level, both essay and objective examinations should be used (see Stanley & Beeman, 1956), along with indices of classroom participation, etc., where feasible. (An extension of this perspective to the question of test validity is provided by Campbell and Fiske, 1959; and Campbell, 1960.)

QUASI-EXPERIMENTAL DESIGNS[5]

There are many natural social settings in which the research person can introduce something like experimental design into his scheduling of data collection procedures (e.g., the *when* and *to whom* of measurement), even though he lacks the full control over the scheduling of experimental stimuli (the *when* and *to whom* of exposure and the ability to randomize exposures) which makes a true experiment possible. Collectively, such situations can be regarded as quasi-experimental designs. One purpose of this chapter is to encourage the utilization of such quasi-experiments and to increase awareness of the kinds of settings in which opportunities to employ them occur. But just because full experimental control *is* lacking, it becomes imperative that the researcher be thoroughly aware of which specific variables his particular design fails to control. It is for this need in evaluating quasi-experiments, more than for understanding true experiments, that the check lists of sources of invalidity in Tables 1, 2, and 3 were developed.

The average student or potential researcher reading the previous section of this chapter probably ends up with more things to worry about in designing an experiment than he had in mind to begin with. This is all to the good if it leads to the design and execution of better experiments and to more circumspection in drawing inferences from results. It is, however, an unwanted side effect if it creates a feeling of hopelessness with regard to achieving experimental control and leads to the abandonment of such efforts in favor of even more informal methods of investigation. Further, this formidable list of sources of invalidity might, with even more likelihood, reduce willingness to undertake quasi-experimental designs, designs in which from the very outset it can be seen that full experimental control is lacking. Such an effect would be the opposite of what is intended.

From the standpoint of the final interpretation of an experiment and the attempt to fit it into the developing science, every experiment is imperfect. What a check list of validity criteria can do is to make an experimenter more aware of the residual imperfections in his design so that on the relevant points he can be aware of competing interpretations of his data. He should, of course, design the very best experiment which the situation makes possible. He should deliberately seek out those artificial and natural laboratories which provide the best opportunities for control. But beyond that he should go ahead with experiment and interpretation, fully aware of the points on which the results are equivocal. While this awareness is important for experiments in which "full" control has been exercised, it is crucial for quasi-experimental designs.

In implementing this general goal, we shall in this portion of the chapter survey the strengths and weaknesses of a heterogeneous collection of quasi-experimental designs, each deemed worthy of use *where better designs are not feasible*. First will be discussed three single-group experimental designs. Following these, five general types of multiple-group experiments will be presented. A separate section will deal with correlation, ex post facto designs, panel studies, and the like.

Some Preliminary Comments on the Theory of Experimentation

This section is written primarily for the educator who wishes to take his research

[5] This section draws heavily upon D. T. Campbell, Quasi-experimental designs for use in natural social settings, in D. T. Campbell, *Experimenting, Validating, Knowing: Problems of Method in the Social Sciences*. New York: McGraw-Hill, in preparation.

out of the laboratory and into the operating situation. Yet the authors cannot help being aware that experimental psychologists may look with considerable suspicion on any effort to sanction studies having less than full experimental control. In part to justify the present activity to such monitors, the following general comments on the role of experiments in science are offered. These comments are believed to be compatible with most modern philosophies of science, and they come from a perspective on a potential general psychology of inductive processes (Campbell, 1959).

Science, like other knowledge processes, involves the proposing of theories, hypotheses, models, etc., and the acceptance or rejection of these on the basis of some external criteria. Experimentation belongs to this second phase, to the pruning, rejecting, editing phase. We may assume an ecology for our science in which the number of potential positive hypotheses very greatly exceeds the number of hypotheses that will in the long run prove to be compatible with our observations. *The task of theory-testing data collection is therefore predominantly one of rejecting inadequate hypotheses.* In executing this task, any arrangement of observations for which certain outcomes would disconfirm theory will be useful, including quasi-experimental designs of less efficiency than true experiments.

But, it may be asked, will not such imperfect designs result in spurious confirmation of inadequate theory, mislead our subsequent efforts, and waste our journal space with the dozens of studies which it seems to take to eradicate one conspicuously published false positive? This is a serious risk, but a risk which we must take. It is a risk shared in kind, if not in the same degree, by "true" experiments of Designs 4, 5, and 6. In a very fundamental sense, experimental results never "confirm" or "prove" a theory—rather, the successful theory is tested and escapes being disconfirmed. The word "prove," by being frequently employed to designate deductive validity, has acquired in our generation a connotation inappropriate both to its older uses and to its application to inductive procedures such as experimentation. The results of an experiment "probe" but do not "prove" a theory. An adequate hypothesis is one that has repeatedly survived such probing—but it may always be displaced by a new probe.

It is by now generally understood that the "null hypothesis" often employed for convenience in stating the hypothesis of an experiment can never be "accepted" by the data obtained; it can only be "rejected," or "fail to be rejected." Similarly with hypotheses more generally—they are technically never "confirmed": where we for convenience use that term we imply rather that the hypothesis was exposed to disconfirmation and was not disconfirmed. This point of view is compatible with all Humean philosophies of science which emphasize the impossibility of deductive proof for inductive laws. Recently Hanson (1958) and Popper (1959) have been particularly explicit upon this point. Many bodies of data collected in research on teaching have little or no probing value, and many hypothesis-sets are so double-jointed that they cannot be disconfirmed by available probes. We have no desire to increase the acceptability of such pseudo research. The research designs discussed below are believed to be sufficiently probing, however, to be well worth employing *where more efficient probes are unavailable.*

The notion that experiments never "confirm" theory, while correct, so goes against our attitudes and experiences as scientists as to be almost intolerable. Particularly does this emphasis seem unsatisfactory vis-à-vis the elegant and striking confirmations encountered in physics and chemistry, where the experimental data may fit in minute detail over numerous points of measurement a complex curve predicted by the theory. And the perspective becomes phenomenologically unacceptable to most of us when extended to the inductive achievements of vision. For example, it is hard to realize that

the tables and chairs which we "see" before us are not "confirmed" or "proven" by the visual evidence, but are "merely" hypotheses about external objects not as yet disconfirmed by the multiple probes of the visual system. There is a grain of truth in these reluctances.

Varying degrees of "confirmation" are conferred upon a theory through the number of *plausible rival hypotheses* available to account for the data. The fewer such plausible rival hypotheses remaining, the greater the degree of "confirmation." Presumably, at any stage of accumulation of evidence, even for the most advanced science, there are numerous possible theories compatible with the data, particularly if all theories involving complex contingencies be allowed. Yet for "well-established" theories, and theories thoroughly probed by complex experiments, few if any rivals may be practically available or seriously proposed. This fewness is the epistemological counterpart of the positive affirmation of theory which elegant experiments seem to offer. A comparable fewness of rival hypotheses occurs in the phenomenally positive knowledge which vision seems to offer in contrast, for example, to the relative equivocality of blind tactile exploration.

In this perspective, the list of sources of invalidity which experimental designs control can be seen as a list of frequently plausible hypotheses which are rival to the hypothesis that the experimental variable has had an effect. Where an experimental design "controls" for one of these factors, it merely renders this rival hypothesis implausible, even though through possible complex coincidences it might still operate to produce the experimental outcome. The "plausible rival hypotheses" that have necessitated the routine use of special control groups have the status of well-established empirical laws: practice effects for adding a control group to Design 2, suggestibility for the placebo control group, surgical shock for the sham-operation control. Rival hypotheses are plausible insofar as we are willing to attribute to them the status of empirical laws. Where controls are lacking in a quasi-experiment, one must, in interpreting the results, consider in detail the likelihood of uncontrolled factors accounting for the results. The more implausible this becomes, the more "valid" the experiment.

As was pointed out in the discussion of the Solomon Four-Group Design 5, the more numerous and independent the ways in which the experimental effect is demonstrated, the less numerous and less plausible any singular rival invalidating hypothesis becomes. The appeal is to parsimony. The "validity" of the experiment becomes one of the relative credibility of rival theories: the theory that X had an effect versus the theories of causation involving the uncontrolled factors. If several sets of differences can all be explained by the single hypothesis that X has an effect, while several separate uncontrolled-variable effects must be hypothesized, a different one for each observed difference, then the effect of X becomes the most tenable. This mode of inference is frequently appealed to when scientists summarize a literature lacking in perfectly controlled experiments. Thus Watson (1959, p. 296) found the evidence for the deleterious effects of maternal deprivation confirmatory because it is supported by a wide variety of evidence-types, the specific inadequacies of which vary from study to study. Thus Glickman (1961), in spite of the presence of plausible rival hypotheses in each available study, found the evidence for a consolidation process impressive just because the plausible rival hypothesis is different from study to study. This inferential feature, commonly used in combining inferences from several studies, is deliberately introduced *within* certain quasi-experimental designs, especially in "patched-up" designs such as Design 15.

The appeal to parsimony is not deductively justifiable but is rather a general assumption about the nature of the world, underlying almost all use of theory in science, even though frequently erroneous in specific applications. Related to it is another

plausibility argument which we will invoke perhaps most specifically with regard to the very widely used Design 10 (a good *quasi*-experimental design, often mistaken for the true Design 4). This is the assumption that, in cases of ignorance, a main effect of one variable is to be judged more likely than the interaction of two other variables; or, more generally, that main effects are more likely than interactions. In the extreme form, we can note that if every highest-order interaction is significant, if every effect is specific to certain values on all other potential treatment dimensions, then a science is not possible. If we are ever able to generalize, it is because the great bulk of potential determining factors can be disregarded. Underwood (1957b, p. 6) has referred to this as the assumption of finite causation. Elsewhere Underwood (1954) has tallied the frequency of main effects and interactions from the *Journal of Experimental Psychology,* confirming the relative rarity of significant interactions (although editorial selection favoring neat outcomes makes his finding suspect).

In what follows, we will first deal with single-group experiments. Since 1920 at least, the dominant experimental design in psychology and education has been a control group design, such as Design 4, Design 6, or perhaps most frequently Design 10, to be discussed later. In the social sciences and in thinking about field situations, the control group designs so dominate as to seem to many persons synonymous with experimentation. As a result, many research workers may give up attempting anything like experimentation in settings where control groups are not available and thus end up with more imprecision than is necessary. There are, in fact, several quasi-experimental designs applicable to single groups which might be used to advantage, with an experimental logic and interpretation, in many situations in which a control group design is impossible. Cooperation and experimental access often come in natural administrative units: a teacher has her own classroom available; a high school principal may be willing to introduce periodic morale surveys, etc. In such situations the differential treatment of segments within the administrative unit (required for the control group experiment) may be administratively impossible or, even if possible, experimentally undesirable owing to the reactive effects of arrangements. For these settings, single-group experiments might well be considered.

