

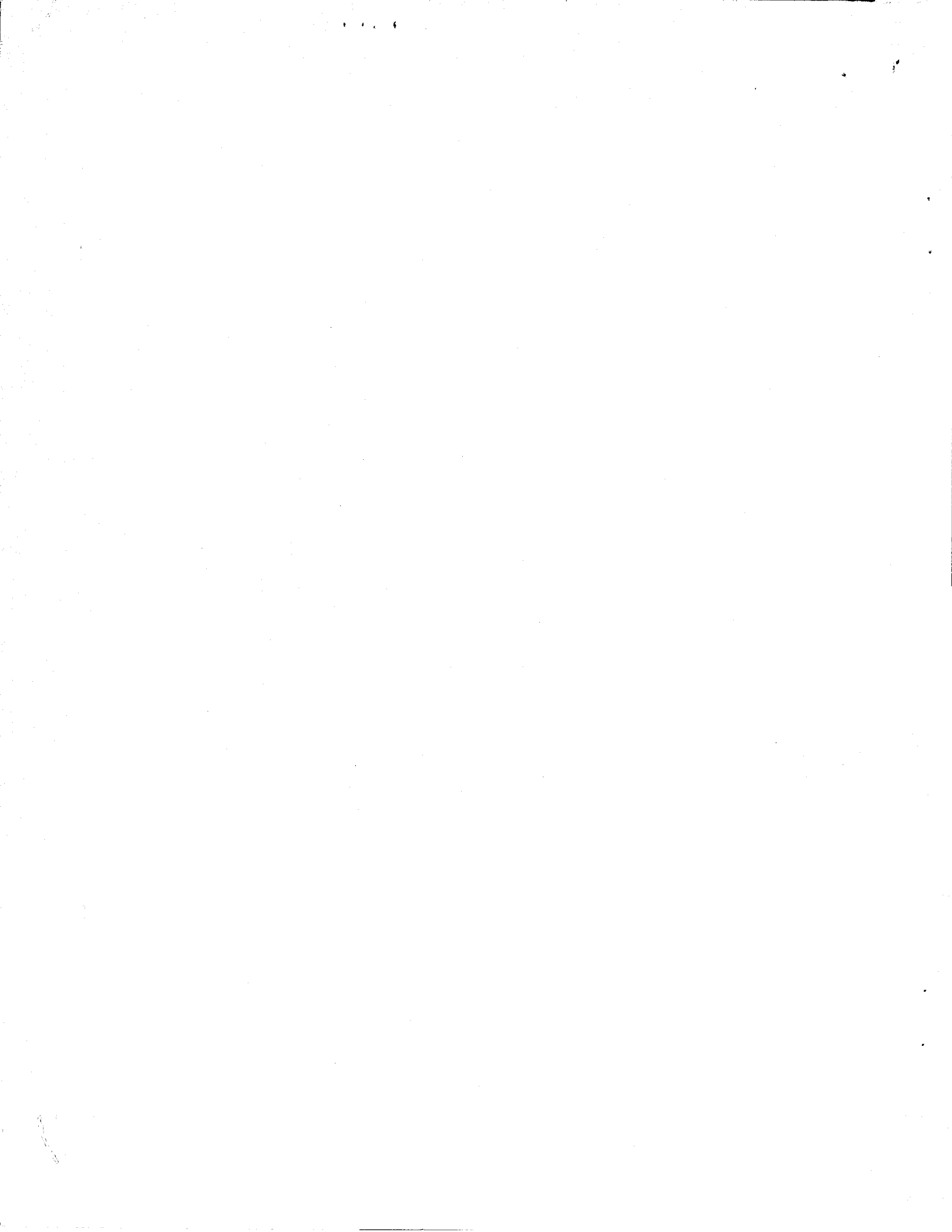
EXPERIMENTING WITH APPELLATE REFORM:

THE SECOND CIRCUIT EXPERIENCE

Jerry Goldman  
Department of Political Science  
Northwestern University  
Evanston, Illinois 60201

Prepared for delivery at the 1978 Meetings  
of the Law and Society Association

50391



EXPERIMENTING WITH APPELLATE REFORM:

THE SECOND CIRCUIT EXPERIENCE

The purpose of this paper is to share the problems, solutions, and frustrations in the design and execution of a legal experiment. The term "experiment" shall have the meaning that Julian Stanley has given to it:

I shall define a true, variable-manipulating, controlled comparative experiment as an investigation in which experimental units are assigned in a simple-random or restrictively random manner to at least some of the experimental combinations.<sup>1</sup>

The legal experiment I executed tested whether and to what extent preappeal conferences conducted by a senior staff attorney could: reduce the proportion of cases that otherwise would be decided by the judges after briefing and argument, improve the quality of appeals that would go on to the judges for decision, and expedite the appellate process. The experiment was conducted in the U. S. Court of Appeals for the Second Circuit from 1974 to 1977.

