women, we recoil from the implication that our harsh course of training is debeautifying, and instead point to the hazards in the way of a beautiful girl's finishing college before getting married. Such an effect is classified here as experimental *mortality*. (Of course, if we consider the *same* girls when they are freshmen and seniors, this problem disappears, and we have Design 2.)

THREE TRUE EXPERIMENTAL DESIGNS

The three basic designs to be treated in this section are the currently recommended designs in the methodological literature. They will also turn out to be the most strongly recommended designs of this presentation, even though this endorsement is subject to many specific qualifications regarding usual practice and to some minus signs in Table 1 under *external validity*. Design 4 is the most used of the three, and for this reason we allow its presentation to be disproportionately extended and to become the locus of discussions more generally applicable. Note that all three of these designs are presented in terms of a single X being compared with *no* X. Designs with more numerous treatments in the Fisher factorial experiment tradition represent important elaborations tangential to the main thread of this chapter and are discussed at the end of this section, subsequent to Design 6. But this perspective can serve to remind us at this point that the comparison of X with *no* X is an oversimplification. The comparison is actually with the specific activities of the control group which have filled the time period corresponding to that in which the experimental group receives the X. Thus the comparison might better be between X_1 and X_c, or between X_1 and X_o, or X_1 and X_2. That these control group activities are often unspecified adds an undesirable ambiguity to the interpretation of the contribution of X. Bearing these comments in mind, we will continue in this section the graphic convention of presenting no X in the control group.

4. The Pretest-Posttest Control Group Design

Controls for Internal Validity

One or another of the above considerations led psychological and educational researchers between 1900 and 1920 to add a control group to Design 2, creating the presently orthodox control group design. McCall (1923), Solomon (1949), and Boring (1954) have given us some of this history, and a scanning of the *Teachers College Record* for that period implies still more, for as early as 1912 control groups were being referred to without need of explanation (e.g., Pearson, 1912). The control group designs thus introduced are classified in this chapter under two heads: the present Design 4 in which equivalent groups as achieved by randomization are employed, and the quasi-experimental Design 10 in which extant intact comparison groups of unassured equivalence are employed. Design 4 takes this form:

$$\begin{array}{llll} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \end{array}$$

Because the design so neatly controls for *all* of the seven rival hypotheses described so far, the presentations of it have usually not made explicit the control needs which it met. In the tradition of learning research, the practice effects of *testing* seem to provide the first recognition of the need for a control group. *Maturation* was a frequent critical focus in experimental studies in education, as well as in the nature-nurture problem in the child development area. In research on attitude change, as in the early studies on the effects of motion pictures, *history* may have been the main necessitating consideration. In any event, it seems desirable here to discuss briefly the way in which, or the conditions under which, these factors are controlled.

History is controlled insofar as general historical events that might have produced an O_1—O_2 difference would also produce an O_3—O_4 difference. Note, however, that

many supposed utilizations of Design 4 (or 5 or 6) do *not* control for unique *intrasession history*. If all of the randomly assigned students in the experimental group are treated in a single session, and similarly the control students in another single session, then the irrelevant unique events in either session (the obstreperous joke, the fire across the street, the experimenter's introductory remarks, etc.) become rival hypotheses explaining the O_1—O_2 versus O_3—O_4 difference. Such an experiment is *not* a true experiment, even when presented, as was Solomon's (1949) experiment on the teaching of spelling, as an illustrative paradigm. (To be fair, we point out that it was chosen to illustrate a different point.) Thinking over our "best practice" on this point may make this seem a venial sin, but our "best practice" is producing experiments too frequently unreplicable, and this very source of "significant" but extraneous differences might well be an important fault. Furthermore, the typical experiment in the *Journal of Experimental Psychology* does achieve control of intrasession history through testing students and animals individually and through assigning the students and experimental periods at random to experimental or control conditions. Note, however, that even with individual sessions, history can be uncontrolled if all of the experimental group is run before the control group, etc. Design 4 calls for simultaneity of experimental and control sessions. If we actually run sessions simultaneously, then different experimenters must be used, and experimenter differences can become a form of intrasession history confounded with X.

The optimal solution is a randomization of experimental occasions, with such restrictions as are required to achieve balanced representation of such highly likely sources of bias as experimenters, time of day, day of week, portion of semester, nearness to examinations, etc. The common expedient of running experimental subjects in small groups rather than individually is inadmissible if this grouping is disregarded in the statistical analysis. (See the section on assigning intact groups to treatments, below.) All those in the same session share the same intrasession history, and thus have sources of similarity other than X. If such sessions have been assigned at random, the correct statistical treatment is the same as that discussed below for the assignment of intact classrooms to treatments. (For some studies involving group testing, the several experimental treatments can be randomly distributed within one face-to-face group, as in using multiple test forms in a study of the effect of the order of difficulty of items. In such cases, the specificities of intrasession history are common to both treatments and do not become a plausible rival hypothesis confounded with X in explaining the differences obtained.)

Maturation and *testing* are controlled in that they should be manifested equally in experimental and control groups. *Instrumentation* is easily controlled where the conditions for the control of intrasession history are met, particularly where the O is achieved by student responses to a fixed instrument such as a printed test. Where observers or interviewers are used, however, the problem becomes more serious. If observers are few enough not to be randomly assignable to the observation of single sessions, then not only should each observer be used for both experimental and control sessions, but in addition, the observers should be kept ignorant as to which students are receiving which treatments, lest the knowledge bias their ratings or records. That such bias tendencies are "dependable" sources of variance is affirmed by the necessity in medical research of the second blind in the double-blind experiment, by recent research (Rosenthal, 1959), and by older studies (e.g., Kennedy & Uphoff, 1939; Stanton & Baker, 1942). The use of recordings of group interaction, so that judges may judge a series of randomized sections of pretest, posttest, experimental, and control group transcriptions, helps to control instrumentation in research on classroom behavior and group interaction.