7. The Time-Series Experiment

The essence of the time-series design is the presence of a periodic measurement process on some group or individual and the introduction of an experimental change into this time series of measurements, the results of which are indicated by a discontinuity in the measurements recorded in the time series. It can be diagramed thus:

$$O_1\ O_2\ O_3\ O_4 X O_5\ O_6\ O_7\ O_8$$

This experimental design typified much of the classical nineteenth-century experimentation in the physical sciences and in biology. For example, if a bar of iron which has remained unchanged in weight for many months is dipped in a nitric acid bath and then removed, the inference tying together the nitric acid bath and the loss of weight by the iron bar would follow some such experimental logic. There may well have been "control groups" of iron bars remaining on the shelf that lost no weight, but the measurement and reporting of these weights would typically not be thought necessary or relevant. Thus it seems likely that this experimental design is frequently regarded as valid in the more successful sciences even though it rarely has accepted status in the enumerations of available experimental designs in the social sciences. (See, however, Maxwell, 1958; Underwood, 1957b, p. 133.) There are good reasons for this differential status and a careful consideration of them will provide a better understanding of the

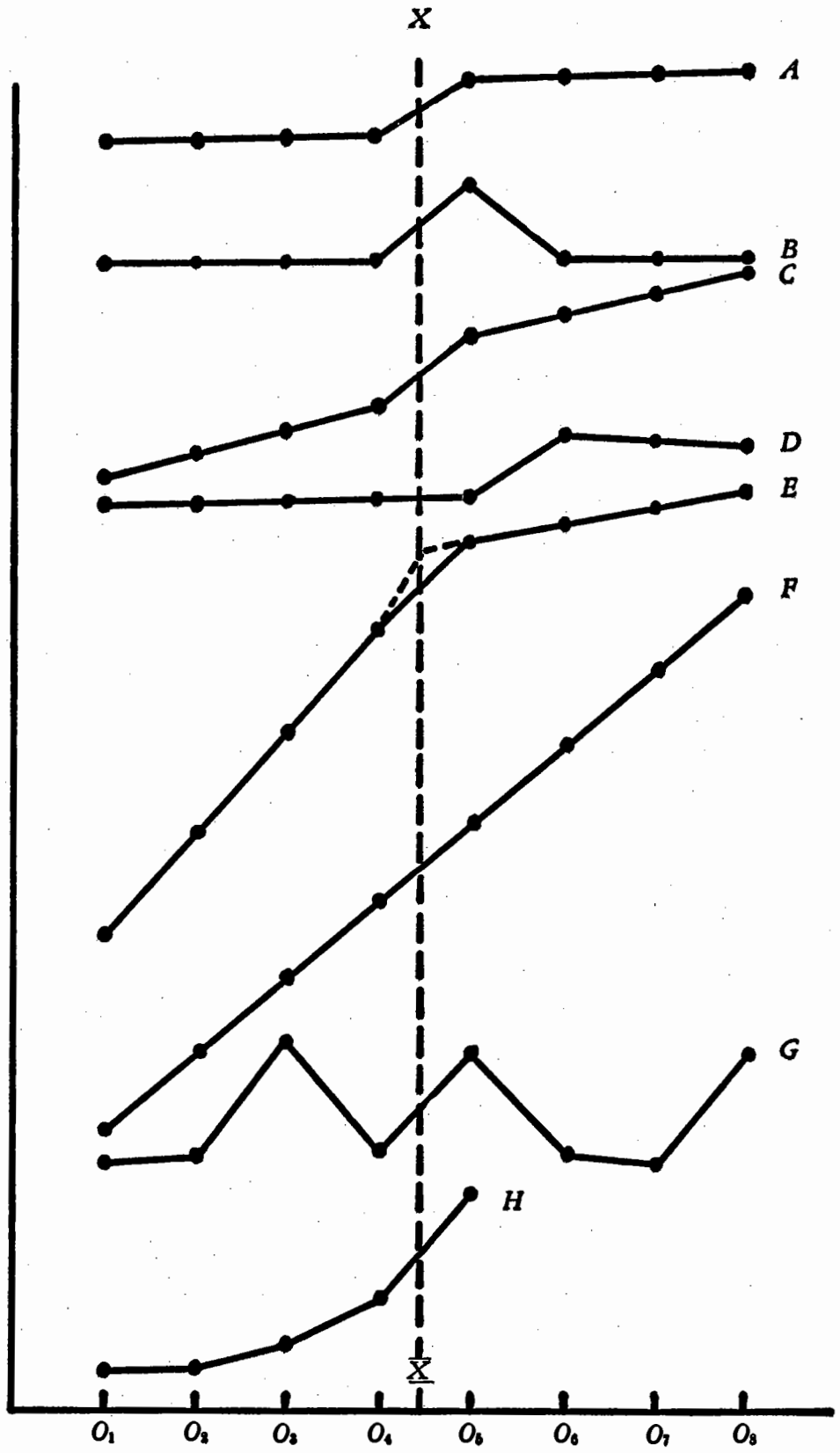

Fig. 3. Some Possible Outcome Patterns from the Introduction of an Experimental Variable at Point X into a Time Series of Measurements, O_1—O_8. Except for D, the O_4—O_5 gain is the same for all time series, while the legitimacy of inferring an effect varies widely, being strongest in A and B, and totally unjustified in F, G, and H.

conditions under which the design might meaningfully be employed by social scientists when more thorough experimental control is impossible. The design is typical of the classic experiments of the British Industrial Fatigue Research Board upon factors affecting factory outputs (e.g., Farmer, Brooks, & Chambers, 1923).

Figure 3 indicates some possible outcome patterns for time series into which an experimental alteration had been introduced as indicated by the vertical line *X*. For purposes of discussion let us assume that one will be tempted to infer that *X* had some effect in time series with outcomes such as *A* and *B* and possibly *C, D,* and *E,* but that one would not be tempted to infer an effect in time series such as *F, G,* and *H,* even were the jump in values from O_4 to O_5 as great and as statistically stable as were the O_4 to O_5 differences in *A* and *B,* for example. While discussion of the problem of statistical tests will be postponed for a few paragraphs, it is assumed that the problem of internal validity boils down to the question of plausible competing hypotheses that offer likely alternate explanations of the shift in the time series other than the effect of *X*. A tentative check-off of the controls provided by this experiment under these optimal conditions of outcome is provided in Table 2. The strengths of the time-series design are most apparent in contrast with Design 2, to which it has a superficial similarity in lacking a control group and in using before-and-after measures.

Scanning the list of problems of internal validity in Table 2, we see that failure to control history is the most definite weakness of Design 7. That is, the rival hypothesis exists that not *X* but some more or less simultaneous event produced the shift. It is upon the plausibility of ruling out such extraneous stimuli that credence in the interpretation of this experiment in any given instance must rest. Consider an experiment involving repeated measurements and the effect of a documentary film on students' optimism about the likelihood of war. Here the failure to provide a clear-cut control on *history* would seem very serious indeed since it is obvious that the students are exposed daily to many potentially relevant sources of stimulation beyond those under the experimenter's control in the classroom. Of course even here, were the experiment to be accompanied by a careful log of nonexperimental stimuli of possible relevance, plausible interpretation making the experiment worth doing might be possible. As has been noted above, the variable *history* is the counterpart of what in the physical and biological science laboratory has been called *experimental isolation*. The plausibility of *history* as an explanation for shifts such as those found in time-series *A* and *B* of Fig. 3 depends to a considerable extent upon the degree of experimental isolation which the experimenter can claim. Pavlov's conditioned-reflex studies with dogs, essentially "one-group" or "one-animal" experiments, would have been much less plausible as support of Pavlov's theories had they been conducted on a busy street corner rather than in a soundproof laboratory. What constitutes experimental isolation varies with the problem under study and the type of measuring device used. More precautions are needed to establish experimental isolation for a cloud chamber or scintillation counter study of subatomic particles than for the hypothetical experiment on the weight of bars of iron exposed to baths of nitric acid. In many situations in which Design 7 might be used, the experimenter could plausibly claim experimental isolation in the sense that he was aware of the possible rival events that might cause such a change and could plausibly discount the likelihood that they explained the effect.

Among other extraneous variables which might for convenience be put into *history* are the effects of weather and the effects of season. Experiments of this type are apt to extend over time periods that involve seasonal changes and, as in the studies of worker output, the seasonal fluctuations in illumination, weather, etc., may be confounded with the introduction of experimental change.

TABLE 2

SOURCES OF INVALIDITY FOR QUASI-EXPERIMENTAL DESIGNS 7 THROUGH 12

	Sources of Invalidity											
	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation, etc.	Interaction of Testing and X	Interaction of Selection and X	Reactive Arrangements	Multiple-X Interference
Quasi-Experimental Designs:												
7. Time Series O O O OXO O O O	−	+	+	?	+	+	+	+	−	?	?	
8. Equivalent Time Samples Design X_1O X_0O X_1O X_0O, etc.	+	+	+	+	+	+	+	+	−	?	−	−
9. Equivalent Materials Samples Design M_aX_1O M_bX_0O M_cX_1O M_dX_0O, etc.	+	+	+	+	+	+	+	+	−	?	?	−
10. Nonequivalent Control Group Design O X O --- O O	+	+	+	+	?	+	+	−	−	?	?	
11. Counterbalanced Designs X_1O X_2O X_3O X_4O --- X_2O X_4O X_1O X_3O --- X_3O X_1O X_4O X_2O --- X_4O X_3O X_2O X_1O	+	+	+	+	+	+	+	?	?	?	?	−
12. Separate-Sample Pretest-Posttest Design R O (X) R X O	−	−	+	?	+	+	−	−	+	+	+	
12*a*. R O (X) R X O --- R O (X) R X O	+	−	+	?	+	+	−	+	+	+	+	
12*b*. R O_1 (X) R O_2 (X) R X O_3	−	+	+	?	+	+	−	?	+	+	+	
12*c*. R O_1 X O_2 R X O_3	−	−	+	?	+	+	+	−	+	+	+	

Perhaps best also included under *history*, although in some sense akin to *maturation*, would be periodical shifts in the time series related to institutional customs of the group such as the weekly work-cycles, pay-period cycles, examination periods, vacations, and student festivals. The observational series should be arranged so as to hold known cycles constant, or else be long enough to include several such cycles in their entirety.

To continue with the factors to be controlled: *maturation* seems ruled out on the grounds that if the outcome is like those in illustrations *A* and *B* of Fig. 3, maturation does not usually provide plausible rival hypotheses to explain a shift occurring between O_4 and O_5 which did not occur in the previous time periods under observation. (However, maturation may not always be of a smooth, regular nature. Note how the abrupt occurrence of menarche in first-year junior high school girls might in a Design 7 appear as an effect of the shift of schools upon physiology records, did we not know better.) Similarly, testing seems, in general, an implausible rival hypothesis for a jump between O_4 and O_5. Had one only the observations at O_4 and O_5, as in Design 2, this means of rendering maturation and test-retest effects implausible would be lacking. Herein lies the great advantage of this design over Design 2.

In a similar way, many hypotheses invoking changes in *instrumentation* would lack a specific rationale for expecting the instrument error to occur on this particular occasion, as opposed to earlier ones. However, the question mark in Table 2 calls attention to situations in which a change in the calibration of the measurement device could be misinterpreted as the effect of *X*. If the measurement procedure involves the judgments of human observers who are aware of the experimental plan, pseudo confirmation of the hypothesis can occur as a result of the observer's expectations. Thus, the experimental change of putting into office a new principal may produce a change in the recording of discipline infractions rather than in the infraction rate itself. Design 7 may frequently be employed to measure effects of a major change in administrative policy. Bearing this in mind, one would be wise to avoid shifting measuring instruments at the same time he shifts policy. In most instances, to preserve the interpretability of a time series, it would be better to continue to use a somewhat antiquated device rather than to shift to a new instrument.

