University of Chicago Law School

Chicago Unbound

Journal Articles

Faculty Scholarship

1974

Reflections on Experimental Techniques in the Law

Hans Zeisel

Follow this and additional works at: https://chicagounbound.uchicago.edu/journal_articles

Part of the [Law Commons](#)

Recommended Citation

Hans Zeisel, "Reflections on Experimental Techniques in the Law," 15 *Jurimetrics Journal* 256 (1974).

This Article is brought to you for free and open access by the Faculty Scholarship at Chicago Unbound. It has been accepted for inclusion in Journal Articles by an authorized administrator of Chicago Unbound. For more information, please contact unbound@law.uchicago.edu.

REFLECTIONS ON EXPERIMENTAL TECHNIQUES IN THE LAW*

Hans Zeisel†

Much of the law's reasoning turns on the effects of the rules it creates or enforces. Normally these effects are alleged or implied, but rarely proven. Yet during the last decades, first the natural sciences and, learning from them, the social sciences have developed techniques of experimentation that would allow firmer determinations of what a rule of law does or does not accomplish.¹

The basic tool (and also the paradigm for retrospective analysis) is the controlled experiment. Its essence is simple: The objects or people on which the experiment is to be performed are divided randomly, that is, by some lottery device, into two groups; the experimental treatment is then applied to the one group and withheld from the other, the control group. If an effect is found in the experimental group that is absent from the control group, it can be attributed to the experimental treatment, since at the time of the prior random selection the two groups had been strictly comparable, and in fact interchangeable.² And randomization assures that whatever the outcome of the experiment, it is possible to calculate the degree of certainty that can be attached to it.

Several controlled legal experiments have been or are being conducted, as we shall see; but sometimes the controlled experiment, employed in its pure form for legal investigations, encounters obstacles on the part of the law. Since by definition it gives something to one group

*Reprinted with permission from the January, 1973 issue of *The Journal of Legal Studies*. Copyright © 1973 by *The Journal of Legal Studies*.

†Hans Zeisel is Professor Emeritus, University of Chicago Law School, and serves as Consultant to the American Bar Foundation and as Vice President in charge of Research, Vera Institute of Justice. The author wishes to express his indebtedness to his colleague William H. Kruskal for his critical reading of the paper.

¹E.g., Ronald A. Fisher, *The Design of Experiments* (7th ed. 1960); D. R. Cox, *Planning of Experiments* (1958).

²While the principle of the controlled experiment is simple, the statistical details of its design and execution comprise a considerable body of learning which in this context we can allude to. Cf. William G. Cochran, G. E. P. Box & Donald T. Campbell, "Experimental Design," 5 *Int'l Enc. Soc. Sci.* 245 (1968).

and withholds it from another, it would seem to violate the basic guarantee of equality before the law. For example, while one might want to know what difference it makes to the fate of a defendant if he can afford to hire his own counsel, one can hardly undertake an experiment in which one group of defendants is assured of individual counsel and the other group compelled to accept a public defender. To take another example, one might like to know what effects it has on the future of convicted defendants if instead of being sent to prison they are put on probation, or vice versa. An attempt to answer this question through a controlled experiment in which these two types of sentence were assigned randomly might again encounter objections.

Several circumstances, nevertheless, argue in favor of the experiment, even in legally sensitive areas. First, it is by no means certain that the experimental treatment makes any difference; the uncertainty may in fact be the reason for the experiment. Second, discrimination in an experiment is by definition only temporary; after it is concluded and the best solution is found, the solution will be applied to all.³ Finally, there are areas at the fringe of the law where random assignment is more tolerable because the right involved is not sufficiently important to merit special protection.

Certain modifications have been developed that make controlled experimentation tolerable even in sensitive areas. They operate in three dimensions: abandonment of direct control of the experimental treatment in favor of the indirect experiment, simulation of part of the experimental design, and the use of natural experimental conditions.

THE INDIRECT EXPERIMENT

The essence of the indirect experiment is that it removes direct control over the experimental variable from the investigator while maintaining the controlled character of the experiment, albeit at a price. Three examples follow.

1. The first of the indirect experiments concerned a procedural device designed to reduce the trial load of the courts. Normally, a claimant in our civil courts presents his case both as to liability and size of damages; he is followed by the defendant who presents evidence in rebuttal of both claims. After the parties have had their say the jury retires and decides whether the defendant is liable for damages, and if so how large these damages should be. The question of damages thus becomes relevant only if liability is found. It was proposed⁴ to try the liability issue in its entirety first and ask the jury to decide whether the defendant owed anything at all. Only if liability was affirmed would the trial continue, and in that event the jury would render a second verdict on the amount of damages. Since liability is found in only slightly more

³For a more elaborate statement see "The Case for the Official Experiment," in Hans Zeisel, Harry Kalven, Jr. & Bernard Buchholz, *Delay in the Court* ch. 21 (1959).