Regression is controlled as far as mean differences are concerned, no matter how extreme the group is on pretest scores, if both experimental and control groups are randomly assigned from this same extreme pool. In such a case, the control group regresses as much as does the experimental group. Interpretative lapses due to regression artifacts do frequently occur, however, even under Design 4 conditions. An experimenter may employ the control group to confirm group mean effects of X, and then abandon it while examining which pretest-score subgroups of the experimental group were most influenced. If the whole group has shown a gain, then he arrives at the stimulating artifact that those initially lowest have gained most, those initially highest perhaps not at all. This outcome is assured because under conditions of total group mean gain, the regression artifact supplements the gain score for the below-mean pretest scorers, and tends to cancel it for the high pretest scorers. (If there was no over-all gain, then the experimenter may mistakenly "discover" that this was due to two mutually cancelling effects, for those low to gain, those high to lose.) One cure for these misinterpretations is to make parallel analyses of extreme pretest scorers in the control group, and to base differential gain interpretations on comparisons of the posttest scores of the corresponding experimental and control pretest subgroups. (Note, however, that skewed distributions resulting from selection make normal-curve statistics of dubious appropriateness.)

Selection is ruled out as an explanation of the difference to the extent that randomization has assured group equality at time R. This extent is the extent stated by our sampling statistics. Thus the assurance of equality is greater for large numbers of random assignments than for small. To the extent indicated by the error term for the no-difference hypothesis, this assumption will be wrong occasionally. In Design 4, this means that there will occasionally be an apparently "significant" difference between the pretest scores. Thus, while simple or stratified randomization assures unbiased assignment of experimental subjects to groups, it is a less than perfect way of assuring the initial equivalence of such groups. It is nonetheless the only way of doing so, and the essential way. This statement is made so dogmatically because of a widespread and mistaken preference in educational research over the past 30 years for equation through matching. McCall (1923) and Peters and Van Voorhis (1940) have helped perpetuate this misunderstanding. As will be spelled out in more detail in the discussion of Design 10 and the ex post facto analysis below, matching is no real help when used to overcome initial group differences. This is not to rule out matching as an adjunct to randomization, as when one gains statistical precision by assigning students to matched pairs, and then randomly assigning one member of each pair to the experimental group, the other to the control group. In the statistical literature this is known as "blocking." See particularly the discussions of Cox (1957), Feldt (1958), and Lindquist (1953). But matching as a substitute for randomization is taboo even for the quasi-experimental designs using but two natural intact groups, one experimental, the other control: even in this weak "experiment," there are better ways than matching for attempting to correct for initial mean differences in the two samples.

The data made available by Design 4 make it possible to tell whether *mortality* offers a plausible explanation of the O_1—O_2 gain. Mortality, lost cases, and cases on which only partial data are available, are troublesome to handle, and are commonly swept under the rug. Typically, experiments on teaching methods are spread out over days, weeks, or months. If the pretests and posttests are given in the classrooms from which experimental group and control group are drawn, and if the experimental condition requires attendance at certain sessions, while the control condition does not, then the differential attendance on the three occasions (pretest, treatment, and posttest) produces "mortality" which can introduce subtle sample biases.

If, of those initially designated as experimental group participants, one eliminates those who fail to show up for experimental sessions, then one selectively shrinks the experimental group in a way not comparably done in the control group, biasing the experimental group in the direction of the conscientious and healthy. The preferred mode of treatment, while not usually employed, would seem to be to use all of the selected experimental and control students who completed both pretest and posttest, including those in the experimental group who failed to get the *X*. This procedure obviously attenuates the apparent effect of the *X*, but it avoids the sampling bias. This procedure rests on the assumption that no simpler mortality biases were present; this assumption can be partially checked by examining both the number and the pretest scores of those who were present on pretest but not on posttest. It is possible that some *X*s would affect this drop-out rate rather than change individual scores. Of course, even where drop-out rates are the same, there remains the possibility of complex interactions which would tend to make the character of the drop-outs in the experimental and control groups differ.

The mortality problem can be seen in a greatly exaggerated form in the *invited remedial treatment* study. Here, for example, one sample of poor readers in a high school is invited to participate in voluntary remedial sessions, while an equivalent group are not invited. Of the invited group, perhaps 30 per cent participate. Posttest scores, like pretest scores, come from standard reading achievement tests administered to all in the classrooms. It is unfair to compare the 30 per cent volunteers with the total of the control group, because they represent those most disturbed by their pretest scores, those likely to be most vigorous in self-improvement, etc. But it is impossible to locate their exact counterparts in the control group. While it also seems unfair to the hypothesis of therapeutic effectiveness to compare the total invited group with the total uninvited group, this is an acceptable, if conservative, solution. Note, however, the possibility that the invitation itself, rather than the therapy, causes the effect. In general, the uninvited control group should be made just as aware of its standing on the pretest as is the invited group. Another alternative is to invite all those who need remedial sessions and to assign those who accept into true and placebo remedial treatment groups; but in the present state of the art, any placebo therapy which is plausible enough to look like help to the student is apt to be as good a therapy as is the treatment we are studying. Note, however, the valid implication that experimental tests of the relative efficacy of two therapeutic procedures are much easier to evaluate than the absolute effectiveness of either. The only solution in actual use is that of creating experimental and control groups from among seekers of remedial treatment by manipulating waiting periods (e.g., Rogers & Dymond, 1954). This of course sometimes creates other difficulties, such as an excessive drop-out from the postponed-therapy control group. For a successful and apparently nonreactive use of a lottery to decide on an immediate or next-term remedial reading course, see Reed (1956).