Regression effects are usually a negatively accelerated function of elapsed time and are therefore implausible as explanations of an effect at O_5 greater than the effects at O_2, O_3, and O_4. *Selection* as a source of main effects is ruled out in both this design and in Design 2, if the same specific persons are involved at all *O*s. If data from a group is basically collected in terms of individual group members, then mortality may be ruled out in this experiment as in Design 2. However, if the observations consist of collective products, then a record of the occurrence of absenteeism, quitting, and replacement should be made to insure that coincidences of personnel change do not provide plausible rival hypotheses.

Regarding external validity, it is clear that the experimental effect might well be specific to those populations subject to repeated testing. This is hardly likely to be a limitation in research on teaching in schools, unless the experiment is conducted with artificial *O*s not common to the usual school setting. Furthermore, this design is particularly appropriate to those institutional settings in which records are regularly kept and thus constitute a natural part of the environment. Annual achievement tests in the public schools, illness records, etc., usually are nonreactive in the sense that they are typical of the universe to which one wants to generalize. The *selection-X* interaction refers to the limitation of the effects of the experimental variable to that specific sample and to the possibility that this reaction would not be typical of some more general universe of interest for which the naturally aggregated exposure-group was a biased sample. For example, the data requirements may limit one to those students who have had perfect attendance records over long periods, an obviously select subset. Further, if novel *O*s have been used, this repetitive occurrence may have provoked absenteeism.

If such time series are to be interpreted as experiments, it seems essential that the experimenter must specify in advance the expected time relationship between the introduction of the experimental variable and the

manifestation of an effect. If this had been done, the pattern indicated in time-series *D* of Fig. 3 could be almost as definitive as that in *A*. Exploratory surveys opportunistically deciding upon interpretations of delayed effect would require cross-validation before being interpretable. As the time interval between *X* and effect increases, the plausibility of effects from extraneous historical events also increases.

It also seems imperative that the *X* be specified before examining the outcome of the time series. The post hoc examination of a time series to infer what *X* preceded the most dramatic shift must be ruled out on the grounds that the opportunistic capitalization on chance which it allows makes any approach to testing the significance ot effects difficult if not impossible.

The prevalence of this design in the more successful sciences should give us some respect for it, yet we should remember that the facts of "experimental isolation" and "constant conditions" make it more interpretable for them than for us. It should also be remembered that, in their use of it, a single experiment is never conclusive. While a control group may never be used, Design 7 is repeated in many different places by various researchers before a principle is established. This, too, should be our use of it. *Where nothing better controlled is possible,* we will use it. We will organize our institutional bookkeeping to provide as many time series as possible for such evaluations and will try to examine in more detail than we have previously the effects of administrative changes and other abrupt and arbitrary events as *X*s. But these will not be regarded as definitive until frequently replicated in various settings.

Tests of Significance for the Times-Series Design

If the more advanced sciences use tests of significance less than do psychology and education, it is undoubtedly because the magnitude and the clarity of the effects with which they deal are such as to render tests of significance unnecessary. If our conventional tests of significance were applied, high degrees of significance would be found. It seems typical of the ecology of the social sciences, however, that they must work the low-grade ore in which tests of significance are necessary. It also seems likely that wherever common sense or intuitive considerations point to a clear-cut effect, some test of significance that formalizes considerations underlying the intuitive judgment is usually possible. Thus tests of significance of the effects of *X* that would distinguish between the several outcomes illustrated in Fig. 3, judging *A* and *B* to be significant and *F* and *G* not significant, may be available. We shall discuss a few possible approaches.

First, however, let us reject certain conceivable approaches as inadequate. If the data in Fig. 3 represent group means, then a simple significance test of the difference between the observations of O_4 and O_5 is insufficient. Even if in series *F* and *G*, these provided *t* ratios that were highly significant, we would not find the data evidence of effect of *X* because of the presence of other similar significant shifts occurring on occasions for which we had no matching experimental explanation. Where one is dealing with the kind of data provided in national opinion surveys, it is common to encounter highly significant shifts from one survey to the next which are random noise from the point of view of the interpreting scientist, inasmuch as they represent a part of the variation in the phenomena for which he has no explanation. The effect of a clear-cut event or experimental variable must rise above this ordinary level of shift in order to be interpretable. Similarly, a test of significance involving the pooled data for all of the pre-*X* and post-*X* observations is inadequate, inasmuch as it would not distinguish between instances of type *F* and instances of type *A*.

There is a troublesome nonindependence involved which must be considered in developing a test of significance. Were such nonindependence homogeneously distributed across all observations, it would be no threat

to internal validity, although a limitation to external validity. What is troublesome is that in almost every time series it will be found that adjacent observations are more similar than nonadjacent ones (i.e., that the autocorrelation of lag 1 is greater than that for lag 2, etc.). Thus, an extraneous influence or random disturbance affecting an observation point at, say, O_5 or O_6, will also disturb O_7 and O_8, so that it is illegitimate to treat them as several independent departures from the extrapolation of the O_1—O_4 trend.

The test of significance employed will, in part, depend upon the hypothesized nature of the effect of X. If a model such as line B is involved, then a test of the departure of O_5 from the extrapolation of O_1—O_4 could be used. Mood (1950, pp. 297–298) provides such a test. Such a test could be used for all instances, but it would seem to be unnecessarily weak where a continuous improvement, or increased rate of gain, were hypothesized. For such cases, a test making use of all points would seem more appropriate. There are two components which might enter into such tests of significance. These are intercept and slope. By intercept we refer to the jump in the time series at X (or at some specified lag after X). Thus lines A and C show an intercept shift with no change in slope. Line E shows a change in slope but no change in intercept in that the pre-X extrapolation to X and the post-X extrapolation to X coincide. Often both intercept, and slope would be changed by an effective X. A pure test of intercept might be achieved in a manner analogous to working the Mood test from both directions at once. In this case, two extrapolated points would be involved, with both pre-X and post-X observations being extrapolated to a point X halfway between O_4 and O_5.

Statistical tests would probably involve, in all but the most extended time series, linear fits to the data, both for convenience and because more exact fitting would exhaust the degrees of freedom, leaving no opportunity to test the hypothesis of change. Yet frequently the assumption of linearity may not be appropriate. The plausibility of inferring an effect of X is greatest adjacent to X. The more gradual or delayed the supposed effect, the more serious the confound with history, because the possible extraneous causes become more numerous.

8. The Equivalent Time-Samples Design

The most usual form of experimental design employs an equivalent sample of persons to provide a baseline against which to compare the effects of the experimental variable. In contrast, a recurrent form of one-group experimentation employs two equivalent samples of occasions, in one of which the experimental variable is present and in another of which it is absent. This design can be diagramed as follows (although a random rather than a regular alternation is intended):

$$X_1O \quad X_0O \quad X_1O \quad X_0O$$

This design can be seen as a form of the time-series experiment with the repeated introduction of the experimental variable. The experiment is most obviously useful where the effect of the experimental variable is anticipated to be of transient or reversible character. While the logic of the experiment may be seen as an extension of the time-series experiment, the mode of statistical analysis is more typically similar to that of the two-group experiment in which the significance of the difference between the means of two sets of measures is employed. Usually the measurements are quite specifically paired with the presentations of the experimental variable, frequently being concomitant, as in studies of learning, work production, conditioning, physiological reaction, etc. Perhaps the most typical early use of this experimental design, as in the studies of efficiency of students' work under various conditions by Allport (1920) and Sorokin (1930), involved the comparison of two experimental variables with each other, i.e., X_1 versus X_2 rather than

one with a control. For most purposes, the simple alternation of conditions and the employment of a consistent time spacing are undesirable, particularly when they may introduce confounding with a daily, weekly, or monthly cycle, or when through the predictable periodicity an unwanted conditioning to the temporal interval may accentuate the difference between one presentation and another. Thus Sorokin made sure that each experimental treatment occurred equally often in the afternoon and the forenoon.

Most experiments employing this design have used relatively few repetitions of each experimental condition, but the type of extension of sampling theory represented by Brunswik (1956) calls attention to the need for large, representative, and equivalent random samplings of time periods. Kerr (1945) has perhaps most nearly approximated this ideal in his experiments on the effects of music upon industrial production. Each of his several experiments involved a single experimental group with a randomized, equivalent sample of days over periods of months. Thus, in one experiment he was able to compare 56 music days with 51 days without music, and in another he was able to compare three different types of music, each represented by equivalent samples of 14 days.

As employed by Kerr, for example, Design 8 seems altogether internally valid. *History,* the major weakness of the time-series experiment, is controlled by presenting X on numerous separate occasions, rendering extremely unlikely any rival explanation based on the coincidence of extraneous events. The other sources of invalidity are controlled by the same logic detailed for Design 7. With regard to external validity, generalization is obviously possible only to frequently tested populations. The reactive effect of arrangements, the awareness of experimentation, represents a particular vulnerability of this experiment. Where separate groups are getting the separate Xs, it is possible (particularly under Design 6) to have them totally unaware of the presence of an experiment or of the treatments being compared. This is not so when a single group is involved, and when it is repeatedly being exposed to one condition or another, e.g., to one basis for computing payment versus another in Sorokin's experiment; to one condition of work versus another in Allport's; to one kind of ventilation versus another in Wyatt, Fraser, and Stock's (1926) studies; and to one kind of music versus another in Kerr's (although Kerr took elaborate precautions to make varied programing become a natural part of the working environment). As to the interaction of *selection* and X: there is as usual the limitation of the generalization of the demonstrated effects of X to the particular type of population involved.

This experimental design carries a hazard to external validity which will be found in all of those experiments in this paper in which multiple levels of X are presented to the *same* set of persons. This effect has been labeled "multiple-X interference." The effect of X_1, in the simplest situation in which it is being compared with X_0, can be generalized only to conditions of repetitious and spaced presentations of X_1. No sound basis is provided for generalization to possible situations in which X_1 is continually present, or to the condition in which it is introduced once and once only. In addition, the X_0 condition or the absence of X is not typical of periods without X in general, but is only representative of absences of X interspersed among presences. If X_1 has some extended effect carrying over into the non-X periods, as usually would seem likely, the experimental design may underestimate the effect of X_1 as compared with a Design 6 study, for example. On the other hand, the very fact of frequent shifts may increase the stimulus value of an X over what it would be under a continuous, homogeneous presentation. Hawaiian music in Kerr's study might affect work quite differently when interspersed for a day among days of other music than it would as a continuous diet. Ebbinghaus' (1885) experimental designs may be regarded as essentially of this type and, as Underwood (1957a) has pointed out, the

laws which he found are limited in their generalizability to a population of persons who have learned dozens of other highly similar lists. Many of his findings do not in fact hold for persons learning a single list of nonsense syllables. Thus, while the design is internally valid, its external validity may be seriously limited for some types of content. (See also Kempthorne, 1952, Ch. 29.)

Note, however, that many aspects of teaching on which one would like to experiment may very well have effects limited for all practical purposes to the period of actual presence of X. For such purposes, this design might be quite valuable. Suppose a teacher questions the value of oral recitation versus individual silent study. By varying these two procedures over a series of lesson units, one could arrange an interpretable experiment. The effect of the presence of a parent-observer in the classroom upon students' volunteer discussion could be studied in this way. Awareness of such designs can place an experimental testing of alternatives within the grasp of an individual teacher. This could pilot-test procedures which if promising might be examined by larger, more coordinated experiments.