⁴Hans Zeisel, Harry Kalven, Jr. & Bernard Buchholz, *supra* note 3, at 99.

than half of all cases, this mode of trial was expected to dispense with about half of all damage litigation. The U.S. District Court for Northern Illinois was sufficiently intrigued by this idea to permit by rule the conduct of split trials, and asked the University of Chicago Law School to assess its effect on the workload of the court.

The ideal design for this investigation would have been random assignment of cases to the traditional and split mode of trial. But this the judges, quite properly, considered their decision to make. Once random assignment is placed by deliberate judicial decisions, the difficulties begin. Suppose the judges decided, sensibly enough, to order a split trial whenever they believed that the jury was likely to deny liability. Such a principle of selection would make it impossible to compare the duration of the split trials with that of the normal ones, because we would know that the split trials concerned different kinds of cases to begin with.

At first glance, the lack of a control group seemed to vitiate the experiment. But by slightly shifting the focus of the inquiry, it was possible to save the integrity of the experiment. As the cases come to the court they are randomly assigned to the individual judges. This means that the cases coming before Judge A would not differ, in the long run, from the cases coming before Judge B. And it so happened that the various judges of the court made differential use of the split rule, in that some used it for almost all their cases, some for hardly any, and some for varying proportions in between. If it were true that the application of the split trial rule saved time, the judges who applied the rule more often should require on the average less trial time than those who applied it less often.

This turned out to be the case, as the following table shows:⁵

TABLE 1
Proportion of Split Trials and Average Trial Time in
Personal Injury Trials

Judge	Proportion of Cases Tried under Split Trial (Per Cent)	Average Length of All Trials Before This Judge (Days)	Number of Trials Before This Judge*
A	89	3.2	(26)
B	51	3.3	(41)
C	38	3.5	(26)
D	14	3.8	(22)
E	7	3.9	(27)
F	7	4.3	(14)

*Only judges with more than 10 trials were included.

The regression line based on these data indicated that if a judge were to conduct all trials under the split trial rule, his average trial time would be cut by about 20 per cent. This would be, then, the amount of time that could be saved through general application of the split trial rule.

This particular experimental design, however, contained a hidden

⁵Hans Zeisel & Thomas Callahan, "Split Trials and Time Saving: A Statistical Analysis," 76 *Harv. L. Rev.* 1606 (1963).

flaw. Conceivably, the judges who tried their cases in general more expeditiously might also have made more frequent use of the split trial rule. Such self-selection would have biased the results of the experiment. Supplemental evidence was therefore adduced. It was clear that any savings had to come from the elimination of the damage trial, so the frequency of damage trials was determined both for the regular and for the split trials, as shown in Table 2.

TABLE 2
Disposition of Cases in Regular and Split Trials

	Regular Trials	Separate Trials
	(Per Cent)	(Per Cent)
Complete trial on liability and damages	76	15
Trial ended in liability verdict	—	58
(a) Because verdict was for defendant	—	43
(b) Because damages were settled after verdict on liability	—	15
Other dispositions (settlement during trial, directed verdicts)	24	27
	100%	100%

The dispositions of the two groups of cases differ drastically: 76 per cent of all regular cases go through a complete trial as against only 15 per cent of the split trials. In 58 per cent of the split cases, trial of the damage issue was avoided *because* of the intermediate liability verdict. In 43 per cent of the cases the trial simply ended when liability was denied and in another 15 per cent there was no trial of the damages because the case was settled after the jury had found liability. The convergence of the two sets of data left no room for doubting the efficacy of the split trial rule.

2. A similar problem arose when the courts of New Jersey decided to appraise the value of what has become known as pretrial. In most of our courts civil suits (occasionally also criminal cases) before they come to trial are scheduled for pretrial. There, counsel for both sides, occasionally with their clients present, meet with the judge to present briefly the issues under dispute and air the possibilities of a settlement. Tradition has it that the pretrial facilitates settlement and hence is a most desirable means of reducing the trial load and thereby the congestion of our metropolitan courts.

But analysis of the available evidence raised doubts as to whether pretrial in fact saved court time since, to begin with, it uses up court time. We suggested that a controlled experiment could provide the answer.⁶ At the time, the New Jersey courts had a rule that made pretrial obligatory. The state's distinguished Chief Justice and the Administrative Director of the Courts became sympathetic to the idea and commissioned Professor Maurice Rosenberg, then Director of the Project for Effective

⁶Also in Hans Zeisel, Harry Kalven, Jr. & Bernard Buchholz, *supra* note 3, at 154.

Justice at Columbia University, to design and conduct the experiment.⁷

Again, the ideal experimental design was simple: randomly assigning one part of the cases to pretrial, omitting pretrial with the other part. But there was concern as to the constitutionality of a procedure that denied the right to pretrial to a group of litigants who could not be distinguished on the merits from the members of the group to which pretrial was granted.