The results of this test demonstrated that the program: did not substantially reduce the proportion of cases decided

---

<sup>1</sup>Stanley, "On Improving Certain Aspects of Educational Experimentation," in Stanley (ed.), Experimental Design and Statistical Analysis 5 (1967).

after briefs and argument, did affect the quality of appeals but only in a trivial way, and did not significantly expedite the process for appeals that run the gamut from notice through decision.<sup>2</sup> On the basis of this evidence, or perhaps in spite of it, the Second Circuit enlarged its program, which continues to operate today as it did during the period of the experiment

### Mounting the Experiment

In 1973, the Chief Judge in the Second Circuit-- Irving R. Kaufman--observed that many of the civil cases appealed to his court seemed amenable to resolution short of a decision on the merits by a panel of appellate judges. No court of appeals had ever made efficacious and systematic use of prehearing conferences to encourage informal dispute resolution. Rule 33 of the Federal Rules of Appellate Procedure seemed to provide sufficient authority for settlement discussion at the appellate level.<sup>3</sup> Would intervention by the court early in the appellate process effectively induce

---

<sup>2</sup>See Goldman, "Informal Dispute Resolution of Appellate Litigation: A Controlled Field Experiment" (mimeo, 1978).

<sup>3</sup>Rule 33, Prehearing Conference:

"The court may direct the attorneys for the parties to appear before the court or a judge thereof for a prehearing conference to consider the simplification of the issues and such other matters as may aid in the disposition of the proceeding by the court. The court or judge shall make an order which recites the action taken at the conference and the agreements made by the parties as to any of the matters considered and which limits the issues to those not disposed of by admissions or agreements

the parties to resolve their differences, and thus reduce the proportion of cases presented to the court for decision?<sup>4</sup>

In late 1973, Judge Kaufman wrote to Chief Justice Warren E. Burger about his experiences with the Rule 33 procedure:

To determine whether restrained and dignified encouragement by the court would facilitate settlement, I decided to experiment personally with the procedure [of preargument conferences authorized by Rule 33]. . . . Five cases were selected at random from among a group of cases which seemed to lend themselves to private dispute resolution and which were in the early stages of the appellate process. I met with the attorneys and, although my role was limited to that of catalyst, all five cases were terminated by settlements that were entirely satisfactory to the parties.<sup>5</sup>

Judge Kaufman's remarkable experience developed into a larger program designed to manage civil appeals. The program--as approved by the Circuit Council--would have the following components: one part would make each civil case subject to a scheduling order notifying the parties when certain stages in their appeal would take place (e.g., filing of the record,

---

of counsel, and such order when entered controls the subsequent course of the proceeding, unless modified to prevent manifest injustice."

<sup>4</sup>The theoretical justification for the use of the prehearing conference to resolve disputes without judicial decisions will be found in Mack, "Settlement Procedures in the U. S. Court of Appeals: A Proposal," 1 The Justice System Journal 17 (1975) (issue 2).

<sup>5</sup>Letter from Irving R. Kaufman to Warren E. Burger (Nov. 30, 1973).

filing of briefs, date for oral argument); the other part would systematically utilize Rule 33 by holding preargument conferences in selected appeals in order to explore settlement possibilities, to otherwise improve the quality of the appeal if it was to be argued, and to facilitate supervision of the appeal.

Judge Kaufman sought the Federal Judicial Center's financial support to experiment with this departure from appellate practice. The Center's funds would be used to hire a senior attorney and staff, whose major job would be to conduct the preargument conferences for settling or otherwise improving appeals. Judge Kaufman noted there was wide disagreement about the possible effectiveness of a staff attorney in the settlement process. Some critics felt only a judge would have sufficient prestige to resolve such disputes; others were convinced that a staff attorney could do the job. "Testing under controlled conditions," said Judge Kaufman, "may be the only way to settle questions of this nature."<sup>6</sup>

The board of the Federal Judicial Center, at its December, 1973 meeting, approved the Second Circuit's request for support of a one-year experiment to utilize the preargument conference procedure under Rule 33. The board allocated

---

<sup>6</sup>Id.