This approach could be applied to a sampling of occasions for a single individual. While tests of significance are not typically applied, this is a recurrent design in physiological research, in which a stimulus is repeatedly applied to one animal, with care taken to avoid any periodicity in the stimulation, the latter feature corresponding to the randomization requirement for occasions demanded by the logic of the design. Latin squares rather than simple randomization may also be used (e.g., Cox, 1951; Maxwell, 1958).

Tests of Significance for Design 8

Once again, we need appropriate tests of significance for this particular type of design. Note that two dimensions of generalization are implied: generalization across occasions and generalization across persons. If we consider an instance in which only one person is employed, the test of significance will obviously be limited to generalizations about this particular person and will involve a generalization across instances, for which purpose it will be appropriate to use a t with degrees of freedom equal to the number of occasions less two. If one has individual records on a number of persons undergoing the same treatment, all a part of the same group, then data are available also for generalization across persons. In this usual situation two strategies seem common. A wrong one is to generate for each individual a single score for each experimental treatment, and then to employ tests of significance of the difference between means with correlated data. While tests of significance were not actually employed, this is the logic of Allport's and Sorokin's analyses. But where only one or two repetitions of each experimental condition are involved, sampling errors of occasions may be very large or the control of history may be very poor. Chance sampling errors of occasions could contribute what would appear under this analysis to be significant differences among treatments. This seems to be a very serious error if the effect of occasions is significant and appreciable. One could, for example, on this logic get a highly significant difference between X_1 and X_2 where each has been presented only once and where on one occasion some extraneous event had by chance produced a marked result. It seems essential therefore that at least two occasions be "nested" within each treatment and that degrees of freedom between occasions within treatments be represented. This need is probably most easily met by initially testing the difference between treatment means against a between-occasions-within-treatments error term. After the significance of the treatment effect has been established in this way, one could proceed to find for what proportion of the subjects it held, and thus obtain evidence relevant to the generalizability of the effect across persons. Repeated measurements and sampling of occasions

pose many statistical problems, some of them still unresolved (Collier, 1960; Cox, 1951; Kempthorne, 1952).

9. The Equivalent Materials Design

Closely allied to the equivalent time-samples design is Design 9, basing its argument on the equivalence of samples of materials to which the experimental variables being compared are applied. Always or almost always, equivalent time samples are also involved, but they may be so finely or intricately interspersed that there is practical temporal equivalence. In a one-group repeated-X design, equivalent materials are required whenever the nature of the experimental variables is such that the effects are enduring and the different treatments and repeats of treatments must be applied to nonidentical content. The design may be indicated in this fashion:

$$M_aX_1O \quad M_bX_0O \quad M_cX_1O \quad M_dX_0O \text{ etc.}$$

The Ms indicate specific materials, the sample M_a, M_c, etc., being, in sampling terms, equal to the sample M_b, M_d, etc. The importance of the sampling equivalence of the two sets of materials is perhaps better indicated if the design is diagramed in this fashion:

$$\text{one person or group} \begin{cases} \text{Materials Sample } A\ (O)\ X_0\ O \\ \text{Materials Sample } B\ (O)\ X_1\ O \end{cases}$$

The Os in parentheses indicate that in some designs a pretest will be used and in others not.

Jost's (1897) early experiment on massed versus distributed practice provides an excellent illustration. In his third experiment, 12 more or less randomly assembled lists of 12 nonsense syllables each were prepared. Six of the lists were assigned to distributed practice and six to massed practice. These 12 were then simultaneously learned over a seven-day period, their scheduling carefully intertwined so as to control for fatigue, etc. Seven such sets of six distributed and six massed lists were learned over a period lasting from November 6, 1895, to April 7, 1896. In the end, Jost had results on 40 different nonsense syllable lists learned under massed practice and 40 learned under distributed practice. The interpretability of the differences found on the one subject, Professor G. E. Müller, depends upon the sampling equivalence of the nonidentical lists involved. Within these limits, this experiment seems to have internal validity. The findings are of course restricted to the psychology of Professor G. E. Müller in 1895 and 1896 and to the universe of memory materials sampled. To enable one to generalize across persons in achieving a more general psychology, replication of the experiment on numerous persons is of course required.

Another illustration comes from early studies of conformity to group opinion. For example, Moore (1921) obtained a "control" estimate of retest stability of questionnaire responses from one set of items, and then compared this with the change resulting when, with another set of items, the retest was accompanied by a statement of majority opinion. Or consider a study in which students are asked to express their opinions on a number of issues presented in a long questionnaire. These questions are then divided into two groups as equivalent as possible. At a later time the questionnaires are handed back to the students and the group vote for each item indicated. These votes are falsified, to indicate majorities in opposite directions for the two samples of items. As a post-X measure, the students are asked to vote again on all items. Depending upon the adequacy of the argument of sampling equivalence of the two sets of items, the differences in shifts between the two experimental treatments would seem to provide a definitive experimental demonstration of the effects of the reporting of group opinions, even in the absence of any control group of persons.

Like Design 8, Design 9 has internal validity on all points, and in general for the same reasons. We may note, with regard to exter-

nal validity, that the effects in Design 9, like those in all experiments involving repeated measures, may be quite specific to persons repeatedly measured. In learning experiments, the measures are so much a part of the experimental setting in the typical method used today (although not necessarily in Jost's method, in which the practices involved controlled numbers of readings of the lists) that this limitation on generalization becomes irrelevant. Reactive arrangements seem to be less certainly involved in Design 9 than in Design 8 because of the heterogeneity of the materials and the greater possibility that the subjects will not be aware that they are getting different treatments at different times for different items. This low reactivity would not be found in Jost's experiment but it would be found in the conformity study. Interference among the levels of the experimental variable or interference among the materials seems likely to be a definite weakness for this experiment, as it is for Design 8.

We have a specific illustration of the kind of limitation thus introduced with regard to Jost's findings. He reported that spaced learning was more efficient than massed practice. From the conditions of his experimentation in general, we can see that he was justified in generalizing only to persons who were learning many lists, that is, persons for whom the general interference level was high. Contemporary research indicates that the superiority of spaced learning is limited to just such populations, and that for persons learning highly novel materials for the first time, no such advantage is present (Underwood & Richardson, 1958).

Statistics for Design 9

The sampling of materials is obviously relevant to the validity and the degree of proof of the experiment. As such, the N for the computation of the significance of the differences between the means of treatment groups should probably have been an N of lists in the Jost experiment (or an N of items in the conformity study) so as to represent this relevant sampling domain. This must be supplemented by a basis for generalizing across persons. Probably the best practice at the present time is to do these seriatim, establishing the generalization across the sample of lists or items first, and then computing an experimental effects score for each particular person and employing this as a basis for generalizing across persons. (Note the cautionary literature cited above for Design 8.)

10. The Nonequivalent Control Group Design

One of the most widespread experimental designs in educational research involves an experimental group and a control group both given a pretest and a posttest, but in which the control group and the experimental group do not have pre-experimental sampling equivalence. Rather, the groups constitute naturally assembled collectives such as classrooms, as similar as availability permits but yet not so similar that one can dispense with the pretest. The assignment of X to one group or the other is assumed to be random and under the experimenter's control.

$$\frac{O}{O} \quad \frac{X}{} \quad \frac{O}{O}$$

Two things need to be kept clear about this design: First, it is not to be confused with Design 4, the Pretest-Posttest Control Group Design, in which experimental subjects are assigned *randomly* from a common population to the experimental and the control group. Second, in spite of this, Design 10 should be recognized as well worth using in many instances in which Designs 4, 5, or 6 are impossible. In particular it should be recognized that the addition of even an unmatched or nonequivalent control group reduces greatly the equivocality of interpretation over what is obtained in Design 2, the One-Group Pretest-Posttest Design. The more similar the experimental and the con-

trol groups are in their recruitment, and the more this similarity is confirmed by the scores on the pretest, the more effective this control becomes. Assuming that these desiderata are approximated for purposes of internal validity, we can regard the design as controlling the main effects of history, maturation, testing, and instrumentation, in that the difference for the experimental group between pretest and posttest (if greater than that for the control group) cannot be explained by main effects of these variables such as would be found affecting both the experimental and the control group. (The cautions about intrasession history noted for Design 4 should, however, be taken very seriously.)

An effort to explain away a pretest-posttest gain specific to the experimental group in terms of such extraneous factors as history, maturation, or testing must hypothesize an interaction between these variables and the specific selection differences that distinguish the experimental and control groups. While in general such interactions are unlikely, there are a number of situations in which they might be invoked. Perhaps most common are interactions involving *maturation*. If the experimental group consists of psychotherapy patients and the control group some other handy population tested and retested, a gain specific to the experimental group might well be interpreted as a spontaneous remission process specific to such an extreme group, a gain that would have occurred even without *X*. Such a selection-maturation interaction (or a selection-history interaction, or a selection-testing interaction) could be mistaken for the effect of *X*, and thus represents a threat to the *internal* validity of the experiment. This possibility has been represented in the eighth column of Table 2 and is the main factor of *internal* validity which distinguishes Designs 4 and 10.

A concrete illustration from educational research may make this point clear. Sanford and Hemphill's (1952) study of the effects of a psychology course at Annapolis provides an excellent illustration of Design 10. In this study, the Second Class at Annapolis provided the experimental group and the Third Class the control group. The greater gains for the experimental group might be explained away as a part of some general sophistication process occurring maximally in the first two classes and only in minimal degree in the Third and Fourth, thus representing an interaction between the selection factors differentiating the experimental and control groups and natural changes (maturation) characteristic of these groups, rather than any effect of the experimental program. The particular control group utilized by Sanford and Hemphill makes possible some check on this rival interpretation (somewhat in the manner of Design 15 below). The selection-maturation hypothesis would predict that the Third Class (control group) in its initial test would show a superiority to the pretest measures for the Second Class (experimental group) of roughly the same magnitude as that found between the experimental group pretest and posttest. Fortunately for the interpretation of their experiment, this was not generally so. The class differences on the pretest were in most instances not in the same direction nor of the same magnitude as the pretest-posttest gains for the experimental group. However, their finding of a significant gain for the experimental group in confidence scores on the social situations questionnaire can be explained away as a selection-maturation artifact. The experimental group shows a gain from 43.26 to 51.42, whereas the Third Class starts out with a score of 55.82 and goes on to a score of 56.78.

The hypothesis of an interaction between selection and maturation will occasionally be tenable even where the groups are identical in pretest scores. The commonest of these instances will be where one group has a higher rate of maturation or autonomous change than the other. Design 14 offers an extension of 10 which would tend to rule this out.

Regression provides the other major internal validity problem for Design 10. As indicated by the "?" in Table 2, this hazard

is avoidable but one which is perhaps more frequently tripped over than avoided. In general, if either of the comparison groups has been selected for its extreme scores on *O* or correlated measures, then a difference in degree of shift from pretest to posttest between the two groups may well be a product of regression rather than the effect of *X*. This possibility has been made more prevalent by a stubborn, misleading tradition in educational experimentation, in which matching has been regarded as the appropriate and sufficient procedure for establishing the pre-experimental equivalence of groups. This error has been accompanied by a failure to distinguish Designs 4 and 10 and the quite different roles of matching on pretest scores under the two conditions. In Design 4, matching can be recognized as a useful adjunct to randomization but not as a substitute for it: in terms of scores on the pretest or on related variables, the total population available for experimental purposes can be organized into carefully matched pairs of subjects; members of these pairs can then be assigned *at random* to the experimental or the control conditions. Such matching plus subsequent randomization usually produces an experimental design with greater precision than would randomization alone.