Factors Jeopardizing External Validity

The factors of internal invalidity which have been described so far have been factors which directly affected *O* scores. They have been factors which by themselves could produce changes which might be mistaken for the results of *X*, i.e., factors which, once the control group was added, would produce effects manifested by themselves in the control group and added onto the effects of *X* in the experimental group. In the language of analysis of variance, *history, maturation, testing,* etc., have been described as main effects, and as such have been controlled in Design 4, giving it *internal* validity. The threats to *external* validity, on the other hand, can be called interaction effects, involving *X* and some other variable. They thus represent a

potential specificity of the effects of X to some undesirably limited set of conditions. To anticipate: in Design 4, for all we know, the effects of X observed may be specific to groups warmed up by the pretest. We are logically unable to generalize to the larger unpretested universe about which we would prefer to be able to speak.

In this section we shall discuss several such threats to generalizability, and procedures for reducing them. Thus since there are valid designs avoiding the pretest, and since in many settings (but not necessarily in research on teaching) it is to unpretested groups that one wants to generalize, such designs are preferred on grounds of *external* validity or generalizability. In the area of teaching, the doubts frequently expressed as to the applicability in actual practice of the results of highly artificial experiments are judgments about *external* validity. The introduction of such considerations into the discussion of optimal experimental designs thus strikes a sympathetic note in the practitioner who rightly feels that these considerations have been unduly neglected in the usual formal treatise on experimental methodology. The ensuing discussion will support such views by pointing out numerous ways in which experiments can be made more valid externally, more appropriate bases of generalization to teaching practice, without losing *internal* validity.

But before entering this discussion, a caveat is in order. This caveat introduces some painful problems in the science of induction. The problems are painful because of a recurrent reluctance to accept Hume's truism that *induction or generalization is never fully justified logically*. Whereas the problems of *internal* validity are solvable within the limits of the logic of probability statistics, the problems of external validity are not logically solvable in any neat, conclusive way. Generalization always turns out to involve extrapolation into a realm not represented in one's sample. Such extrapolation is made by *assuming* one knows the relevant laws. Thus, if one has an internally valid Design 4, one has demonstrated the effect only for those specific conditions which the experimental and control group have in common, i.e., only for pretested groups of a specific age, intelligence, socioeconomic status, geographical region, historical moment, orientation of the stars, orientation in the magnetic field, barometric pressure, gamma radiation level, etc.

Logically, we cannot generalize beyond these limits; i.e., we cannot generalize at all. But we do attempt generalization by guessing at laws and checking out some of these generalizations in other equally specific but different conditions. In the course of the history of a science we learn about the "justification" of generalizing by the cumulation of our experience in generalizing, but this is not a logical generalization deducible from the details of the original experiment. Faced by this, we do, in generalizing, make guesses as to yet unproven laws, including some not even explored. Thus, for research on teaching, we are quite willing to assume that orientation in the magnetic field has no effect. But we know from scattered research that pretesting has often had an effect, and therefore we would like to remove it as a limit to our generalization. If we were doing research on iron bars, we would know from experience that an initial weighing has never been found to be reactive, but that orientation in magnetic field, if not systematically controlled, might seriously limit the generalizability of our discoveries. The sources of external invalidity are thus guesses as to general laws in the science of a science: guesses as to what factors lawfully interact with our treatment variables, and, by implication, guesses as to what can be disregarded.

In addition to the specifics, there is a general empirical law which we are assuming, along with all scientists. This is the modern version of Mill's assumption as to the lawfulness of nature. In its modern, weaker version, this can be stated as the assumption of the "stickiness" of nature: we assume that the closer two events are in time, space, and measured value on any or all dimensions, the more they tend to follow the same laws.

While complex interactions and curvilinear relationships are expected to confuse attempts at generalization, they are more to be expected the more the experimental situation differs from the setting to which one wants to generalize. Our call for greater external validity will thus be a call for that maximum similarity of experiments to the conditions of application which is compatible with internal validity.

While stressing this, we should keep in mind that the "successful" sciences such as physics and chemistry made their strides without any attention to representativeness (but with great concern for repeatability by independent researchers). An ivory-tower artificial laboratory science is a valuable achievement even if unrepresentative, and artificiality may often be essential to the analytic separation of variables fundamental to the achievements of many sciences. But certainly, if it does not interfere with internal validity or analysis, external validity is a very important consideration, especially for an applied discipline such as teaching.

Interaction of testing and X. In discussions of experimental design per se, the threat of the pretest to external validity was first presented by Solomon (1949), although the same considerations had earlier led individual experimenters to the use of Design 6, which omits the pretest. Especially in attitude-change studies, where the attitude tests themselves introduce considerable amounts of unusual content (e.g., one rarely sees in cold print as concentrated a dose of hostile statements as is found in the typical prejudice test), it is quite likely that the person's attitudes and his susceptibility to persuasion are changed by a pretest. As a psychologist, one seriously doubts the comparability of one movie audience seeing *Gentlemen's Agreement* (an antiprejudice film) immediately after having taken a 100-item anti-Semitism test with another audience seeing the movie without such a pretest. These doubts extend not only to the main effect of the pretest, but also to its effect upon the response to persuasion. Let us assume that that particular movie was so smoothly done that some persons could enjoy it for its love interest without becoming aware of the social problem it dealt with. Such persons would probably not occur in a pretested group. If a pretest sensitized the audience to the problem, it might, through a focusing of attention, increase the educational effect of the *X*. Conceivably, such an *X* might be effective only for a pretested group.

While such a sensitizing effect is frequently mentioned in anecdotal presentations of the effect, the few published research results show either no effect (e.g., Anderson, 1959; Duncan, et al., 1957; Glock, 1958; Lana, 1959a, 1959b; Lana & King, 1960; Piers, 1955; Sobol, 1959; Zeisel, 1947) or an interaction effect of a dampening order. Thus Solomon (1949) found that giving a pretest reduced the efficacy of experimental spelling training, and Hovland, Lumsdaine, and Sheffield (1949) suggested that a pretest reduced the persuasive effects of movies. This interaction effect is well worth avoiding, even if not as misleading as sensitization (since false positives are more of a problem in our literature than false negatives, owing to the glut of published findings [Campbell, 1959, pp. 168–170]).