\$50,000 for the project<sup>7</sup> and the Center staff was charged with evaluating it. "The hope was expressed," according to the board minutes, "that the evaluation would be able to distinguish between cases which might have been settled or otherwise disposed of without intervention by the court."<sup>8</sup> The evaluation of the Civil Appeals Management Plan (CAMP) was begun with this mandate.

After reviewing many approaches, I proposed the evaluation be conducted experimentally, i.e., that cases meriting CAMP procedures be assigned by a truly random process to a number of groups. The division of all cases meriting CAMP attention into groups by a random process provides the greatest possible assurance that the groups are equivalent. This is so because a random process of division is blind to the characteristics of the cases. It is akin to flipping a coin, but much more exact. The administration of CAMP procedures, either separately or collectively, to one group (the "experimental" or "treatment" group), and the withholding of CAMP procedures from another group (the "control" group) provides the clearest proof that observed

---

<sup>7</sup>The board approved an additional sum of not more than \$40,484 in Jan., 1975 in order to continue the program during the evaluation. Since Center support was terminated, staff salaries have been budgeted through the Administrative Office of the United States Courts as part of the regular federal court appropriation.

<sup>8</sup>Federal Judicial Center, Minutes of Board Meeting 2-3 (Dec. 15-16, 1973).

differences between the experimental group and the control group are caused by CAMP procedures. For example, if none of the experimental cases reached a panel of judges for decision, but all of the control cases went to the panel, it could be said with some assurance that CAMP caused the reduction in panel considerations. Similarly, if half the cases in each group reached three-judge panels, it could be said with some assurance that the CAMP procedures had no effect. In a controlled experiment, such cause and effect statements are warranted because the experimental and control groups of cases are equivalent in all relevant respects except one: only the experimental cases are subject to the CAMP procedures.

The random assignment of cases to experimental and control groups also permits a precise calculation of the probability that differences between the groups are due to chance. The researcher's willingness to tolerate such differences is determined by the degree of risk he takes when he draws inferences from the data. This risk is dependent on the number of cases and the magnitude of sampling or chance fluctuation in each group. Probability estimates are essential when the researcher is faced with results falling somewhere between the two extreme examples offered in the preceding paragraph. Thus, the researcher's first task was to determine whether to accept the hypothesis



that CAMP is effective. The next step was to estimate the magnitude of CAMP effects.

How many groups of cases should be created for such an experiment? Obviously, one for each "treatment," and a control group to provide for a basis for comparison. In the CAMP experiment, it was necessary to specify exactly what a "treatment" was. The proposal submitted by Judge Kaufman, and the CAMP Rules adopted by the Circuit Council, suggested a variety of "treatments" or procedures worthy of experimentation.

The program was based on the use of two separate procedures: first, the use of a scheduling order to notify attorneys of deadlines in the processing of their appeals, with the threat of dismissal in the event of default; and, second, the use of Rule 33 preargument conferences to discuss settlement, withdrawal, or other matters that might improve the appeal if it should be decided by a panel of judges. Of course, CAMP emphasized the conference procedures, but it is at least arguable that the scheduling procedure would discourage some appeals. Hence, it seemed only reasonable to study the effects of each procedure separately and in combination.

Judge Kaufman introduced another variation by noting there was a division of opinion on the effectiveness of having a senior staff attorney conduct the Rule 33 conferences.

"Testing under controlled conditions," wrote Judge Kaufman, "may be the only way to settle questions of this nature."<sup>9</sup> This suggestion implies that in order to test the effectiveness of the senior staff attorney, the evaluation should include separate judge participation in the Rule 33 conferences. This would provide the needed proof of whether the senior attorney was more, less, or as effective as an appellate judge in reducing the proportion of cases that are fully briefed and argued, or in improving the quality of those cases that are briefed and argued.