Not to be confused with this ideal is the procedure under Design 10 of attempting to compensate for the differences between the nonequivalent experimental and control groups by a procedure of matching, when random assignment to treatments is not possible. If in Design 10 the means of the groups are substantially different, then the process of matching not only fails to provide the intended equation but in addition insures the occurrence of unwanted regression effects. It becomes predictably certain that the two groups will differ on their posttest scores altogether independently of any effects of *X*, and that this difference will vary directly with the difference between the total populations from which the selection was made and inversely with the test-retest correlation. Rulon (1941), Stanley and Beeman (1958), and R. L. Thorndike (1942) have discussed this problem thoroughly and have called attention to covariance analysis and to other statistical techniques suggested by Johnson and Neyman (see Johnson & Jackson, 1959, pp. 424–444) and by Peters and Van Voorhis (1940) for testing the effects of the experimental variable without the procedure of matching. Recent cautions by Lord (1960) concerning the analysis of covariance when the covariate is not perfectly reliable should be considered, however. Simple gain scores are also applicable but usually less desirable than analysis of covariance. Application of analysis of covariance to this Design 10 setting involves assumptions (such as that of homogeneity of regression) less plausible here than in Design 4 settings (Lindquist, 1953).

In interpreting published studies of Design 10 in which matching was used, it can be noted that the direction of error is predictable. Consider a psychotherapy experiment using ratings of dissatisfaction with one's own personality as *O*. Suppose the experimental group consists of therapy applicants and the matched control group of "normal" persons. Then the control group will turn out to represent extreme low scores from the normal group (selected because of their extremity), will regress on the posttest in the direction of the normal group average, and thus will make it less likely that a significant effect of therapy can be shown, rather than produce a spurious impression of efficacy for the therapeutic procedure.

The illustration of psychotherapy applicants also provides an instance in which the assumptions of homogeneous regression and of sampling from the same universe, except for extremity of scores, would seem likely to be inappropriate. The inclusion of normal controls in psychotherapy research is of some use, but extreme caution must be employed in interpreting results. It seems important to distinguish two versions of Design 10, and to give them different status as approximations of true experimentation. On the one hand, there is the situation in which the ex-

perimenter has two natural groups available, e.g., two classrooms, and has free choice in deciding which gets *X*, or at least has no reason to suspect differential recruitment related to *X*. Even though the groups may differ in initial means on *O*, the study may approach true experimentation. On the other hand, there are instances of Design 10 in which the respondents clearly are self-selected, the experimental group having deliberately sought out exposure to *X*, with no control group available from this same population of seekers. In this latter case, the assumption of uniform regression between experimental and control groups becomes less likely, and selection-maturation interaction (and the other selection interactions) become more probable. The "self-selected" Design 10 is thus much weaker, but it does provide information which in many instances would rule out the hypothesis that *X* has an effect. The control group, even if widely divergent in method of recruitment and in mean level, assists in the interpretation.

The threat of testing to external validity is as presented for Design 4 (see page 188). The question mark for interaction of selection and *X* reminds us that the effect of *X* may well be specific to respondents selected as the ones in our experiment have been. Since the requirements of Design 10 are likely to put fewer limitations on our freedom to sample widely than do those of Design 4, this specificity will usually be less than it would be for a laboratory experiment. The threat to external validity represented by reactive arrangements is present, but probably to a lesser degree than in most true experiments, such as Design 4.

Where one has the alternative of using two intact classrooms with Design 10, or taking random samples of the students out of the classrooms for different experimental treatments under a Design 4, 5, or 6, the latter arrangement is almost certain to be the more reactive, creating more awareness of experiment, I'm-a-guinea-pig attitude, and the like.

The Thorndike studies of formal discipline and transfer (e.g., E. L. Thorndike & Woodworth, 1901; Brolyer, Thorndike, & Woodyard, 1927) represent applications of Design 10 to *X*s uncontrolled by the experimenter. These studies avoided in part, at least, the mistake of regression effects due to simple matching, but should be carefully scrutinized in terms of modern methods. The use of covariance statistics would probably have produced stronger evidence of transfer from Latin to English vocabulary, for example.

In the other direction, the usually positive, albeit small, transfer effects found could be explained away not as transfer but as the selection into Latin courses of those students whose annual rate of vocabulary growth would have been greater than that of the control group even without the presence of the Latin instruction. This would be classified here as a selection-maturation interaction. In many school systems, this rival hypothesis could be checked by extending the range of pre-Latin *O*s considered, as in a Design 14. These studies were monumental efforts to get experimental thinking into field research. They deserve renewed attention and extension with modern methods.

11. Counterbalanced Designs

Under this heading come all of those designs in which experimental control is achieved or precision enhanced by entering all respondents (or settings) into all treatments. Such designs have been called "rotation experiments" by McCall (1923), "counterbalanced designs" (e.g., Underwood, 1949), cross-over designs (e.g., Cochran & Cox, 1957; Cox, 1958), and switch-over designs (Kempthorne, 1952). The Latin-square arrangement is typically employed in the counterbalancing. Such a Latin square is employed in Design 11, diagramed here as a quasi-experimental design, in which four experimental treatments are applied in a restrictively *randomized* manner in turn to four naturally assembled groups or even to four individuals (e.g., Maxwell, 1958):

	Time 1	*Time 2*	*Time 3*	*Time 4*
Group A	X_1O	X_2O	X_3O	X_4O
Group B	X_2O	X_4O	X_1O	X_3O
Group C	X_3O	X_1O	X_4O	X_2O
Group D	X_4O	X_3O	X_2O	X_1O

The design has been diagramed with posttests only, because it would be especially preferred where pretests were inappropriate, and designs like Design 10 were unavailable. The design contains three classifications (groups, occasions, and *X*s or experimental treatments). Each classification is "orthogonal" to the other two in that each variate of each classification occurs equally often (once for a Latin square) with each variate of each of the other classifications. To begin with, it can be noted that each treatment (each *X*) occurs once, and only once in each column and only once in each row. The same Latin square can be turned so that *X*s become row or column heads, e.g.:

	X_1	X_2	X_3	X_4
Group A	t_1O	t_2O	t_3O	t_4O
Group B	t_3O	t_1O	t_4O	t_2O
Group C	t_2O	t_4O	t_1O	t_3O
Group D	t_4O	t_3O	t_2O	t_1O

Sums of scores by *X*s thus are comparable in having each time and each group represented in each. The differences in such sums could not be interpreted simply as artifacts of the initial group differences or of practice effects, history, etc. Similarly comparable are the sums of the rows for intrinsic group differences, and the sums of the columns of the first presentation for the differences in occasions. In analysis of variance terms, the design thus appears to provide data on three main effects in a design with the number of cells usually required for two. Thinking in analysis of variance terms makes apparent the cost of this greater efficiency: What appears to be a significant main effect for any one of the three classification criteria could be instead a significant interaction of a complex form between the other two (Lindquist, 1953, pp. 258–264). The apparent differences among the effects of the *X*s could instead be a specific complex interaction effect between the group differences and the occasions. Inferences as to effects of *X* will be dependent upon the plausibility of this rival hypothesis, and will therefore be discussed in more detail.

First, let us note that the hypothesis of such interaction is more plausible for the quasi-experimental application described than for the applications of Latin squares in the true experiments described in texts covering the topic. In what has been described as the dimension of groups, two possible sources of systematic effects are confounded. First, there are the systematic selection factors involved in the natural assemblage of the groups. These factors can be expected both to have main effects and to interact with history, maturation, practice effects, etc. Were a fully controlled experiment to have been organized in this way, each person would have been assigned to each group independently and at random, and this source of both main and interaction effects would have been removed, at least to the extent of sampling error. It is characteristic of the quasi-experiment that the counterbalancing was introduced to provide a kind of equation just because such random assignment was not possible. (In contrast, in fully controlled experiments, the Latin square is employed for reasons of economy or to handle problems specific to the sampling of land parcels.) A second possible source of effects confounded with groups is that associated with specific sequences of treatments. Were all replications in a true experiment to have followed the same Latin square, this source of main and interaction effects would also have been present. In the typical *true* experiment, however, some replication sets of respondents would have been assigned different specific Latin squares, and the sys-

tematic effect of specific sequences eliminated. This also rules out the possibility that a specific systematic interaction has produced an apparent main effect of *X*s.

Occasions are likely to produce a main effect due to repeated testing, maturation, practice, and cumulative carry-overs, or transfer. History is likewise apt to produce effects for occasions. The Latin-square arrangement, of course, keeps these main effects from contaminating the main effects of *X*s. But where main effects symptomatize significant heterogeneity, one is probably more justified in suspecting significant interactions than when main effects are absent. Practice effects, for example, may be monotonic but are probably nonlinear, and would generate both main and interaction effects. Many uses of Latin squares in true experiments, as in agriculture, for instance, do not involve repeated measurements and do not typically produce any corresponding systematic column effects. Those of the cross-over type, however, share this potential weakness with the quasi-experiments.

These considerations make clear the extreme importance of replication of the quasi-experimental design with different specific Latin squares. Such replications in sufficient numbers would change the quasi-experiment into a true experiment. They would probably also involve sufficient numbers of groups to make possible the random assignment of intact groups to treatments, usually a preferable means of control. Yet, lacking such possibilities, a single Latin square represents an intuitively satisfying quasi-experimental design, because of its demonstration of all of the effects in all of the comparison groups. With awareness of the possible misinterpretations, it becomes a design well worth undertaking where better control is not possible. Having stressed its serious weaknesses, now let us examine and stress the relative strengths.

Like all quasi-experiments, this one gains strength through the consistency of the internal replications of the experiment. To make this consistency apparent, the main effects of occasions and of groups should be removed by expressing each cell as a deviation from the row (group) and column (time) means: $M_{gt}-M_{g.}-M_{.t}+M_{..}$ Then rearrange the data with treatments (*X*s) as column heads. Let us assume that the resulting picture is one of gratifying consistency, with the same treatment strongest in all four groups, etc. What are the chances of this being no true effect of treatments, but instead an interaction of groups and occasions? We can note that most possible interactions of groups and occasions would reduce or becloud the manifest effect of *X*. An interaction that imitates a main effect of *X* would be an unlikely one, and one that becomes more unlikely in larger Latin squares.

One would be most attracted to this design when one had scheduling control over a very few naturally aggregated groups, such as classrooms, but could not subdivide these natural groups into randomly equivalent subgroups for either presentation of *X* or for testing. For this situation, if pretesting is feasible, Design 10 is also available; it also involves a possible confounding of the effects of *X* with interactions of selection and occasions. This possibility is judged to be less likely in the counterbalanced design, because all comparisons are demonstrated in each group and hence several matched interactions would be required to imitate the experimental effect.

Whereas in the other designs the special responsiveness of just one of the groups to an extraneous event (history) or to practice (maturation) might simulate an effect of X_1, in the counterbalanced design such coincident effects would have to occur on separate occasions in each of the groups in turn. This assumes, of course, that we would not interpret a main effect of *X* as meaningful if inspection of the cells showed that a statistically significant main effect was primarily the result of a very strong effect in but one of the groups. For further discussion of this matter, see the reports of Wilk and Kempthorne (1957), Lubin (1961), and Stanley (1955).