The effect of the pretest upon *X* as it restricts external validity is of course a function of the extent to which such repeated measurements are characteristic of the universe to which one wants to generalize. In the area of mass communications, the researcher's interview and attitude-test procedures are quite atypical. But in research on teaching, one is interested in generalizing to a setting in which testing is a regular phenomenon. Especially if the experiment can use regular classroom examinations as *O*s, but probably also if the experimental *O*s are similar to those usually used, no undesirable interaction of *testing* and *X* would be present. Where highly unusual test procedures are used, or where the testing procedure involves deception, perceptual or cognitive restructuring, surprise, stress, etc., designs having unpretested groups remain highly desirable if not essential.

Interaction of selection and X. While Design 4 controls for the effects of selection at the level of explaining away experimental and control group differences, there remains the possibility that the effects validly demonstrated hold only for that unique population from which the experimental and control groups were jointly selected. This possibility becomes more likely as we have more difficulty in getting subjects for our experiment. Consider the implications of an experiment on teaching in which the researcher has been turned down by nine school systems and is finally accepted by a tenth. This tenth almost certainly differs from the other nine, and from the universe of schools to which we would like to generalize, in many specific ways. It is, thus, nonrepresentative. Almost certainly its staff has higher morale, less fear of being inspected, more zeal for improvement than does that of the average school. And the effects we find, while internally valid, might be specific to such schools. To help us judge on these matters, it would seem well for research reports to include statements as to how many and what kind of schools and classes were asked to cooperate but refused, so that the reader can estimate the severity of possible selective biases. Generally speaking, the greater the amount of cooperation involved, the greater the amount of disruption of routine, and the higher our refusal rate, the more opportunity there is for a selection-specificity effect.

Let us specify more closely just what the "*interaction* of selection and *X*" means. If we were to conduct a study within a single volunteered school, using random assignment of subjects to experimental and control groups, we would not be concerned about the "main effect" of the school itself. If both experimental and control group means were merely elevated equally by this, then no harm would be done. If, however, there were characteristics of the school that caused the experimental treatment to be more effective there than it would be in the target population of schools, this could be serious. We want to know that the interaction of school characteristics (probably related to voluntarism) with experimental treatments is negligible. Some experimental variables might be quite sensitive to (interact with) school characteristics; others might not. Such interaction *could* occur between schools with similar mean IQs, or it could be absent when IQ differences were great. We would expect, however, that interactions would be more likely if the schools differed markedly in various characteristics than if they were similar.

Often stringent sampling biases occur because of the inertia of experimenters who do not allow a more representative selection of schools the opportunity to refuse to participate. Thus most research on teaching is done in those schools with the highest percentage of university professors' children enrolled. While sampling representativeness is impossible of perfect achievement and is almost totally neglected in many sciences (in most studies appearing in the *Journal of Experimental Psychology,* for example), it both can and should be emphasized as a desideratum in research on teaching. One way to increase it is to reduce the number of students or classrooms participating from a given school or grade and to increase the number of schools and grades in which the experiment is carried on. It is obvious that we are never going to conduct experiments on samples representatively drawn from all United States classrooms, or all world classrooms. We will learn how far we can generalize an internally valid finding only piece by piece through trial and error of generalization efforts. But these generalization efforts will succeed more often if in the initial experiment we have demonstrated the phenomenon over a wide variety of conditions.

With reference to the pluses and minuses of Table 1, it is obvious that nothing firm can be entered in this column. The column is presented, however, because the requirements of some designs exaggerate or ameliorate this problem. Design 4 in the social-attitudes realm is so demanding of cooperation on the part of respondents or subjects as to end up

with research done only on captive audiences rather than the general citizen of whom one would wish to speak. For such a setting, Design 4 would rate a minus for selection. Yet for research on teaching, our universe of interest is a captive population, and for this, highly representative Design 4s can be done.

Other interactions with X. In parallel fashion, the interaction of *X* with the other factors can be examined as threats to *external* validity. Differential *mortality* would be a product of *X* rather than interactive with it. *Instrumentation* interacting with *X* has been implicitly included in the discussion of *internal* validity, since an instrumentation effect specific to the presence of *X* would counterfeit a true effect of *X* (e.g., where observers make ratings, know the hypothesis, and know which students have received *X*). A threat to external validity is the possibility of the specificity of effects to the specific instruments (tests, observers, meters, etc.) used in the study. If multiple observers or interviewers are used across treatments, such interactions can be studied directly (Stanley, 1961a). *Regression* does not enter as interacting with *X*.

Maturation has implications of a selection-specificity nature: the results may be specific to those of this given age level, fatigue level, etc. The interaction of *history* and *X* would imply that the effect was specific to the historical conditions of the experiment, and while validly observed there, would not be found upon other occasions. The fact that the experiment was done during wartime, or just following an unsuccessful teachers' strike, etc., might produce a responsiveness to *X* not to be found upon other occasions. If we were to produce a sampling model for this problem, we should want the experiment replicated over a random sample of past and future occasions, which is obviously impossible. Furthermore, we share with other sciences the empirical assumption that there are no truly time-dependent laws, that the effects of *history* where found will be due to the specific combinations of stimulus conditions at that time, and thus ultimately will be incorporated under time-independent general laws (Neyman, 1960). ("Expanding universe" cosmologies may seem to require qualification of this statement, but not in ways relevant to this discussion.) Nonetheless, successful replication of research results across times as well as settings increases our confidence in a generalization by making interaction with *history* less likely.