A complete and exhaustive evaluation of CAMP would require a complex experiment (or series of experiments) in which eligible groups of cases would be given CAMP procedures separately and in combination, in order to assess the effectiveness of each procedure and of the combined procedures. Additional groups would be subject to Rule 33 conferences administered by a judge, to settle the issue of staff versus judge effectiveness. This complex experiment was presented to the court in April, 1973, and it was quickly rejected on the ground that "The design is fine for research, but poor for administration." This cryptic message was deciphered to mean that no judge would agree to participate in the test, in order to put to rest the a priori judgments of staff-versus-judge effectiveness.

---

<sup>9</sup>Supra note 5.

As an alternative, the Second Circuit suggested a simple one observation pre-test, one observation post-test interrupted time-series design. I rejected this offer as too ambiguous a demonstration; moreover, it was likely to lead to a negative judgment since there was substantial variability in the critical response variables measured across a longer time-series.

In an effort to strike a compromise, I suggested forfeiting the separate tests for each of the independent variables, leaving the rest of the combined variables intact. I also urged retention of staff and judge participation. This compromise was also rejected, at which point I suggested eliminating judge participation in the experiment despite Kaufman's claims about a priori judgments. Should the experiment with the staff attorney fail, I argued, a new experiment with judge participation could then be mounted. Throughout these negotiations, I did not waver from one crucial feature: randomization. In some form, a true experiment would be superior to any possible alternative.

During this negotiation stage--from April through August--the program was launched under the direction of the staff attorney. This occurred because the release of funds was not conditional on acceptance of the research design.

In September, the Second Circuit consented to a scaled-down version of the classic, controlled experiment.

This experiment would have two main components: (1) a single experimental group, in which eligible cases would merit both scheduling orders and Rule 33 conferences under the auspices of the senior staff attorney, now known as the staff counsel; and (2) a control group of eligible cases, in which both scheduling orders and preargument conferences would be withheld. Judge participation in the preargument conferences was sacrificed from the experiment. Opinion was still divided about the effectiveness of staff in relation to judge. An attempt to estimate the separate effects of each CAMP procedure was also eliminated from the study.

The Second Circuit was also concerned that the senior staff attorney hired by the court to run CAMP would be involved in only a portion of eligible cases. In short, the program would be difficult to justify if the senior attorney's energies were not fully consumed. Of course, I argued that in order to determine whether the senior attorney was effective, it would be necessary to establish appropriate comparisons to gauge that effect.

Another compromise had to be fashioned to obtain the court's consent to the experiment. In order to minimize the threat of underutilizing the staff attorney, eligible cases would be randomly assigned so that substantially more cases would be designated experimental than control. The chief disadvantage of this approach was that it would take longer

than originally contemplated to establish a control group large enough to test the program's effectiveness.

With the money in the hands of the court and with the CAMP up and running, why did the court consent to the experimental design. It would be naive to think that the compelling logic of experimental science won out over the quasi- and nonexperimental competitors. A more plausible explanation is that the court recognized that the initial funds would not allow sufficient time to adopt the cost of the program into the regular budget for the judiciary.<sup>10</sup> In order to keep the program alive, the court would have to make another request for funds from the Center. Thus if the court did not go along with the experiment, the probability of renewed funding would be predictably low.<sup>11</sup>

The lesson to be learned from this experience in convincing an institution to adopt a controlled experiment is that funds to the institution should not be released until the design details have been approved. Had this policy been in force when my involvement began, a much stronger experiment would have been launched with less delay than I needed to endure. It was also fortuitous that the Second

---

<sup>10</sup>See supra n. 7.

<sup>11</sup>In insisting on the randomization feature as the kernel of the study, I had the strongest possible support from the Center's Research Director, William B. Eldridge. Without that support, I doubt that an experiment would have been launched.

Circuit's proposal was cast in the experimental mold and that the Center's Board of Directors spoke with clarity on the isolation of cause and effect in mandating the CAMP evaluation.

Compromises must be fashioned in most field settings. I recognized the great advantage of randomization and refused the attempts to eliminate it. My compromise was to eliminate separate estimates for the effect of each procedure, but I retained the strongest test of the procedures at issue by experimenting with them in combination.

#### Random Assignment<sup>12</sup>

The random assignment of cases began on October 21, 1974. From early September through mid-October, procedures were devised to assure that true randomization would be achieved. The success of the experiment hinged on these procedures, yet there was no ready way to accomplish the task. Moreover, the staff attorney continued to obstruct the evaluation, and refused to permit his staff to be used for any evaluation purpose.

