of the participating circuits have been in vain. The burden estimates they provided have shown that the questionable reliability of appellate court statistical reporting is a severe impediment to the analysis of appellate court management innovations. Those estimates should also be helpful in testing the reliability of a revised statistical reporting plan. Moreover, they have suggested that caseload weighting may be of little utility to the appellate courts, a suggestion which, if confirmed, will undoubtedly result in substantial savings of time and money.