These several factors have not been entered as column headings in Table 1, because they do not provide bases of discrimination among alternative designs.

Reactive arrangements. In the usual psychological experiment, if not in educational research, a most prominent source of unrepresentativeness is the patent artificiality of the experimental setting and the student's knowledge that he is participating in an experiment. For human experimental subjects, a higher-order problem-solving task is generated, in which the procedures and experimental treatment are reacted to not only for their simple stimulus values, but also for their role as clues in divining the experimenter's intent. The play-acting, outguessing, up-for-inspection, I'm-a-guinea-pig, or whatever attitudes so generated are unrepresentative of the school setting, and seem to be qualifiers of the effect of *X*, seriously hampering generalization. Where such *reactive arrangements* are unavoidable, internally valid experiments of this type should by all means be continued. But if they can be avoided, they obviously should be. In stating this, we in part join the typical anti-experimental critic in the school system or the education faculty by endorsing his most frequent protest as to the futility of "all this research." Our more moderate conclusion is not, however, that research should be abandoned for this reason, but rather that it should be improved on this score. Several suggestions follow.

Any aspect of the experimental procedure may produce this *reactive arrangements* effect. The pretesting in itself, apart from its contents, may do so, and part of the *pretest* interaction with *X* may be of this nature, although there are ample grounds to suspect

the content features of the testing process. The process of randomization and assignment to treatments may be of such a nature: consider the effect upon a classroom when (as in Solomon, 1949) a randomly selected half of the pupils in a class are sent to a separate room. This action, plus the presence of the strange "teachers," must certainly create expectations of the unusual, with wonder and active puzzling as to purpose. The presentation of the treatment *X*, if an out-of-ordinary event, could have a similar effect. Presumably, even the posttest in a posttest-only Design 6 could create such attitudes. The more obvious the connection between the experimental treatment and the posttest content, the more likely this effect becomes.

In the area of public opinion change, such reactive arrangements may be very hard to avoid. But in much research on teaching methods there is no need for the students to know that an experiment is going on. (It would be nice to keep the teachers from knowing this, too, in analogy to medicine's double-blind experiment, but this is usually not feasible.) Several features may make such disguise possible. If the *X*s are variants on usual classroom events occurring at plausible periods in the curriculum calendar, then one-third of the battle is won when these treatments occur without special announcement. If the *O*s are similarly embedded as regular examinations, the second requirement is achieved. If the *X*s are communications focused upon individual students, then randomization can be achieved without the physical transportation of randomly equivalent samples to different classrooms, etc.

As a result of such considerations, and as a result of personal observations of experimenters who have published data in spite of having such poor rapport that their findings were quite misleading, the present authors are gradually coming to the view that experimentation within schools must be conducted by regular staff of the schools concerned, whenever possible, especially when findings are to be generalized to other classroom situations.

At present, there seem to be two main types of "experimentation" going on within schools: (1) research "imposed" upon the school by an outsider, who has his own ax to grind and whose goal is not immediate action (change) by the school; and (2) the so-called "action" researcher, who tries to get teachers themselves to be "experimenters," using that word quite loosely. The first researcher gets results that may be rigorous but not applicable. The latter gets results that may be highly applicable but probably not "true" because of extreme lack of rigor in the research. An alternative model is for the ideas for classroom research to originate with teachers and other school personnel, with designs to test these ideas worked out cooperatively with specialists in research methodology, and then for the bulk of the experimentation to be carried out by the idea-producers themselves. The appropriate statistical analyses could be done by the research methodologist and the results fed back to the group via a trained intermediary (supervisor, director of research in the school system, etc.) who has served as intermediary all along. Results should then be relevant and "correct." How to get *basic* research going under such a pattern is largely an unsolved problem, but studies could become less and less ad hoc and more and more theory-oriented under a competent intermediary.

While there is no intent in this chapter to survey either good or bad examples in the literature, a recent study by Page (1958) shows such an excellent utilization of these features (avoiding reactive arrangements, achieving sampling representativeness, and avoiding testing-*X* interactions) that it is cited here as a concrete illustration of optimal practice. His study shows that brief written comments upon returned objective examinations improve subsequent objective examination performance. This finding was demonstrated across 74 teachers, 12 school systems, 6 grades (7–12), 5 performance levels (A, B, C, D, F), and a wide variety of subjects, with almost no evidence of inter-

action effects. The teachers and classes were randomly selected. The earliest regular objective examination in each class was used as the pretest. By rolling a specially marked die the teacher assigned students to treatment groups, and correspondingly put written comments on the paper or did not. The next normally scheduled objective test in the class became the posttest. As far as could be told, not one of the 2,139 students was aware of experimentation. Few instructional procedures lend themselves to this inconspicuous randomization, since usually the oral communication involved is addressed to a whole class, rather than to individuals. (Written communications do allow for randomized treatment, although student detection of varied treatments is a problem.) Yet, holding these ideals in mind, research workers can make experiments nonreactive in many more features than they are at present.

Through regular classroom examinations or through tests presented as regular examinations and similar in content, and through alternative teaching procedures presented without announcement or apology in the regular teaching process, these two sources of reactive arrangements can probably be avoided in most instances. Inconspicuous randomization may be the more chronic problem. Sometimes, in large high schools or colleges, where students sign up for popular courses at given hours and are then assigned arbitrarily to multiple simultaneous sections, randomly equivalent sections might be achieved through control of the assignment process. (See Siegel & Siegel, 1957, for an opportunistic use of a natural randomization process.) However, because of unique intragroup histories, such initially equivalent sections become increasingly nonequivalent with the passage of long periods of time.