The procedure used here offers a breakthrough for experiments in which units to be randomly assigned (in this study, eligible civil cases) trickle into the court on a daily basis.

---

<sup>12</sup>Some of this discussion is drawn from Goldman, "A Randomization Procedure in 'Trickle-Process' Evaluations," Eval. Q. 493 (1977).

In most experiments, the units to be assigned are enumerated in advance and then randomly assigned to groups, but in this experiment, it was not known from one day to the next how many cases would have to be assigned or how and when to randomly divide them after they entered the appellate process. These were the choices:

1. One out of every four cases deemed eligible for CAMP by the staff counsel would be withheld from CAMP to establish the control group. This idea was rejected because it might give the program administrator considerable discretion to alter the equivalence of the controls to the experimentals. For example, perhaps some cases are very good candidates for settlement or withdrawal and others are not. Indeed, it was known before the start of CAMP that some appeals are settled or withdrawn. If the person responsible for the random assignment selected as control cases those that were unlikely candidates for settlement, and designated as experimental cases those that were likely to settle or withdraw anyway, then no doubt at the end of the experiment, there would be proportionally more control cases that were fully briefed and argued. The unwarranted conclusion would then be reached that CAMP caused a reduction in cases that otherwise would be decided by the court, when in truth this effect would be a result of the assignment procedure.

2. Another possibility was to use the last digit of

each case's docket number to determine the random assignment. But the cases would have to be screened to determine eligibility for the experiment.<sup>13</sup> Thus it was still possible--although unlikely--that the program personnel could alter the random assignment by providing different eligibility requirements for experimental cases than for control cases. This approach, too, was rejected because there was an increased risk that the assignment procedure might produce an unwarranted conclusion.

3. Yet another technique for achieving the random assignment was to accumulate a batch of eligible cases at fixed intervals (for instance, every week), and then have someone from the evaluation staff oversee the random assignment. This alternative was rejected for two reasons. It would have introduced delay in the processing of appeals, which the staff attorney viewed as unwise; and it would have tended to create distrust between CAMP personnel and Center employees, who would have been charged with overseeing the random assignment.

4. With all known conventional techniques eliminated for one reason or another, a technique was developed that assured truly random assignment but without supervision and its attendant costs. All civil appeals entering the Second

---

<sup>13</sup>Eligibility standards are discussed later in this section.



Circuit were reviewed after the appropriate forms were filed and, in nearly all circumstances, the docketing fee paid.<sup>14</sup> Once these threshold requirements were met, the case materials were then examined by the staff attorney. If, in his judgment, a case merited both a scheduling order and a preargument conference, it entered the pool of eligible cases for random assignment.

Some may wonder why there was not a more specific eligibility criterion, such as a money judgment for the plaintiff in the district court. The staff attorney argued that there were many factors to consider in deciding to apply CAMP procedures, especially the preargument conferences. Some cases met a few requirements, others met more. Yet there was no calculable, uniform, and objective standard that, when applied to all cases, would separate the eligible from the noneligible cases. Indeed, CAMP was designed to permit this flexibility.<sup>15</sup> A handbook on appeals in the Second Circuit describes the process of selection:

The staff counsel [i.e., staff attorney] will make the determination as to whether or not the case is

---

<sup>14</sup>One form provides information about the nature of the case, its disposition in the district court, and, to some extent, the issues to be raised on appeal. A second form provides information on the ordering of the transcript. These forms must be filed and the docket fee paid within 10 days of the filing of a notice by appeal in the district court, with dismissal by the clerk in the event of default (CAMP rules 3 and 7(a)).

<sup>15</sup>CAMP Rule 5(a).

appropriate for a preargument conference on the basis of his study of [the forms], and a copy of the docket sheet from the District Court. Such a conference will normally be held in a private action seeking a monetary judgment, and in other actions which, in the judgment of staff counsel, seem susceptible to settlement or simplification of issues.<sup>16</sup>

Rather than impose arguable, objective standards as part of the experiment, the decision as to eligibility was left to the staff attorney. Under most conditions in the evaluation, the extent to which he would err in his judgment by including too many or too few cases did not matter, since more of the experimental than the control cases were expected to terminate short of panel consideration. Of course, if the pool of cases deemed eligible by the staff attorney contained a substantial number that did not merit CAMP procedures, the program's effect would tend to be masked. It was reasonable to expect that the staff attorney's identification of eligible cases would be based on the strong likelihood that CAMP would lead to settlement, withdrawal, or improvement in quality of those cases.