12. The Separate-Sample Pretest-Posttest Design

For large populations, such as cities, factories, schools, and military units, it may often happen that although one cannot randomly segregate subgroups for differential experimental treatments, one can exercise something like full experimental control over the *when* and *to whom* of the *O*, employing random assignment procedures. Such control makes possible Design 12:

$$\begin{array}{lll} R & O & (X) \\ R & X & O \end{array}$$

In this diagram, rows represent randomly equivalent subgroups, the parenthetical *X* standing for a presentation of *X* irrelevant to the argument. One sample is measured prior to the *X*, an equivalent one subsequent to *X*. The design is not inherently a strong one, as is indicated by its row in Table 2. Nevertheless, it may frequently be all that is feasible, and is often well worth doing. It has been used in social science experiments which remain the best studies extant on their topics (e.g., Star & Hughes, 1950). While it has been called the "simulated before-and-after design" (Selltiz, Jahoda, Deutsch, & Cook, 1959, p. 116), it is well to note its superiority over the ordinary before-and-after design, Design 2, through its control of both the main effect of testing and the interaction of testing with *X*. The main weakness of the design is its failure to control for history. Thus in the study of the Cincinnati publicity campaign for the United Nations and UNESCO (Star & Hughes, 1950), extraneous events on the international scene probably accounted for the observed decrease in optimism about getting along with Russia.

It is in the spirit of this chapter to encourage "patched-up" designs, in which features are added to control specific factors, more or less one at a time (in contrast with the neater "true" experiments, in which a single control group controls for all of the threats to internal validity). Repeating Design 12 in different settings at different times, as in Design 12*a* (see Table 2, p. 210), controls for history, in that if the same effect is repeatedly found, the likelihood of its being a product of coincidental historical events becomes less likely. But consistent secular historical trends or seasonal cycles still remain uncontrolled rival explanations. By replicating the effect under other settings, one can reduce the possibility that the observed effect is specific to the single population initially selected. However, if the setting of research permits Design 12*a*, it will also permit Design 13, which would in general be preferred.

Maturation, or the effect of the respondents' growing older, is unlikely to be invoked as a rival explanation, even in a public opinion survey study extending over months. But, in the sample survey setting, or even in some college classrooms, the samples are large enough and ages heterogeneous enough so that subsamples of the pretest group differing in maturation (age, number of semesters in college, etc.) can be compared. Maturation, and the probably more threatening possibility of secular and seasonal trends, can also be controlled by a design such as 12*b* which adds an additional earlier pretest group, moving the design closer to the time-series design, although without the repeated testing. For populations such as psychotherapy applicants, in which healing or spontaneous remission might take place, the assumptions of linearity implicitly involved in this control might not be plausible. It is more likely that the maturational trend will be negatively accelerated, hence will make the O_1—O_2 maturational gain larger than that for O_2—O_3, and thus work against the inter pretation that *X* has had an effect.

Instrumentation represents a hazard in this design when employed in the sample survey setting. If the same interviewers are employed in the pretest and in the posttest, it usually happens that many were doing their first interviewing on the pretest and are more experienced, or perhaps more cynical, on the posttest. If the interviewers differ on each

wave and are few, differences in interviewer idiosyncrasies are confounded with the experimental variable. If the interviewers are aware of the hypothesis, and whether or not the *X* has been delivered, then interviewer expectations may create differences, as Stanton and Baker (1942) and Smith and Hyman (1950) have shown experimentally. Ideally, one would use equivalent random samples of different interviewers on each wave, and keep the interviewers in ignorance of the experiment. In addition, the recruitment of interviewers may show differences on a seasonal basis, for instance, because more college students are available during summer months, etc. Refusal rates are probably lower and interview lengths longer in summer than in winter. For questionnaires which are self-administered in the classroom, such instrument error may be less likely, although test-taking orientations may shift in ways perhaps better classifiable as instrumentation than as effects of *X* upon *O*.

For pretests and posttests separated in time by several months, mortality can be a problem in Design 12. If both samples are selected at the same time (point *R*), as time elapses, more members of the selected sample can be expected to become inaccessible, and the more transient segments of the population to be lost, producing a population difference between the different interviewing periods. Differences between groups in the number of noncontacted persons serve as a warning of this possibility.

Perhaps for studies over long periods the pretest and posttest samples should be selected independently and at appropriately different times, although this, too, has a source of systematic bias resulting from possible changes in the residential pattern of the universe as a whole. In some settings, as in schools, records will make possible the elimination of the pretest scores of those who have become unavailable by the time of the posttest, thus making the pretest and posttest more comparable. To provide a contact making this correction possible in the sample survey, and to provide an additional confirmation of effect which mortality could not contaminate, the pretest group can be retested, as in Design 12*c*, where the O_1—O_2 difference should confirm the O_1—O_3 comparison. Such was the study by Duncan, et al. (1957) on the reduction in fallacious beliefs effected by an introductory course in psychology. (In this design, the retested group does not make possible the examination of the gains for persons of various initial scores because of the absence of a control group to control for regression.)

It is characteristic of this design that it moves the laboratory into the field situation to which the researcher wishes to generalize, testing the effects of *X* in its natural setting. In general, as indicated in Tables 1 and 2, Designs 12, 12*a*, 12*b*, and 12*c* are apt to be superior in external validity or generalizability to the "true" experiments of Designs 4, 5, and 6. These designs put so little demand upon the respondents for cooperation, for being at certain places at certain times, etc., that representative sampling from populations specified in advance can be employed.

In Designs 12 and 13 (and, to be sure, in some variants on Designs 4 and 6, where *X* and *O* are delivered through individual contacts, etc.) representative sampling is possible. The pluses in the selection -*X* interaction column are highly relative and could, in justice, be changed to question marks, since in general practice the units are not selected for their theoretical relevance, but often for reasons of cooperativeness and accessibility, which make them likely to be atypical of the universe to which one wants to generalize.

It was not to Cincinnati but rather to Americans in general, or to people in general, that Star and Hughes (1950) wanted to generalize, and there remains the possibility that the reaction to *X* in Cincinnati was atypical of these universes. But the degree of such accessibility bias is so much less than that found in the more demanding designs that a comparative plus seems justified.

13. The Separate-Sample Pretest-Posttest Control Group Design

It is expected that Design 12 will be used in those settings in which the *X*, if presented at all, must be presented to the group as a whole. If there are comparable (if not equivalent) groups from which *X* can be withheld, then a control group can be added to Design 12, creating Design 13:

$$\begin{array}{l} R\ O\ (X)\ O \\ R\ \ \ X\ O \\ \hline R\ O \\ R\ \ \ \ \ O \end{array}$$

This design is quite similar to Design 10, except that the same specific persons are not retested and thus the possible interaction of testing and *X* is avoided. As with Design 10, the weakness of Design 13 for internal validity comes from the possibility of mistaking for an effect of *X* a specific local trend in the experimental group which is, in fact, unrelated. By increasing the number of the social units involved (schools, cities, factories, ships, etc.) and by assigning them in some number and with randomization to the experimental and control treatments, the one source of invalidity can be removed, and a true experiment, like Design 4 except for avoiding the retesting of specific individuals, can be achieved. This design can be designated 13*a*. Its diagraming (in Table 3) has been complicated by the two levels of equivalence (achieved by random assignment) which are involved. At the level of respondents, there is within each social unit the equivalence of the separate pretest and posttest samples, indicated by the point of assignment *R*. Among the several social units receiving either treatment, there is no such equivalence, this lack being indicated by the dashed line. The *R'* designates the equation of the experimental group and the control group by the random assignment of these numerous social units to one or another treatment.

As can be seen by the row for 13*a* in Table 3, this design receives a perfect score for both internal and external validity, the latter on grounds already discussed for Design 12 with further strength on the selection-*X* interaction problem because of the representation of numerous social units, in contrast with the use of a single one. As far as is known, this excellent but expensive design has not been used.

14. The Multiple Time-Series Design

In studies of major administrative change by time-series data, the researcher would be wise to seek out a similar institution not undergoing the *X*, from which to collect a similar "control" time series (ideally with *X* assigned randomly):

$$\begin{array}{l} O\ O\ O\ O X O\ O\ O\ O \\ \hline O\ O\ O\ O\ \ O\ O\ O\ O \end{array}$$

This design contains within it (in the *O*s bracketing the *X*) Design 10, the Nonequivalent Control Group Design, but gains in certainty of interpretation from the multiple measures plotted, as the experimental effect is in a sense twice demonstrated, once against the control and once against the pre-*X* values in its own series, as in Design 7. In addition, the selection-maturation interaction is controlled to the extent that, if the experimental group showed in general a greater rate of gain, it would show up in the pre-*X* *O*s. In Tables 2 and 3 this additional gain is poorly represented, but appears in the final internal validity column, which is headed "Interaction of Selection and Maturation." Because maturation is controlled for both experimental and control series, by the logic discussed in the first presentation of the Time-Series Design 7 above, the difference in the selection of the groups operating in conjunction with maturation, instrumentation, or regression, can hardly account for an apparent effect. An interaction of the se-

TABLE 3
SOURCES OF INVALIDITY FOR QUASI-EXPERIMENTAL DESIGNS 13 THROUGH 16

	Sources of Invalidity											
	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation, etc.	Interaction of Testing and X	Interaction of Selection and X	Reactive Arrangements	Multiple-X Interference
Quasi-Experimental Designs Continued:												
13. Separate-Sample Pretest-Posttest Control Group Design R O (X) O R X O --- R O R O	+	+	+	+	+	+	+	−	+	+	+	
13a. R O (X) O R X O --- R O (X) O R X O --- R O (X) O R X O R O R O --- R O R O --- R O R O	+	+	+	+	+	+	+	+	+	+	+	
14. Multiple Time-Series O O O X O O O --- O O O O O O O	+	+	+	+	+	+	+	+	−	−	?	
15. Institutional Cycle Design Class A X O_1 --- Class B_1 R O_2 X O_3 Class B_2 R X O_4 --- Class C O_5 X --- [a]Gen. Pop. Con. Cl. B O_6 [a]Gen. Pop. Con. Cl. C O_7												
$O_2 < O_1$ $O_5 < O_4$	+	−	+	+	?	−	?		+	?	+	
$O_2 < O_3$	−	−	−	?	?	+	+		−	?	+	
$O_2 < O_4$	−	−	+	?	?	+	?		+	?	?	
$O_6 = O_7$ $O_{2y} = O_{2o}$		+						−				
16. Regression Discontinuity	+	+	+	?	+	+	?	+	+	−	+	+

[a] General Population Controls for Class B, etc.

lection difference with history remains, however, a possibility.

As with the Time-Series Design 7, a minus has been entered in the external validity column for testing-X interaction, although as with Design 7, the design would often be used where the testing was nonreactive. The standard precaution about the possible specificity of a demonstrated effect of X to the population under study is also recorded in Table 3. As to the tests of significance, it is suggested that differences between the experimental and control series be analyzed as Design 7 data. These differences seem much more likely to be linear than raw time-series data.