A compromise was agreed upon that allowed random assignment of one half of the cases to obligatory pretrial and the other half to optional pretrial, where pretrial would be held only if at least one of the litigants requested it.

This design, which allowed the litigants and counsel to have a pretrial whenever they wanted it, had from the experimenter's viewpoint two shortcomings. It blunted the objective of the experiment by comparing obligatory with optional pretrial, and therefore allowed no direct inference as to the effect of pretrial itself. And, of course, if all litigants in the optional group had opted for pretrial, we would have learned nothing.

The experiment was conducted and 2,954 cases assigned randomly to the two groups. Only about one-half of all litigants in the optional group asked for pretrial, compared with the obligatory group where all cases were pretried. The results of the experiment were as follows:

	<u>Obligatory Pretrial</u>	<u>Optional Pretrial</u>
Suits settled before they reached the trial stage	76%	78%

The conclusion emerged simple and clean: contrary to the widely held belief, obligatory pretrial failed to increase the settlement ratio; there were just as many settlements without it. From the point of view of reducing the number of trials, obligatory pretrial was found to be a waste of time, and the State of New Jersey forthwith changed its rule and made pretrial optional.

3. The third experiment had a major impact on the criminal law. It revolutionized one of its most solid traditions: the practice of setting bail for arraigned defendants. Bail is set, as a matter of constitutional right, for virtually all defendants. If they can post it, they are set free; if not, they must remain in jail. Whether or not they *can* post it as a rule depends—since criminal defendants seldom have substantial sums of money—on the bondsman who, against a premium of some ten per cent will or will not take the risk of providing the demanded bail.

The system has been criticized because it favors the well-to-do and the criminal accused who work for the well-to-do, because it surrenders the actual power of decision to the bondsman, and because it keeps an inordinate proportion of defendants in jail, many of whom are subsequently acquitted. Nevertheless, the system stood fast until the creation of a foundation which conducted a controlled experiment with the cooperation of the New York judiciary and the assistance of New York

⁷Maurice Rosenberg, *The Pretrial Conference and Effective Justice* (1964).

University Law School.⁸ All defendants arraigned in the felony court of Manhattan were interviewed to assess the risk of their failing to appear at their trial if the court freed them without requiring bail. On the basis of these interviews, the defendants were classified into two groups: those for whom such a release without bail could be recommended to the court, and those for whom such a recommendation could not be made. The recommendable group was then divided at random in half. For one of the halves the recommendation to release the defendant without bail was actually transmitted to the arraignment judge; with respect to the other half, the judge was told nothing. In all cases the judge made the ultimate decision as to bail, in the one group guided by the recommendation, in the other group without such guidance. In this latter group only 14 per cent of all defendants were freed without bail, as against 60 per cent in the recommended one-half. The remaining question was whether at the time of trial a larger proportion of the group of which 60 per cent had been freed without bail would fail to appear in court than of the group where only 14 per cent were freed without bail. When trial time came, in both groups not more than one per cent of all defendants deliberately failed to appear. The experiment demonstrated that the number of defendants released without bail could be approximately quadrupled without increasing the likelihood of defendant's not appearing at the trial.

The aftermath of this experiment was dramatic. The City of New York adopted the interviewing procedure as a permanent feature of the criminal justice system; and today most major cities and many rural areas have adopted the interviewing procedure and with it the practice of release without bail. The experiment had fallen on usually fertile soil. Everyone—except the bondsmen—stood to profit from the liberalization: the municipalities by saving money on jails; the defendants by being spared unnecessary hardship; not least, the ends of justice were advanced at a point where the traditional approach was ripe for reform.

In all three of these experiments the direct control over the experimental variable had been left to the judges or the litigants: The decision whether or not to hold a pretrial was left to the litigants; the decision whether or not to order a split trial was left to the judge; and so was the decision whether or not to release a defendant without bail. But because of the prior random assignment of the cases to experimental and control groups, all three studies retained their character as controlled experiments. Two disadvantages result from such indirect control. The first is a lack of directness in the measurement. The New Jersey experiment permitted no direct observation of the effect of pretrial itself; it measured only the difference between obligatory and optional pretrial. More importantly, the lack of direct control could have vitiated the experiment. If all litigants in the optional group had insisted on pretrial, if all judges in the split trial experiment had made use of it with the same frequency,

⁸Louis Schweizer founded what is now the Vera Institute of Justice in 1961. Its first major effort was the Manhattan Bail Project. See Vera Inst. of Justice, *Programs in Criminal Justice Reform, Ten-Year Report 1961-1971*, at 18-43.

and if the judges in the bail experiment had paid no heed to the Vera interviewers' recommendations, there would have been no difference between experimental and control group and hence no experiment.

This shift in the goal, coupled with the risk of failure, are the price to be paid for the privilege of conducting an experiment in an area where the law cannot surrender certain decisions.