The eligibility issue was not ignored, however. It was expected that the staff attorney would learn from his experience at the eligibility stage and, over time, sharpen his decisions. The experiment tested this "learning curve"

---

<sup>16</sup>Appeals to the Second Circuit 15-16 (1975)  
(prepared by the Committee on Federal Courts of the Association of the Bar of the City of New York).

hypothesis in order to minimize possible concern over the eligibility decision.

Following the staff attorney's decision that a case merited both a scheduling order and a preargument conference, a staff member from the circuit executive's office would enter the docket number with the date in a log book. The Research Division of the Center maintained a duplicate log book in Washington, but with one important difference. Each line in this log book had been designated as a control or an experimental unit. When the staff member in New York completed his log entry, he would call the Center to transmit that information to the duplicate log. Only after the docket number and date were entered in Washington was the designation of experimental or control released to New York. This technique provided the greatest possible assurance that the random assignment had been made objectively.<sup>17</sup>

Cases entered the list in chronological order. If the ratio of experimental cases to control cases remained fixed for serially ordered sub-sets on the list, then it would be possible to test hypotheses based on time dependent effects. For example, the staff attorney might improve his effectiveness over time. At the start of the experiment, he might have little effect; but at the end, his effect might be

---

<sup>17</sup>I am indebted to Anthony Partridge, who first suggested this solution to the 'trickle-process' problem.

considerable. Separate analyses of the sub-sets permits a judgment on this "learning curve hypothesis." It is essential, however, that the ratio of experimentals to controls remain fixed across sets. This can be accomplished without difficulty by randomly assigning serially ordered sub-sets of the list. These can later be aggregated.

In summary, the double-list procedure assures randomization at minimal cost in experimental settings where units to be assigned "trickle in." Moreover, the procedure permits a test of time-dependent hypotheses by structuring the randomization into a series of separate, but comparable, replications.

#### Confronting the Ethical Issues

Can these appellate procedures be denied to anyone who is deemed eligible to receive them? This issue is akin to the problem faced by medical experimenters who must weigh the implications of withholding a theoretically valuable therapy from patients in order to estimate effects. The problem is answered by the use of informed consent, i.e., by explaining the nature of the experiment to the patient and the risks and benefits involved from the therapy and its absence (or, more likely, a competing therapy). But the analogy to the medical experiment is not apposite in the CAMP experiment. First, the conference procedure remains

entirely at the discretion of the court. Counsel would have to claim a right to policies that are solely within the court's power to dispense. Second, the theoretical plausibility of the CAMP procedures has no empirical basis to support it; indeed, what evidence does exist about pretrial procedures would suggest that the preappeal conference would be ineffective.<sup>18</sup>

A question still remained whether attorneys in the control cases should be notified that those cases were not to be subject to CAMP procedures. The proponents of notification took the position that CAMP had been in operation for nearly six months. During this period, some unknown number of attorneys could have altered their expectations about Second Circuit procedures to the extent that they might violate the Federal Rules of Appellate Procedure in anticipation of a CAMP scheduling order or a preargument conference.

I argued in opposition that the notice would affect attorney behavior by encouraging greater attention to the Federal Rules of Appellate Procedure and the local rules, thus altering the control cases, which should ideally reflect only the absence of CAMP. In weighing the possibility of introducing positive bias (in experimental research, this is known as the Hawthorne effect) in relation to the possibility

---

<sup>18</sup>Rosenberg, The Pre-Trial Conference and Effective Justice (1964).

of jeopardizing the appeal because of Federal Rules of Appellate Procedure violations, the importance of the notice outweighed the bias it might introduce.

This notice excluded control cases from the scheduling order requirement for all civil appeals.<sup>19</sup> Since the CAMP Rules left convening the preargument conference entirely to the staff attorney's discretion, it was unnecessary to mention withholding the conference in the notice.<sup>20</sup> But what would happen if an attorney in a control case requested a scheduling order or a conference? Could the court withhold procedures deemed by counsel to be beneficial to an appeal? I would have urged the court to deny the request, but, fortunately, this problem never arose. Perhaps it was a signal as to the efficacy of the procedures themselves that no attorney requested them.