In general, this is an excellent quasi-experimental design, perhaps the best of the more feasible designs. It has clear advantages over Designs 7 and 10, as noted immediately above and in the Design 10 presentation. The availability of repeated measurements makes the Multiple Time Series particularly appropriate to research in schools.

15. The Recurrent Institutional Cycle Design: A "Patched-Up" Design

Design 15 illustrates a strategy for field research in which one starts out with an inadequate design and then adds specific features to control for one or another of the recurrent sources of invalidity. The result is often an inelegant accumulation of precautionary checks, which lacks the intrinsic symmetry of the "true" experimental designs, but nonetheless approaches experimentation. As a part of this strategy, the experimenter must be alert to the rival interpretations (other than an effect of X) which the design leaves open and must look for analyses of the data, or feasible extensions of the data, which will rule these out. Another feature often characteristic of such designs is that the effect of X is demonstrated in several different manners. This is obviously an important feature where each specific comparison would be equivocal by itself.

The specific "patched-up" design under discussion is limited to a narrow set of questions and settings, and opportunistically exploits features of these settings. The basic insight involved can be noted by an examination of the second and third rows of Table 1, in which it can be seen that the patterns of plus and minus marks for Designs 2 and 3 are for the most part complementary, and that hence the right combination of these two inadequate arguments might have considerable strength. The design is appropriate to those situations in which a given aspect of an institutional process is, on some cyclical schedule, continually being presented to a new group of respondents. Such situations include schools, indoctrination procedures, apprenticeships, etc. If in these situations one is interested in evaluating the effects of such a global and complex X as an indoctrination program, then the Recurrent Institutional Cycle Design probably offers as near an answer as is available from the designs developed thus far.

The design was originally conceptualized in the context of an investigation of the effects of one year's officer and pilot training upon the attitudes toward superiors and subordinates and leadership functions of a group of Air Force cadets in the process of completing a 14-month training cycle (Campbell & McCormack, 1957). The restriction precluding a true experiment was the inability to control who would be exposed to the experimental variable. There was no possibility of dividing the entering class into two equated halves, one half of which would be sent through the scheduled year's program, and the other half sent back to civilian life. Even were such a true experiment feasible (and opportunistic exploitation of unpredicted budget cuts might have on several occasions made such experiments possible), the reactive effects of such experimental arrangements, the disruption in the lives of those accepted, screened, and brought to the air base and then sent home, would have made them far from an ideal control group. The difference between them and the experimental group receiving indoctrination would

hardly have been an adequate base from which to generalize to the normal conditions of recruitment and training. There remained, however, the experimenter's control over the scheduling of the *when* and *to whom* of the observational procedures. This, plus the fact that the experimental variable was recurrent and was continually being presented to a new group of respondents, made possible some degree of experimental control. In that study two kinds of comparisons relevant to the effect of military experience on attitudes were available. Each was quite inadequate in terms of experimental control, but when both provided confirmatory evidence they were mutually supportive inasmuch as they both involved different weaknesses. The first involved comparisons among populations measured at the same time but varying in their length of service. The second involved measures of the same group of persons in their first week of military training and then again after some 13 months. In idealized form this design is as follows:

Class A	X	O_1	
Class B		O_2	X O_3

This design combines the "longitudinal" and "cross-sectional" approaches commonly employed in developmental research. In this it is assumed that the scheduling is such that at one and the same time a group which has been exposed to X and a group which is just about to be exposed to it can be measured; this comparison between O_1 and O_2 thus corresponds to the Static-Group Comparison, Design 3. Remeasuring the personnel of Class B one cycle later provides the One-Group Pretest-Posttest segment, Design 2. In Table 3, on page 226, the first two rows dealing with Design 15 show an analysis of these comparisons. The cross-sectional comparison of $O_1 > O_2$ provides differences which could not be explained by the effects of history or a test-retest effect. The differences obtained could, however, be due to differences in recruitment from year to year (as indicated by the minus opposite selection) or by the fact that the respondents were one year older (the minus for maturation). Where the testing is all done at the same time period, the confounded variable of instrumentation, or shifts in the nature of the measuring instrument, seem unlikely. In the typical comparison of the differences in attitudes of freshmen and sophomores, the effect of mortality is also a rival explanation: O_1 and O_2 might differ just because of the kind of people that have dropped out from Class A but are still represented in Class B. This weakness is avoidable if the responses are identified by individuals, and if the experimenter waits before analyzing his data until Class B has completed its exposure to X and then eliminates from O_2 all of those measures belonging to respondents who later failed to complete the training. The frequent absence of this procedure justifies the insertion of a question mark opposite the mortality variable. The regression column is filled with question marks to warn of the possibility of spurious effects if the measure which is being used in the experimental design is the one on which the acceptance and rejection of candidates for the training course was based. Under these circumstances consistent differences which should not be attributed to the effects of X would be anticipated. The pretest-posttest comparison involved in O_2 and O_3, if it provides the same type of difference as does the $O_2 - O_1$ comparison, rules out the rival hypotheses that the difference is due to a shift in the selection or recruitment between the two classes, and also rules out any possibility that mortality is the explanation. However, were the $O_2 - O_3$ comparison to be used alone, it would be vulnerable to the rival explanations of history and testing.

In a setting where the training period under examination is one year, the most expensive feature of the design is the scheduling of the two sets of measurements a year apart. Given the investment already made in this, it constitutes little additional expense

to do more testing on the second occasion. With this in mind, one can expand the recurrent institutional design to the pattern shown in Table 3. Exercising the power to designate who gets measured and when, Class B has been broken into two equated samples, one measured both before and after exposure, and the other measured only after exposure as in O_4. This second group provides a comparison on carefully equated samples of an initial measure coming before and after, is more precise than the $O_1 - O_2$ comparison as far as selection is concerned, and is superior to the $O_2 - O_3$ comparison in avoiding test-retest effects. The effect of X is thus documented in three separate comparisons, $O_1 > O_2$, $O_2 < O_3$ and $O_2 < O_4$.

Note, however, that O_2 is involved in all of these three, and thus all might appear to be confirmatory just because of an eccentric performance of that particular set of measurements. The introduction of O_5, that is Class C, tested on the second testing occasion prior to being exposed to X, provides another pre-X measure to be compared with O_4 and O_1, etc., providing a needed redundancy. The splitting of Class B makes this O_4—O_5 comparison more clear-cut than would be an O_3—O_5 comparison. Note, however, that the splitting of a class into the tested and the nontested half often constitutes a "reactive arrangement." For this reason a question mark has been inserted for that factor in the $O_2 < O_4$ row in Table 3. Whether or not this is a reactive procedure depends upon the specific conditions. Where lots are drawn and one half of the class is asked to go to another room, the procedure is likely to be reactive (e.g., Duncan, et al., 1957; Solomon, 1949). Where, as in many military studies, the contacts have been made individually, a class can be split into equated halves without this conspicuousness. Where a course consists of a number of sections with separate schedules, there is the possibility of assigning these intact units to the pretest and no-pretest groups (e.g., Hovland, Lumsdaine, & Sheffield, 1949). For a single classroom, the strategy of passing out questionnaires or tests to everyone but varying the content so that a random half would get what would constitute the pretest and the other half get tested on some other instrument may serve to make the splitting of the class no more reactive than the testing of the whole class would be.

The design as represented through measurements O_1 to O_5 uniformly fails to control for maturation. The seriousness of this limitation will vary depending upon the subject material under investigation. If the experiment deals with the acquisition of a highly esoteric skill or competence, the rival hypothesis of maturation—that just growing older or more experienced in normal everyday societal ways would have produced this gain—may seem highly unlikely.

In the cited study of attitudes toward superiors and subordinates (Campbell & McCormack, 1957), however, the shift was such that it might very plausibly be explained in terms of an increased sophistication which a group of that age and from that particular type of background would have undergone through growing older or being away from home in almost any context. In such a situation a control for maturation seems very essential. For this reason O_6 and O_7 have been added to the design, to provide a cross-sectional test of a general maturation hypothesis made on the occasion of the second testing period. This would involve testing two groups of persons from the general population who differ only in age and whose ages were picked to coincide with those of Class B and Class C at the time of testing. To confirm the hypothesis of an effect of X, the groups O_6 and O_7 should turn out to be equal, or at least to show less discrepancy than do the comparisons spanning exposure to X. The selection of these general population controls would depend upon the specificity of the hypothesis. Considering our knowledge as to the ubiquitous importance of social class and educational considerations, these controls might be selected so as to match the institutional recruitment on social class and previous education. They might

also be persons who are living away from home for the first time and who are of the typical age of induction, so that, in the illustration given, the O_6 group would have been away from home one year and the O_7 group just barely on the verge of leaving home. These general population age-mate controls would always be to some extent unsatisfactory and would represent the greatest cost item, since testing within an institutional framework is generally easier than selecting cases from a general population. It is for this reason that O_6 and O_7 have been scheduled with the second testing wave, for if no effect of X is shown in the first body of results (the comparison $O_1 > O_2$), then these expensive procedures would usually be unjustified (unless, for example, one had the hypothesis that the institutional X had suppressed a normal maturational process).

Another cross-sectional approach to the control of maturation may be available if there is heterogeneity in age (or years away from home, etc.) within the population entering the institutional cycle. This would be so in many situations; for example, in studying the effects of a single college course. In this case, the measures of O_2 could be subdivided into an older and younger group to examine whether or not these two subgroups (O_{2_o} and O_{2_y} in Table 3) differed as did O_1 and O_2 (although the ubiquitous negative correlation between age and ability *within* school grades, etc., introduces dangers here). Better than the general population age-mate control might be the comparison with another specific institution, as comparing Air Force inductees with first-year college students. If the comparison is to be made of this type, one reduces one's experimental variable to those features which the two types of institution *do not* have in common. In this case, the generally more efficient Designs 10 and 13 would probably be as feasible.

The formal requirements of this design would seem to be applicable even to such a problem as that of psychotherapy. This possibility reveals how difficult a proper check on the maturation variable is. No matter how the general population controls for a psychotherapy situation are selected, if they are not themselves applicants for psychotherapy they differ in important ways. Even if they are just as ill as a psychotherapy applicant, they almost certainly differ in their awareness of, beliefs about, and faith in psychotherapy. Such an ill but optimistic group might very well have recovery potentialities not typical of any matching group that we would be likely to obtain, and thus an interaction of selection and maturation could be misinterpreted as an effect of X.

For the study of developmental processes per se, the failure to control maturation is of course no weakness, since maturation is the focus of study. This combination of longitudinal and cross-sectional comparisons should be more systematically employed in developmental studies. The cross-sectional study by itself confounds maturation with selection and mortality. The longitudinal study confounds maturation with repeated testing and with history. It alone is probably no better than the cross-sectional, although its greater cost gives it higher prestige. The combination, perhaps with repeated cross-sectional comparisons at various times, seems ideal.

In the diagrams of Design 15 as presented, it is assumed that it will be feasible to present the posttest for one group at the same chronological time as the pretest for another. This is not always the case in situations where we might want to use this design. The following is probably a more accurate portrayal of the typical opportunity in the school situation:

Class A	X	O_1				
	—	—	—	—	—	—
Class B_1			RO_2	X	O_3	
Class B_2			R	X	O_4	
	—	—	—	—	—	—
Class C					O_5	X

Such a design lacks the clear-cut control on history in the $O_1 > O_2$ and the $O_4 > O_5$ comparisons because of the absence of simul-

taneity. However, the explanation in terms of history could hardly be employed if both comparisons show the effect, except by postulating quite a complicated series of coincidences.