We may note here in passing that the same difficulty and the same solution offer themselves where the experiment involves decisions that for other than legal reasons cannot be imposed on the subjects, such as submission to an experimental medical treatment. Here too, the indirect experiment has found its place.⁹

SIMULATION

If indirect experimentation is one line along which the experimenter can meet the concerns of the law half-way, simulation is the other. The three experimental sequences we now turn to fall into that category. They were not undertaken by the law itself, although assisted at crucial points by the judiciary or the litigants. They all came out of the University of Chicago Law School study of the American jury.

The first sequence concerned the broad issue of the difference it would make if all cases now going to the jury were decided by a judge sitting without a jury. The second experiment was designed to explore the jury's handling of the defense of insanity, if instructed under differing rules of law. The third experiment set out to determine whether in civil cases different juries in different parts of the country, given an identical set of facts, award different amounts of damages.

1. The question of the difference between jury and judge verdicts would seem to demand a controlled experiment, where every case is tried twice, once with and once without a jury—an obviously impossible solution. Yet this is almost precisely the experiment that was performed, if one permits the "two trials" to take place simultaneously, since every jury trial takes place also before a judge, the one who presides over it. Hence, all that had to be done was to ask a nationwide sample of trial judges to reveal, for a sample of their trials, how the jury had decided the case and how he, the judge, would have decided it without a jury.

The results of this study, in criminal cases, have been published in *The American Jury*, where the following table summarizes the pattern of agreement and disagreement between jury and judge.¹⁰

It appears that judge and jury agree in $(13 + 62 =) 75$ per cent of all cases; and of the 24 per cent disagreement cases the jury is on the defendant's side in $(17 + 5 =) 22$ per cent, and on the prosecutor's side only in $(1 + 2 =) 3$ per cent. The survey asked, of course, for information

⁹Cf., e.g., *The National Diet Heart Study* (37 *Circulation*, Suppl. No. 1, March 1968) in which three different sets of diets (in the form of actual grocery baskets) were given to three randomly determined groups with the expectation that they would be used. The final decision, of course, remained with the consumers.

¹⁰Harry Kalven, Jr. & Hans Zeisel, *The American Jury* 56 (1966).

TABLE 3
Agreement and Disagreement Between Jury and Judge in Criminal Trials
 (as % of all trials)

Judge would have:	Jury		
	Acquitted	Convicted	Hung
acquitted	13	2	1
convicted	17	62	5

Total = 100%
 Number of Trials (3576)

☒ Agreement

beyond the two verdicts; the questionnaire contained some fifty-odd questions aimed at a detailed description of the case, so that we could understand the reasons why the jury differed from the judge in the one out of four cases where they disagreed. And it is, of course, the adequacy or inadequacy of these reasons on which the merit of the jury verdict rests.

2. The second experiment in which simulation played a major part concerned the competing insanity instructions. In the Anglo-American system the defense of insanity has been embodied for more than a century in the so-called *M'Naghten* rule, under which the defendant is acquitted either if he did not know what he was doing or if he did not know that what he was doing was wrong. In recent years, the rule has come under increasing criticism, primarily from psychiatrists, and in 1954 the U.S. Court of Appeals for the District of Columbia established a new rule in the case of one *Durham*. Under the *Durham* rule, the defense of insanity was established if the criminal act was shown to be the "product of a mental disease or a mental defect."

For some time, the two rules competed for the favor of the nation's judiciary and it was of obvious interest to learn what difference the rule made in actual outcomes.¹¹ The "law" in a criminal jury trial becomes operative through the judge's instruction to the jury, in which the judge spells out the conditions under which the jury is to either acquit or convict the defendant. The question, therefore, as to what difference the law makes means what difference it makes to the jury whether the judge instructs it according to the rule in *M'Naghten* or according to *Durham*. The rule may affect a variety of facets of the trial, but obviously the main question is how the two rules affect the defendant's chances of acquittal.

To test the issue,¹² two trial records were composed, a case of housebreaking, a simplified version of the original *Durham* trial; and an

¹¹In the meantime the U.S. Court of Appeals for the District of Columbia Circuit has modified the *Durham* rule, first in *McDonald v. United States*, 312 F.2d 847 (1962), and more recently in *Browner v. United States* (Jan. 23, 1972). The new rule, conforming to that adopted by other federal courts of appeals, is that of the American Law Institute's Model Penal Code and provides that "a person is not responsible for criminal conduct if . . . as a result of mental disease or defect he lacks substantial capacity to appreciate the wrongfulness of his conduct or to conform his conduct to the requirements of the law."

¹²Rita James Simon, *The Jury and the Defense of Insanity* (1967).

incest case, also an abbreviated version of an actual trial. In both trials the accused's only defense was insanity. The trial evidence was acted out and, with the other elements of the trial, put on recording tape. Three main variants of each case were produced. The tapes were identical except for that part of the judge's instruction that dealt with the defense of insanity and for the concomitant psychiatric testimony. In one version the instruction and psychiatric testimony were according to *M'Naghten*; in the second according to *Durham*; and in the third the instruction left it to the jurors' own judgment as to whether the evidence in the cases supported a defense of insanity, thus forcing the jury to establish its own law of insanity.