There were circumstances, although few in number, when cases deemed eligible for the experiment were not included in the test. These were appeals in which the court believed that the issues were of such moment or the matters were so urgent that designation to the control group might--if the program really worked--pose a threat to the justice of the appeal. When a case of this magnitude arose, it was excluded from the experiment entirely. Fortunately, this

---

<sup>19</sup>CAMP Rule 4(a).

<sup>20</sup>CAMP Rule 5(a).

occurred so infrequently (not more than five times during the year), that these exclusions from the experiment do not bias the judgment reached regarding CAMP effects on the nonexceptional cases.

In summary, the ethical problems in this experiment were resolved by informing attorneys in the control cases of the experiment, but avoiding the matter of consent, which did not seem to apply. The potential issue of selection bias (due to attorneys requesting treatment in control cases) was begged intentionally, never to be squarely addressed. And the removal by the court of exceptional cases in the experiment avoided the "symbolic" issue of the court withholding "just" procedures from appeals that were in the public eye.

#### Estimating Error Type II

Perhaps it bears repeating that the random assignment of cases to experimental and control categories provided the greatest assurance that the two groups were equivalent in all respects save one: CAMP procedures applied to the experimental cases only. Hence, beyond a certain point determined by the laws of chance, observed differences between the experimental and control groups warrant a conclusion that CAMP is effective. In short, when the difference between the two groups of cases is sufficiently large, it can be said with some confidence that CAMP procedures were

responsible for a particular effect, such as a reduction in briefed and argued appeals or an increase in the quality of appeals.

Precisely how are such conclusions reached? The first step is the formulation of a hypothesis, i.e., a statement that a certain situation might be true. An alternative hypothesis, which would necessarily be true if the first hypothesis is rejected as false, is also formulated. The next step is to examine the empirical evidence on the assumption that the initial hypothesis is true. If the evidence would be highly unlikely under the assumption, the initial hypothesis is rejected, and its alternative is accepted.

One hypothesis was that CAMP has no effect on the proportion of briefed and argued cases.<sup>21</sup> (The alternative hypothesis was that CAMP has an effect on the proportion of

---

<sup>21</sup>The "no difference" or "no effect" starting point is a common feature of scientific research.

"This seems like an extremely devious way of proceeding, but we must remember that we shall not be in a position to establish directly that there is a difference [between groups]. To avoid the fallacy of affirming the consequent, we must proceed by the elimination of false hypotheses. In this case there are logically only two possibilities, there either is or is not a difference. If the latter possibility can be eliminated, we can then conclude that some difference in fact exists."

Blalock, Social Statistics 95 (1960).



briefed and argued cases.) If the empirical evidence is consistent with the initial hypothesis, it stands. If the evidence is inconsistent with this hypothesis, it is rejected in favor of its alternative. For example, if the evidence is that 50 percent of the experimental cases and 50 percent of the control cases were briefed and argued, the initial hypothesis (that CAMP has no effect) probably should be retained. If the evidence is that 40 percent of the experimental cases and 75 percent of the control cases were briefed and argued, the initial hypothesis probably should be rejected in favor of its alternative. It is also possible that the evidence might not squarely support either the initial hypothesis or its alternative. In that case, a judgment about program effects would be suspended.

In general, the greater the difference between groups, the less likely that the initial "no effect" hypothesis remains valid. But at what point is the initial view rejected? There is no clear and convincing answer to this question. By convention, most social scientists claim that, given the initial assumption, if the likelihood of observing a difference between groups is less than 5 times in 100, the assumption should be rejected. There is nothing sacred or absolute in the standard of less than 5 times in 100, but there are strong reasons for having adopted this convention in the CAMP experiment.

When a decision to reject or to accept the initial hypothesis is made, the researcher must face the possibility of making either of two errors: rejecting the initial (no effect) hypothesis when it is in fact true (error type I); or accepting the initial hypothesis when it is in fact false (error type II). The 5-in-100 standard minimizes the first error; and, in general, the sizes of the experimental and control groups minimize the second. For social programs, the first error seems to be more threatening than the second. Keeping the potential for the first error small protects against drawing the false inference that CAMP is effective when in fact it is not.