The all-purpose solution to this problem is to move the randomization to the classroom as a unit, and to construct experimental and control groups each constituted of numerous classrooms randomly assigned (see Lindquist, 1940, 1953). Usually, but not essentially, the classrooms would be classified for analysis on the basis of such factors as school, teacher (where teachers have several classes), subject, time of day, mean intelligence level, etc.; from these, various experimental-treatment groups would be assigned by a random process. There have been a few such studies, but soon they ought to become standard. Note that the appropriate test of significance is *not* the pooling of all students as though the students had been assigned at random. The details will be discussed in the subsequent section.

Tests of Significance for Design 4

Good experimental design is separable from the use of statistical tests of significance. It is the art of achieving interpretable comparisons and as such would be required even if the end product were to be graphed percentages, parallel prose case studies, photographs of groups in action, etc. In all such cases, the interpretability of the "results" depends upon control over the factors we have been describing. If the comparison is interpretable, then statistical tests of significance come in for the decision as to whether or not the obtained difference rises above the fluctuations to be expected in cases of no true difference for samples of that size. Use of significance tests presumes but does not prove or supply the comparability of the comparison groups or the interpretability of the difference found. We would thus be happy to teach experimental design upon the grounds of common sense and nonmathematical considerations. We hope that the bulk of this chapter is accessible to students of education still lacking in statistical training. Nevertheless, the issue of statistical procedures is intimately tied to experimental design, and we therefore offer these segregated comments on the topic. (Also see Green & Tukey, 1960; Kaiser, 1960; Nunnally, 1960; and Rozeboom, 1960.)

A wrong statistic in common use. Even though Design 4 is the standard and most widely used design, the tests of significance

used with it are often wrong, incomplete, or inappropriate. In applying the common "critical ratio" or *t* test to this standard experimental design, many researchers have computed two *t*s, one for the pretest-posttest difference in the experimental group, one for the pretest-posttest gain in the control group. If the former be "statistically significant" and the latter "not," then they have concluded that the *X* had an effect, without any direct statistical comparison of the experimental and control groups. Often the conditions have been such that, had a more appropriate test been made, the difference would not have been significant (as in the case where the significance values are borderline, with the control group showing a gain almost reaching significance). Windle (1954) and Cantor (1956) have shown how frequent this error is.

Use of gain scores and covariance. The most widely used acceptable test is to compute for each group pretest-posttest gain scores and to compute a *t* between experimental and control groups on these gain scores. Randomized "blocking" or "leveling" on pretest scores and the analysis of covariance with pretest scores as the covariate are usually preferable to simple gain-score comparisons. Since the great bulk of educational experiments show no significant difference, and hence are frequently not reported, the use of this more precise analysis would seem highly desirable. Considering the labor of conducting an experiment, the labor of doing the proper analysis is relatively trivial. Standard treatments of Fisher-type analyses may be consulted for details. (Also see Cox, 1957, 1958; Feldt, 1958; and Lindquist, 1953.)

Statistics for random assignment of intact classrooms to treatments. The usual statistics are appropriate only where individual students have been assigned at random to treatments. Where intact classes have been assigned to treatments, the above formulas would provide too small an error term because the randomization procedure obviously has been more "lumpy" and fewer chance events have been employed. Lindquist (1953, pp. 172–189) has provided the rationale and formulas for a correct analysis. Essentially, the class means are used as the basic observations, and treatment effects are tested against variations in these means. A covariance analysis would use pretest means as the covariate.

Statistics for internal validity. The above points were introduced to convey the statistical orthodoxy relevant to experimental design. The point to follow represents an effort to expand or correct that orthodoxy. It extends an implication of the distinction between *external* and *internal validity* over into the realm of sampling statistics. The statistics discussed above all imply sampling from an infinitely large universe, a sampling more appropriate to a public opinion survey than to the usual laboratory experiment. In the rare case of a study like Page's (1958), there is an actual sampling from a large predesignated universe, which makes the usual formulas appropriate. At the other extreme is the laboratory experiment represented in the *Journal of Experimental Psychology,* for example, in which *internal validity* has been the only consideration, and in which *all* members of a unique small universe have been exhaustively assigned to the treatment groups. There is in such experiments a great emphasis upon randomization, but not for the purpose of securing representativeness for some larger population. Instead, the randomization is solely for the purpose of equating experimental and control groups or the several treatment groups. The randomization is thus within a very small finite population which is in fact the sum of the experimental plus control groups.

This extreme position on the sampling universe is justified when describing laboratory procedures of this type: volunteers are called for, with or without promises of rewards in terms of money, personality scores, course credit points, or completion of an obligatory requirement which they will have to meet sometime during the term anyway. As volunteers come in, they are randomly assigned to treatments. When some fixed

number of subjects has been reached, the experiment is stopped. There has not even been a random selection from within a much larger list of volunteers. Early volunteers are a biased sample, and the total universe "sampled" changes from day to day as the experiment goes on, as more pressure is required to recruit volunteers, etc. At some point the procedure is stopped, all designatable members of the universe having been used in one or another treatment group. Note that the sampling biases implied do not in the least jeopardize the random equivalence of the treatment groups, but rather only their "representativeness."

Or consider a more conscientious scientist, who randomly draws 100 names from his lecture class of 250 persons, contacting them by phone or mail, and then as they meet appointments assigns them randomly to treatment groups. Of course, some 20 of them cannot conveniently be fitted into the laboratory time schedule, or are ill, etc., so a redefinition of the universe has taken place implicitly. And even if he doggedly gets all 100, from the point of view of representativeness, what he has gained is the ability to generalize with statistical confidence to the 1961 class of Educational Psychology A at State Teachers. This new universe, while larger, is not intrinsically of scientific interest. Its bounds are not the bounds specified by any scientific theory. The important interests in generalization will have to be explored by the sampling of other experiments elsewhere. Of course, since his students are less select, there is more external validity, but not enough gain to be judged worth it by the great bulk of experimental psychologists.