Each of the three versions was then taken into two major metropolitan courts and presented to some 100 juries in turn. A judge called the jurors into his court room and asked them to cooperate in the experiment; by doing so, they were advised, they would oblige the court and also discharge their turn of jury duty. The jurors would then listen to the taped trial and afterwards deliberate and find a verdict. Following is the outcome of the experiment in terms of the jurors' vote on their first ballot, prior to the beginning of the deliberation.

TABLE 4
Percentage of Jurors Voting "Not Guilty-Insane" on First Ballot

	M'Naghten	Durham	No Rule
Housebreaking	57 (120)	65 (120)	76 (120)
Incest	24 (240)	36 (312)	34 (264)

() = Number of jurors

In both trials, the *Durham* rule elicited a higher percentage of acquittals by reason of insanity than the *M'Naghten* rule. That the percentages under *Durham* are very close to those obtained under the "No Rule" instruction suggested—as indeed it has been argued—that *Durham* comes close to being no rule.

3. The last one in this series of experiments involved a complete simulation. Nevertheless, it was probably a very reliable experiment. There is a notion abroad among lawyers that juries in some parts of the country are likely to award larger damages, for a given injury claim, than juries elsewhere. But there was no precise knowledge concerning either the geographic pattern or the degree of variation.

Since there are some sixty thousand civil jury trials held each year in the United States, it seemed tempting simply to compare the awards made in various jurisdictions. But this proved unprofitable partly because such trial statistics were unavailable, but primarily because the variety of injury cases coming before the courts is so great that to compare "average" award levels would be largely meaningless. Instead, a controlled experiment was designed by submitting simulated cases to the representatives of major insurance companies in various parts of the country.¹³

¹³The complete report forms a chapter in Harry Kalven, Jr. & Hans Zeisel, *The Civil Jury* (forthcoming).

Five personal injury claims were composed by slightly modifying and simplifying five actual personal injury reports. They were written up in précis form such as is normally submitted to claim adjusters. The cases were submitted to the claim adjusters with this question: "How much, judging from your experience, would you expect a jury in your court to award for this case?" The question was asked for a number of courts, which constituted a probability sample of regions and community sizes of the United States. To offset personal differences, the requests were submitted to three adjusters in each community, one from each of the three insurance companies that cooperated in the experiment. The following table gives the average of the (3 adjusters $\times$ 5 awards =) 15 awards in terms of deviations from the national average, summarized for major census regions and city sizes:

TABLE 5
Geographic Variation of Jury Awards for Identical Claims
(In per cent deviation from the national average = 100)

City Size	Region			
	West	Midwest	South	East
Large	+20	+2	0	+19
Medium	+8	-11	-9	+10
Rural	0	-21	-15	-6

Awards for identical claims vary between roughly 80 per cent of the national average in the rural South and Midwest to roughly 120 per cent of that average in the large cities on the East and West coast. Translated into a formula, the table says: add 10 per cent to the average if the trial takes place on the East or West coast and another 10 per cent if it is conducted in a large city; subtract 10 per cent if it is conducted in the South or Midwest and another 10 per cent if it takes place in a rural area. The variations were found to correlate highly with the average per capita income of the community.

As these examples illustrate, simulation opens another line of compromise where the experimenter can pacify the law's basic concerns. However, simulation is a more serious deviation from the ideal experiment than the "indirect experiment" is, because its impact cannot be readily measured but must instead be assessed intuitively and through circumstantial evidence.

In the case of the judge-jury study, the element of simulation was relatively minor, the presiding judge's "verdict" being merely the one "he would have rendered had he tried the case without a jury." For a number of reasons we believe this simulation was highly realistic. Presumably the judge, in general, would find it more comfortable to report agreement with the jury, so if he reports disagreement we can accept his statement. In addition, the experiment required that his disagreement be corroborated in a variety of ways. If he would have convicted, he is asked to state the sentence he would have imposed; he also gives reasons for his disagreement. Most importantly, it was felt that the judge in the

course of a jury trial could not help arriving at his own conclusion as to the guilt or innocence, so that our question did not ask the judge to *make* a decision but only to report a judgment he had made anyway.¹⁴

In the verdict-variation experiment, the simulation, although total, was also almost perfect. The submitted *précis* contained exactly the type of information a claim adjuster is normally called to act upon, and the judgment he was asked to make—what a jury in his area would be likely to do with such a case—was the very judgment that the claim adjuster is paid to make accurately.

The simulation in the insanity experiment was more complex. In a taped trial that had been condensed to about one hour, the added instruction by the judge forms a relatively larger part of the total trial than it would in a real one. Moreover, to have only the audio tape probably reduces the impact of the trial itself more than it reduces the impact of the judge's instructions. Thus, it is quite possible that the simulation exaggerated the relative importance of the instructions, thereby producing larger experimental differences than a real trial would have produced. On the other hand, the jurors were real, they were sitting in a real court house, and the recordings of their deliberations leave no doubt about their seriousness. Some of the juries took many hours to return a verdict, there were fights and agonizing as with real juries, and some cases even ended in hung juries.