Of course, it is possible to err by concluding that CAMP has no effect when in fact it does. For experiments in court procedures, however, this second error may be less critical "since the more important policy problem would seem to be how to avoid the disappointment, frustrated effort and wasted resources caused by making [the first error], that is, adopting an ineffective treatment as a social program."<sup>22</sup>

Nevertheless, an estimate of error type II seemed essential in order to design the experiment properly. If the experiment were designed in the absence of type II error calculations, results deemed not significant on the basis of error type I, might pose considerable risk in terms of

---

<sup>22</sup>Social Experimentation: A Method for Planning and Evaluating Social Intervention 77 (Reicken & Boruch eds. 1974).

error type II. But an estimate of error type II cannot be calculated until some anticipated "true" effect is specified. To put this another way, someone must specify how large an effect would be important to identify (some policy-significant minimum perhaps) in order to assure an experimental design that minimizes errors of both types.

How large should the CAMP experiment be? There were two main constraints in answering this question. First, the Center's Board approved a one-year experiment, but without contemplating whether this would be sufficient to minimize the probability of mistaken judgment. And, second, the Second Circuit insisted on a disproportionate division of cases to give the staff attorney enough work, at the sacrifice of error estimates in the control group. When I put the question of anticipated effects to the court, no one would answer. The fear was that an inordinately high estimate of policy-significant effect would pose too demanding a test, and thus the court would have to accept the conclusion that the program did not measure up to its minimum expectations. But the court was equally hesitant to offer a low estimate of policy-significant effects since it had publicly proclaimed that the procedures would have a substantial impact. In the absence of guidance from those who should have rendered it, I set out on my own.

I estimated that 300 cases divided 225 experimentals

and 75 controls would fit within the constraints imposed by the Board and the court. I also calculated a conservative estimate that an observed difference of 10% between groups would permit a rejection of the no-difference hypothesis (with error type I fixed at less than .05). And I also reasoned that in absence of rejection of the no-difference hypothesis, error type II would be less than .50 if the anticipated effect of the program were at least equal to the 10 percent point of rejection.<sup>23</sup> Moreover, if the court's estimated effect of its program exceeded 10 percent, the probability of error type II would decline. Judge Kaufman had gone on record claiming an anticipated effect of 15 percent,<sup>24</sup> which meant that inability to reject the null hypothesis would mean a .25 percent probability of falsely concluding that CAMP failed. A year later, Kaufman offered an even grander estimate of program effectiveness, upping his claim to 25 percent.<sup>25</sup> Based on this claim, failure to reject the no-difference hypothesis would mean about a .01 probability of falsely concluding that CAMP failed.

---

<sup>23</sup>See Hays, Statistics for the Social Sciences 357-386 (2d ed., 1973).

<sup>24</sup>Kaufman, "The Pre-Argument Conference: An Appellate Procedural Reform," 74 Colum. L. Rev. 1094, 1100 n. 17 (1974).

<sup>25</sup>Kaufman, State of the Judicial Business in the Second Circuit 10-11 (1975) (unpublished address to the Judicial Conference of the Second Circuit, Sept. 1975).

Once again, fortuitous circumstances aided the design, in this case a justifiable estimate of type II error, a matter that all too often is shoved under the rug by social scientists. Of course, it should be noted that the constraints placed on the experiment by the Board and the court limited the precision of the experiment. If the court had claimed that a difference of less than 10 percent was nevertheless important to detect, then a much larger (and longer) experiment would have been required to keep error type II below .50. But of course a longer experiment might have been rejected as too costly to undertake.

My estimates proved reliable, and over one year 302 cases were randomly assigned to treatment and control groups. It would still take nearly 18 months to work these cases out of the court and render a judgment on program effects. (I took the Board's one-year limit as a constraint on random assignment of cases. A more restrictive view would have surely doomed the experiment.)

With more than my share of good fortune and support, the first controlled experiment in the appellate courts became a reality. It is my hope that these experiences can guide other researchers who may wish to mount legal experiments in field settings. The problems in design and execution are difficult, but the rewards in terms of elegance

and clarity are substantial.

#### Postscript

The CAMP program was enlarged shortly after the final report on the experiment was submitted in June 1977. The program continues to operate as it has since its inception in April 1974. That it does so in spite of the evidence does not betoken a flaw in design or execution of the experiment. Why has the court ignored the results of this experiment? This question is best addressed in a session on the politics of socio-legal research.



**END**