Note that any general historical trend, such as we certainly do find with social attitudes, is not confounded with clear-cut experimental results. Such a trend would make O_2 intermediate between O_1 and O_3, while the hypothesis that X has an effect requires O_1 and O_3 to be equal, and O_2 to differ from both in the same direction. In general, with replication of the experiment on several occasions, the confound with history is unlikely to be a problem even in this version of the design. But, for institutional cycles of less than a calendar year, there may be the possibility of confounding with seasonal variations in attitudes, morale, optimism, intelligence, or what have you. If the X is a course given only in the fall semester, and if between September and January people generally increase in hostility and pessimism because of seasonal climatic factors, this recurrent seasonal trend is confounded with the effects of X in all of its manifestations. For such settings, Designs 10 and 13 are available and to be preferred.

If the cross-sectional and longitudinal comparisons indicate comparable effects of X, this could not be explained away as an interaction between maturation and the selection differences between the classes. However, because this control does not show up in the segmental presentations in Table 3, the column has been left blank. The ratings on external validity criteria, in general, follow the pattern of the previous designs containing the same fragments. The question marks in the "Interaction of Selection and X" column merely warn that the findings are limited to the institutional cycle under study. Since the X is so complex, the investigation is apt to be made for practical reasons rather than theoretical purposes, and for these practical purposes, it is probably to this one institution that one wants to generalize in this case.

16. Regression-Discontinuity Analysis

This is a design developed in a situation in which ex post facto designs were previously being used. While very limited in range of possible applications, its presentation here seems justified by the fact that those limited settings are mainly educational. It also seems justifiable as an illustration of the desirability of exploring in each specific situation all of the implications of a causal hypothesis, seeking novel outcroppings where the hypothesis might be exposed to test. The setting (Thistlethwaite & Campbell, 1960) is one in which awards are made to the most qualified applicants on the basis of a cutting score on a quantified composite of qualifications. The award might be a scholarship, admission to a university so sought out that all accepted enrolled, a year's study in Europe, etc. Subsequent to this event, applicants receiving and not receiving the award are measured on various Os representing later achievements, attitudes, etc. The question is then asked, Did the award make a difference? The problem of inference is sticky just because almost all of the qualities leading to eligibility for the award (except such factors as need and state of residence, if relevant) are qualities which would have led to higher performance on these subsequent Os. We are virtually certain in advance that the recipients would have scored higher on the Os than the nonrecipients even if the award had not been made.

Figure 4 presents the argument of the design. It illustrates the expected relation of pre-award ability to later achievement, plus the added results of the special educational or motivational opportunities resulting. Let us first consider a true experiment of a Design 6 sort, with which to contrast our quasi-experiment. This true experiment might be rationalized as a tie-breaking process, or as an experiment in extension of program, in which, for a narrow range of scores at or just below the cutting point, random assignment would create an award-winning experimental group and a nonwinning control

group. These would presumably perform as the two circle-points at the cutting line in Fig. 4. For this narrow range of abilities, a true experiment would have been achieved. *Such experiments are feasible and should be done.*

The quasi-experimental Design 16 attempts to substitute for this true experiment by examining the regression line for a discontinuity at the cutting point which the causal hypothesis clearly implies. If the outcome were as diagramed, and if the circle-points in Fig. 4 represented extrapolations from the two halves of the regression line rather than a randomly split tie-breaking experiment, the evidence of effect would be quite compelling, almost as compelling as in the case of the true experiment.

Some of the tests of significance discussed for Design 7 are relevant here. Note that the hypothesis is clearly one of intercept difference rather than slope, and that the location of the step in the regression line must be right at the X point, no "lags" or "spreads" being consistent with the hypothesis. Thus parametric and nonparametric tests avoiding assumptions of linearity are appropriate. Note also that assumptions of linearity are usually more plausible for such regression data than for time series. (For certain types of data, such as percentages, a linearizing transformation may be needed.) This might make a t test for the difference between the two linearly extrapolated points appropriate. Perhaps the most efficient test would be a covariance analysis, in which the award-decision score would be the covariate of later achievement, and award and no-award would be the treatment.

Is such a design likely to be used? It certainly applies to a recurrent situation in which claims for the efficacy of X abound. Are such claims worth testing? One sacrifice required is that all of the ingredients going into the final decision be pooled into a composite index, and that a cutting point be cleanly applied. But certainly we are convinced by now that all of the qualities lead-

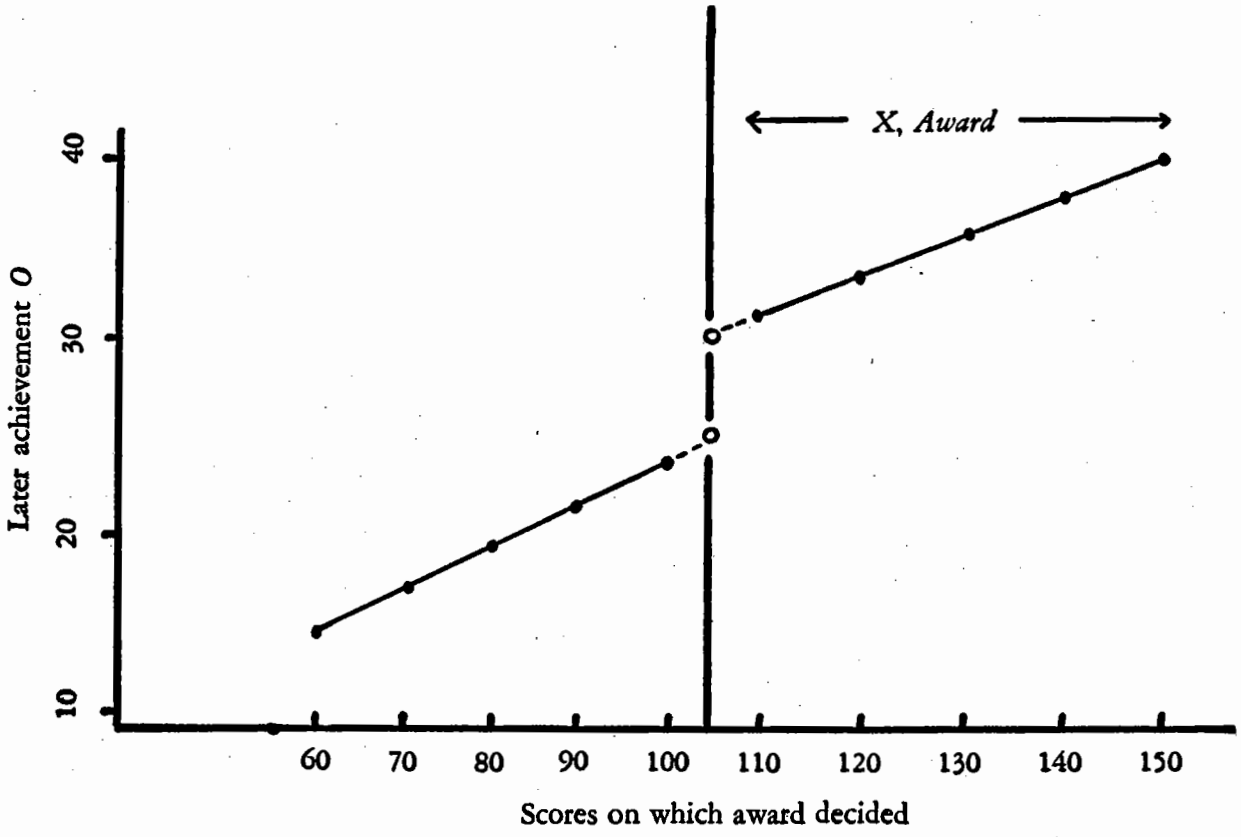

Fig. 4. Regression-Discontinuity Analysis.

ing to a decision—the appearance of the photograph, the class standing discounted by the high school's reputation, the college ties of the father, etc., can be put into such a composite, by ratings if by no more direct way. And we should likewise by now be convinced (Meehl, 1954) that a multiple correlational weighting formula for combining the ingredients (even using past committee decisions as a criterion) is usually better than a committee's case-by-case ponderings. Thus, we would have nothing to lose and much to gain for all purposes by quantifying award decisions of all kinds. If this were done, and if files were kept on awards and rejections, then years later follow-ups of effects could be made.

Perhaps a true parable is in order: A generous foundation interested in improving higher education once gave an Ivy League college half a million dollars to study the impact of the school upon its students. Ten years later, not a single research report remotely touching upon this purpose had appeared. Did the recipients or donors take the specifics of the formal proposal in any way seriously? Was the question in any way answerable? Designs 15 and 16 seem to offer the only possible approximations. But, of course, perhaps no scientist has any real curiosity about the effects of such a global *X*.

To go through the check-off in Table 3: Because of synchrony of experimental and control group, history and maturation seem controlled. Testing as a main effect is controlled in that both the experimental and control groups have received it. Instrumentation errors might well be a problem if the follow-up *O* was done under the auspices which made the award, in that gratitude for the award and resentment for not receiving the award might lead to differing expressions of attitude, differing degrees of exaggeration of one's own success in life, etc. This weakness would also be present in the tie-splitting true experiment. It could be controlled by having the follow-ups done by a separate agency. We believe, following the arguments above, that both regression and selection are controlled as far as their possible spurious contributions to inference are concerned, even though selection is biased and regression present—both have been controlled through representing them in detail, not through equation. Mortality would be a problem if the awarding agency conducted the follow-up measure, in that award recipients, alumni, etc., would probably cooperate much more readily than nonwinners. Note how the usually desirable wish of the researcher to achieve complete representation of the selected sample may be misleading here. If conducting the follow-up with a different letterhead would lead to a drop in cooperation from, say, 90 per cent to 50 per cent, an experimenter might be reluctant to make the shift because his goal is a 100 per cent representation of award winners. He is apt to forget that his true goal is interpretable data, that no data are interpretable in isolation, and that a comparable contrast group is essential to make use of his data on award winners. Both for this reason and because of the instrumentation problem, it might be scientifically better to have independent auspices and a 50 per cent return from both groups instead of a 90 per cent return from award winners and a 50 per cent return from the nonwinners. Again, the mortality problem would be the same for the tie-breaking true experiment. For both, the selection-maturation interaction threat to internal validity is controlled. For the quasi-experiment, it is controlled in that this interaction could not lawfully explain a distinct discontinuity in the regression line at *X*. The external validity threat of a testing-*X* interaction is controlled to the extent that the basic measurements used in the award decision are a part of the universe to which one wants to generalize.

Both the tie-breaking true experiment and the regression-discontinuity analysis are particularly subject to the external-validity limitation of selection-*X* interaction in that the effect has been demonstrated only for a very narrow band of talent, i.e., only for those at the cutting score. For the quasi-experiment, the possibilities of inference may seem broad-

er, but note that the evils of the linear fit assumption are minimal when extrapolated but one point, as in the design as illustrated in Fig. 4. Broader generalizations involve the extrapolation of the below-*X* fit across the entire range of *X* values, and at each greater degree of extrapolation the number of plausible rival hypotheses becomes greater. Also, the extrapolated values of different types of curves fitted to the below-*X* values become more widely spread, etc.