The fact remains that the impairment created by simulation is difficult to assess. At a certain point, therefore, one might prefer to have less simulation even if it means giving up some of the rigor of the experimental design. Here too, it is useful to recall the frequent reliance of the biological sciences on simulation, in the use of animals in experiments designed to test human reactions.¹⁵

THE NATURAL EXPERIMENT

The third escape route from legal constraints is to make use of what may be termed the "natural experiment." Here, there is no need for the experimenter to arrange for random assignment because random assignment has been part of the judicial process itself. A good example is provided by studies comparing judging or sentencing practices of different judges. Whenever a court assigns its cases randomly among its judges, the analyst may attribute with confidence differences in sentences to differences between judges, since the groups of cases that come before one or the other judge are as comparable as an experimenter could make them.¹⁶ Our judge-jury study had elements of a natural experiment in that the rules of criminal procedure require jury and judge to be exposed to

¹⁴Cf. Harry Kalven, Jr. & Hans Zeisel, *supra* note 10, at pp. 51–52.

¹⁵See, e.g., Primary Prevention of the Arteriosclerotic Diseases, 42 *Circulation* A-55 (Dec. 1970).

¹⁶The pioneering study was Frederick Joseph Gaudet, "Individual Differences in the Sentencing Tendencies of Judges," 32 *Archives of Psychology*, no. 230 (1938). Our split-trial study also relied on such random assignment.

the same "experimental treatment," which is to say that both see exactly the same trial.

Randomization will not always be explicit. In some situations, natural events may be relied upon to provide random exposure. An example is the interesting curve describing the relationship of driving accidents of women to their menstrual cycle:

Figure 1

The graph indicates sharply increased accidents on the days before and during the menses. This should be a natural experiment, unless women on the crucial days drive more frequently or under more hazardous conditions, a highly implausible assumption.¹⁷

Because the natural experiment is far less costly than the artificially controlled experiment, the analyst is often tempted to bestow on natural events, without further proof, the character of a controlled experiment. For instance, when it became important to learn how criminal jury verdicts would be affected if the law were to allow majority verdicts instead of insisting on unanimity, it was natural to compare jury verdicts from Oregon (the only state that allowed majority verdicts) with verdicts from other states. The frequency of hung juries in Oregon turned out to be three per cent as against five per cent in other states.¹⁸ Acceptance of the comparison implied that the cases and the jurors in Oregon and in other

¹⁷Actually the bi-modal curve is the composite of two separate uni-modal curves. For reasons not quite known, nulliparous women have their accident peak before, parous women during, their menses. Cf. Katharina Dalton, *The Premenstrual Syndrome* 89 (1964).

¹⁸Harry Kalven, Jr. & Hans Zeisel, "The American Jury: Notes for an English Controversy," 57 *The Round Table*, no. 226, at 158 (April 1967); Hans Zeisel, "... And Then There Were None: The Diminution of the Federal Jury," 38 *U. Chi. L. Rev.* 710 (1971).

states are as comparable *as if* they had been randomly assigned. Or, to put the implication differently, had one conducted an experiment in Oregon, by assigning all jury trials at random to majority rule and unanimity rule proceedings, one would have found the same three and five per cent frequencies of hung juries. But if the cases in Oregon are different from the cases that come to trial in other states, or if Oregon jurors are for some reason more likely to agree with one another, we would attribute falsely to the difference in voting rules what in fact was the result of different types of cases or different types of jurors in the Oregon courts.

The importance of a randomized control group is well illustrated by a recent analysis that did not have one. One of the major cities has a methadone treatment program for which drug addicts may volunteer. To evaluate the effects of the program on the crime rate, the following data were produced: during the 12 months prior to admission to the program the arrest rate per 100 volunteers had been 70; during the 12 months after admission, it had declined to 40. It was tempting to conclude that the methadone treatment was responsible for the decline in the arrest rates. But since there was no control group of volunteers who by random assignment were excluded from admission, there was no way of separating the arrest reduction due to the addict's motive to improve (expressed in his volunteering for treatment) from the effects of the treatment itself.

RETROSPECTIVE ANALYSIS

Situations created by explicit random assignment (e.g., cases to judges) and situations in which random assignment, however likely in fact, is unproven are radically different. In the one situation, the analyst may assign emerging differences without hesitation to the different experimental treatments (e.g., differences between the judges). In the other situation he must start with the opposite assumption: that the cases in the compared groups *may not* have been interchangeable prior to the experimental exposure, that differences in the result may be due not to the difference in treatment but to differences that existed before the experimental treatment began. For example, sentences given to defendants convicted in jury trials are on the average higher than those given in bench trials in the same court system. But it would be incorrect to conclude from this comparison alone that judges reward defendants who waive their right to jury trial with a lower sentence. The cases that are tried before a jury are on the average more serious to begin with; hence cases tried to a judge cannot be compared to cases tried to a jury as if they had been assigned randomly to these two modes of trial.