In general, it is obvious that the dominant purpose of randomization in laboratory experiments is internal validity, not external. Pursuant to this, more appropriate and smaller error terms based upon small finite universes should be employed. Following Kempthorne (1955) and Wilk and Kempthorne (1956), we note that the appropriate model is urn randomization, rather than sampling from a universe. Thus there is available a more appropriate, more precise, nonparametric test, in which one takes the obtained experimental and control group scores and repeatedly assigns them at random to two "urns," generating empirically (or mathematically) a distribution of mean differences arising wholly from random assignment of these particular scores. This distribution is the criterion with which the obtained mean difference should be compared. When "plot-treatment interaction" (heterogeneity of true effects among subjects) is present, this distribution will have less variability than the corresponding distribution assumed in the usual t test.

These comments are not expected to modify greatly the actual practice of applying tests of significance in research on teaching. The exact solutions are very tedious, and usually inaccessible. Urn randomization, for example, ordinarily requires access to high-speed computers. The direction of error is known: using the traditional statistics is too conservative, too inclined to say "no effect shown." If we judge our publications to be overloaded with "false-positives," i.e., claims for effects that won't hold up upon cross-validation (this is certainly the case for experimental and social psychology, if not as yet for research on teaching), this error is in the preferred direction—if error there must be. Possible underestimation of significance is greatest when there are only two experimental conditions and all available subjects are used (Wilk & Kempthorne, 1955, p. 1154).

5. The Solomon Four-Group Design

While Design 4 is more used, Design 5, the Solomon (1949) Four-Group Design, deservedly has higher prestige and represents the first explicit consideration of *external validity* factors. The design is as follows:

R	O_1	X	O_2
R	O_3		O_4
R		X	O_5
R			O_6

By paralleling the Design 4 elements (O_1 through O_4) with experimental and control groups lacking the pretest, both the main effects of *testing* and the interaction of *testing* and X are determinable. In this way, not only is generalizability increased, but in addition, the effect of X is replicated in four different fashions: $O_2 > O_1$, $O_2 > O_4$, $O_5 > O_6$, and $O_5 > O_3$. The actual instabilities of experimentation are such that if these comparisons are in agreement, the strength of the inference is greatly increased. Another indirect contribution to the generalizability of experimental findings is also made, in that through experience with Design 5 in any given research area one learns the general likelihood of testing-by-X interactions, and thus is better able to interpret past and future Design 4s. In a similar way, one can note (by comparison of O_6 with O_1 and O_3) a combined effect of maturation and history.

Statistical Tests for Design 5

There is no singular statistical procedure which makes use of all six sets of observations simultaneously. The asymmetries of the design rule out the analysis of variance of gain scores. (Solomon's suggestions concerning these are judged unacceptable.) Disregarding the pretests, except as another "treatment" coordinate with X, one can treat the posttest scores with a simple 2 × 2 analysis of variance design:

	No X	X
Pretested	O_4	O_2
Unpretested	O_6	O_5

From the column means, one estimates the main effect of X, from row means, the main effect of pretesting, and from cell means, the interaction of testing with X. If the main and interactive effects of pretesting are negligible, it may be desirable to perform an analysis of covariance of O_4 versus O_2, pretest scores being the covariate.

6. The Posttest-Only Control Group Design

While the pretest is a concept deeply embedded in the thinking of research workers in education and psychology, it is not actually essential to true experimental designs. For psychological reasons it is difficult to give up "knowing for sure" that the experimental and control groups were "equal" before the differential experimental treatment. Nonetheless, the most adequate all-purpose assurance of lack of initial biases between groups is randomization. Within the limits of confidence stated by the tests of significance, randomization can suffice without the pretest. Actually, almost all of the agricultural experiments in the Fisher (1925, 1935) tradition are without pretest. Furthermore, in educational research, particularly in the primary grades, we must frequently experiment with methods for the initial introduction of entirely new subject matter, for which pretests in the ordinary sense are impossible, just as pretests on believed guilt or innocence would be inappropriate in a study of the effects of lawyers' briefs upon a jury. Design 6 fills this need, and in addition is appropriate to all of the settings in which Designs 4 or 5 might be used, i.e., designs where true randomization is possible. Its form is as follows:

R	X	O_1
R		O_2

While this design was used as long ago as the 1920's, it has not been recommended in most methodological texts in education. This has been due in part to a confusion of it with Design 3, and due in part to distrust of randomization as equation. The design can be considered as the two last groups of the Solomon Four-Group Design, and it can be seen that it controls for testing as main effect and interaction, but unlike Design 5 it does not measure them. However, such measurement is tangential to the central question of whether or not X did have an effect. Thus,

while Design 5 is to be preferred to Design 6 for reasons given above, the extra gains from Design 5 may not be worth the more than double effort. Similarly, Design 6 is usually to be preferred to Design 4, unless there is some question as to the genuine randomness of the assignment. Design 6 is greatly underused in educational and psychological research.

However, in the repeated-testing setting of much educational research, if appropriate antecedent variates are available, they should certainly be used for blocking or leveling, or as covariates. This recommendation is made for two reasons: first, the statistical tests available for Design 4 are more powerful than those available for Design 6. While the greater effort of Design 4 outweighs this gain for most research settings, it would not do so where suitable antecedent scores were automatically available. Second, the availability of pretest scores makes possible examination of the interaction of X and pretest ability level, thus exploring the generalizability of the finding more thoroughly. Something similar can be done for Design 6, using other available measures in lieu of pretests, but these considerations, coupled with the fact that for educational research frequent testing is characteristic of the universe to which one wants to generalize, may reverse the case for generally preferring Design 6 over Design 4. Note also that for any substantial mortality between R and the posttest, the pretest data of Design 4 offer more opportunity to rule out the hypothesis of differential mortality between experimental and control groups.