The technique of retrospective analysis (where prior natural randomization is the rare exception) is an attempt to arrange comparisons in a way that comes as close as possible to the comparison that would have been obtained through randomization, which would have made the two groups prior to the experimental treatment interchangeable. The analysis tries to approach this goal by breaking the overall group down

into smaller segments which, one hopes, approach comparability. Instead of comparing sentences in *all* jury trials with those in *all* bench trials, one would compare such sentences separately for each particular crime category and for defendants of specified sex, age, and prior criminal record. One would compare, for instance, white, male defendants, under 25 years of age, without prior criminal record found guilty of auto theft—after a jury trial—with white, male defendants under 25, without prior record, found guilty of auto theft—after a bench trial.

The drawback of this procedure of fractioned analysis is that however far it is carried (always the size of the sample sets limits), it never attains certainty. There is no way of knowing for sure that all hidden factors that might render a causal inference spurious have been eliminated.¹⁹

Retrospective analysis encounters still another difficulty that the controlled experiment eliminates. Technically known as the identification problem, it is the difficulty of determining (in retrospect) the temporal sequence of two factors. Suppose, for instance, that one were to find that in states in which the proportion of crimes in which the criminal is apprehended is high, the crime rate tends to be low. One might be tempted to conclude that it is the higher probability of apprehension that is responsible for the lower crime rate. But it could be the other way around: the fewer crimes, the more attention the police can pay to each, hence the greater the chances of apprehension. Or, the observed relationship between the two factors may reflect a mixture of both these effects. A controlled experiment that, at a certain point in time, increased the police forces in some areas but not others would resolve the problem.

Retrospective study uses essentially the same modes of analysis as the controlled experiment. They both use cross-tabulation if the distinguishing factors are attributes (classifications that differ not in gradations but in kind: male-female; black-white; Judge A vs. Judge B; jury vs. judge, etc.). They both use regression analysis if the factors allow for gradation (age, income, apprehension rates, crime rates, etc.).

Whatever technique is used, retrospective analysis must reckon with the two difficulties discussed above—the identification problem, and the required interchangeability prior to the experimental exposure. Because it bypasses both these difficulties the controlled experiment, where feasible, is the preferred research device.

THE PROBLEM OF EXTRAPOLATION AND SOME CAUTIONARY CONCLUSIONS

The controlled experiment would seem to suffer from a peculiar limitation; by its nature, it is an operation with specific limits in time,

¹⁹Cf. also Hans Zeisel, *Say It With Figure 108* (5th ed. 1965). Although it is customary to contrast retrospective studies with controlled experiments, which are as a rule prospective, the terminology is not precise. A controlled experiment may be retrospective, in the case of a natural experiment, and a prospective study might have to resort to "retrospective" analysis if it was designed without proper controls. The relevant distinction is between designs with prior random assignment, and designs without it.

space and operating conditions. Hence, there is always the question as to how far its results can be extrapolated, that is, applied to other places, other times, and other conditions.

Would the pretrial experiment, if conducted in Illinois instead of New Jersey, have yielded the same result? Would jury reactions to the defense of insanity in a murder trial be the same as in a case involving burglary or incest? There are no general rules for answering such questions; each situation requires its own assessment. The limitations of the controlled experiment are the price to be paid for the power of its analytical resolution. The endeavor must always be to choose a clearly, not necessarily narrowly, defined experimental variable, and then test it under as many environmental conditions as possible. If, for instance, the experimental treatment consists of an ill defined conglomerate of services that can never exactly be duplicated again, one is in trouble.²⁰

The limiting circumstances of controlled experiments are not a problem peculiar to the law or the social sciences in general; they hold for the so-called natural sciences as well. All one can say is that the judgment on how far a particular finding can be extrapolated will be the more correct the more developed the particular area of scientific endeavor is. That is why extrapolations in physics or chemistry can be made with greater confidence than, for instance, in biology, and in biology with greater confidence than in the social sciences.

Reading this report on a variety of legal experiments may create in some readers' minds the expectation that the law is rapidly approaching a stage where a great many of its premises will be tested through controlled experimentation. The law is very far from such a stance, and it would be a mistake to repeat such a development.

There are many legal and practical constraints that set limits to experimentation. Both the costs of experimentation and the time it requires will often prove insurmountable obstacles: legislators and courts are usually in a hurry and short of money. There is also the ever-present danger that even a generously designed experiment may run into unforeseen barriers and thereby end inconclusively or fail altogether. There is also the occasional danger that an effect attributed to the experimental treatment was in fact caused by the subjects' awareness of being part of an experiment, the so-called Hawthorne effect named after a famous first example.²¹ Even if the experiment succeeds, it seldom answers the

²⁰The British Home Office Research Unit has recently published an interesting report on a controlled experiment that seems to have failed, primarily because the experimental variable had been poorly defined. The report is in many ways valuable and instructive. But by allowing it to appear under a broadly suggestive title, the authors have come close to committing the very error their paper warns against—to generalize from limited experience. It is precisely because they have located with accuracy the causes of their limited failure that there was less need than before to choose a misleadingly discouraging title. R. V. G. Clarke and D. B. Cornish, *The Controlled Trial in Institutional Research—Paradigm or Pitfall for Penal Evaluators?*, A Home Office Research Unit Report (no. 15), 1972.

²¹A study conducted in the Hawthorne, Chicago works of the Western Electric Company, F. J. Roethlisberger & W. F. Dickson, *Management and the Worker* (1939).

"ultimate question." It provides as a rule only part of the answer, and there is always the question of just how valuable that part is.

Despite these obstacles, a number of controlled experiments are now under way, two of them designed to explore the effects of potentially far-reaching reforms. The first was designed to discover whether a negative income tax would reduce the recipient's incentive to work.²² The other experiment, conducted by the Vera Institute of Justice under a grant from the National Institute for Mental Health, explores the potential effects of a new idea for the rehabilitation of ex-drug addicts and ex-convicts. Since these men and women have a high propensity to fail in normal job situations, special working situations were arranged in which these people work among themselves, under appropriate counseling supervision. They are being paid a normal wage, although they are not quite earning it, the difference being made up from public funds. The expectation is that this *supported work* might be the most productive way—in both economic and human terms—of spending welfare money. Three hundred men will be employed in this fashion, but 600 will be screened and one-half of them will be randomly assigned to a control group. The project keeps track of all 600 men in order to learn what, if any, difference such supported work makes in terms of return to drugs or crime, economic productivity, and personal relations.

A third experiment, conducted by the University of Chicago Law School under a grant from the National Science Foundation, is designed to test the effects of peremptory challenges of jurors on the verdicts of juries. In the Federal District Court for Northern Illinois, trials are now being conducted before the juries, two of them mock juries seated in the spectator section; one consists of the jurors who have been excused by both sides, the other, nicknamed the "English Jury," consists of a jury for which no challenges have been allowed.

Thus, experimentation in the law covers a considerable variety of problems in many areas, both at its core and at the fringe. Harry Kalven has attempted to describe the area in which empirical studies should prove most useful:²³

Robert Merton some years ago observed that in their current state the social sciences could aspire only to theories in the middle range. We can adapt his remark to the application of social science to legal problems: it can only aspire to facts in the middle range. Some premises are too deeply held for factual footnoting, and some facts are too well and accessibly known for professional inquiry.

²²This experiment, incidentally, in its New Jersey phase, was to an unknown degree damaged when the experimental group of tax recipients in New Jersey was able to learn through a well publicized television program of the hoped-for result of the experiment. Moreover, because of a new welfare program introduced in New Jersey nine months after the experiment began, an alternative source of welfare became available to the experimental subjects.

²³Harry Kalven, Jr., "The Quest for the Middle Range: Empirical Inquiry and Legal Policy," in *Law in a Changing America* (Geoffrey C. Hazard, Jr., ed. 1968).

What remains then as the critical area is the middle range where the premises are not that unshakeable and where the facts are not that accessible.

There is one particular place in the law where the controlled experiment should play an increasingly important role. Occasionally, when a legal innovation is instituted, there is some lingering doubt as to whether it will work. Often such a tentative innovation is even labeled "experimental" in the sense of a try-out. For instance, before Great Britain abolished capital punishment permanently, it suspended it for a trial period of several years to see what would happen. But "to see what would happen" is very different from building an experimental design into the innovation in order to enable a precise evaluation of the enacted rule.²⁴

Whenever the expected consequences of a rule are at issue, the lawmaker should at least consider the possibilities of a controlled experiment. Even if, for any number of good reasons, he should decide not to undertake the experiment, the mere process of thinking through such a possible experimental design will have been useful. The growing vogue of insistence on "evaluation" of innovating projects, so that the funding agency may learn what has been accomplished, will help. Eventually it will lead to building into such projects the elements of controlled experimentation.

Controlled experimentation, of course, is but one of the approaches, and in a sense the sharpest one, of exploring legal issues empirically. The borderline between the prospective experiment and retrospective analysis, as we have seen, is blurred; there is a spectrum rather than a dichotomy, especially when one considers the trade-offs created by simulations. In the end, the law's attitude toward experimentation is necessarily but part of its more general attitude toward systematically collected, scientifically analyzed data. And in this perspective the signs are encouraging.

²⁴Donald T. Campbell has been a major advocate of this type of experimentation. Cf. his "Legal Reforms as Experiments," 23 *J. Legal Ed.* 217 (1970).