Even so, many problems exist for which pretests are unavailable, inconvenient, or likely to be reactive, and for such purposes the legitimacy of Design 6 still needs emphasis in many quarters. In addition to studies of the mode of teaching novel subject materials, a large class of instances remains in which (1) the X and posttest O can be delivered to students or groups as a single natural package, and (2) a pretest would be awkward. Such settings frequently occur in research on testing procedures themselves, as in studies of different instructions, different answer-sheet formats, etc. Studies of persuasive appeals for volunteering, etc., are similar. Where student anonymity must be kept, Design 6 is usually the most convenient. In such cases, randomization is handled in the mixed ordering of materials for distribution.

The Statistics for Design 6

The simplest form would be the t test. Design 6 is perhaps the only setting for which this test is optimal. However, covariance analysis and blocking on "subject variables" (Underwood, 1957b) such as prior grades, test scores, parental occupation, etc., can be used, thus providing an increase in the power of the significance test very similar to that provided by a pretest. Identicalness of pretest and posttest is not essential. Often these will be different forms of "the same" test and thus less identical than a repetition of the pretest. The gain in precision obtained corresponds directly to the degree of covariance, and while this is usually higher for alternate forms of "the same" test than for "different" tests, it is a matter of degree, and something as reliable and factorially complex as a grade-point average might turn out to be superior to a short "pretest." Note that a grade-point average is not usually desirable as a posttest measure, however, because of its probable insensitivity to X compared with a measure more specifically appropriate in content and timing. Whether such a pseudo pretest design should be classified as Design 6 or Design 4 is of little moment. It would have the advantages of Design 6 in avoiding an experimenter-introduced pretest session, and in avoiding the "giveaway" repetition of identical or highly similar unusual content (as in attitude change studies). It is for such reasons that the entry for Design 6 under "reactive arrangements" should be slightly more positive than that for Designs 4 and 5. The case for this differential is, of course, much stronger for the social sciences in gen-

eral than for research on educational instruction.

Factorial Designs

On the conceptual base of the three preceding designs, but particularly of Designs 4 and 6, the complex elaborations typical of the Fisher factorial designs can be extended by adding other groups with other Xs. In a typical single-classification criterion or "one-way" analysis of variance we would have several "levels" of the treatment, e.g., X_1, X_2, X_3, etc., with perhaps still an X_0 (no-X) group. If the control group be regarded as one of the treatments, then for Designs 4 and 6 there would be one group for each treatment. For Design 5 there would be two groups (one pretested, one not) for each treatment, and a two-classification ("two-way") analysis of variance could still be performed. We are not aware that more-than-two-level designs of the Design 5 type have been done. Usually, if one were concerned about the pretest interaction, Design 6 would be employed because of the large number of groups otherwise required. Very frequently, two or more treatment variables, each at several "levels," will be employed, giving a series of groups that could be designated X_{a1} X_{b1}, X_{a1} X_{b2}, X_{a1} X_{b3}, ..., X_{a2} X_{b1}, etc.

Such elaborations, complicated by efforts to economize through eliminating some of the possible permutations of X_a by X_b, have produced some of the traumatizing mysteries of factorial design (randomized blocks, split plots, Greco-Latin squares, fractional replication, confounding, etc.) which have created such a gulf between advanced and traditional research methodologies in education. We hope that this chapter helps bridge this gulf through continuity with traditional methodology and the common-sense considerations which the student brings with him. It is also felt that a great deal of what needs to be taught about experimental design can best be understood when presented in the form of two-treatment designs, without interference from other complexities. Yet a full presentation of the problems of traditional usage will generate a comprehension of the need for and place of the modern approaches. Already, in searching for the most efficient way of summarizing the widely accepted old-fashioned Design 4, we were introduced to a need for covariance analysis, which has been almost unused in this setting. And in Design 5, with a two-treatment problem elaborated only to obtain needed controls, we moved away from critical ratios or t tests into the related analysis-of-variance statistics.

The details of statistical analyses for factorial designs cannot be taught or even illustrated in this chapter. Elementary aspects of these methods are presented for educational researchers by Edwards (1960), Ferguson (1959), Johnson and Jackson (1959), and Lindquist (1953). It is hoped, however, that the ensuing paragraphs may convey some understanding of certain alternatives and complexities particularly relevant for the design issues discussed in this chapter. The complexities to be discussed do not include the common reasons for using Latin squares and many other incomplete designs where knowledge concerning certain interactions is sacrificed merely for reasons of cost. (But the use of Latin squares as a substitute for control groups where randomization is not possible will be discussed as quasi-experimental Design 11 below.) The reason for the decision to omit such incomplete designs is that detailed knowledge of interactions is highly relevant to the external validity problem, particularly in a science which has experienced trouble in replicating one researcher's findings in another setting (see Wilk & Kempthorne, 1957). The concepts which we seek to convey in this section are interaction, nested versus crossed classifications, and finite, fixed, random, and mixed factorial models.

Interaction

We have already used this concept in contexts where it was hoped the untrained

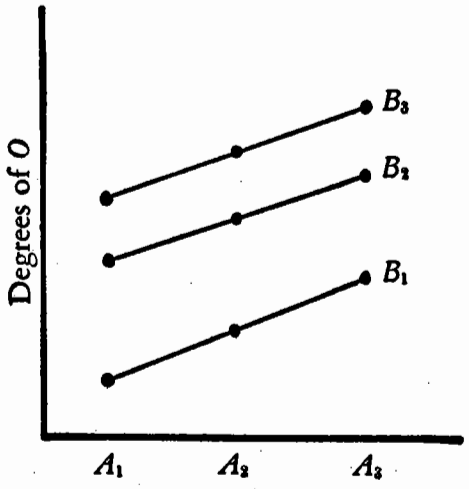

Fig. 2*a*.

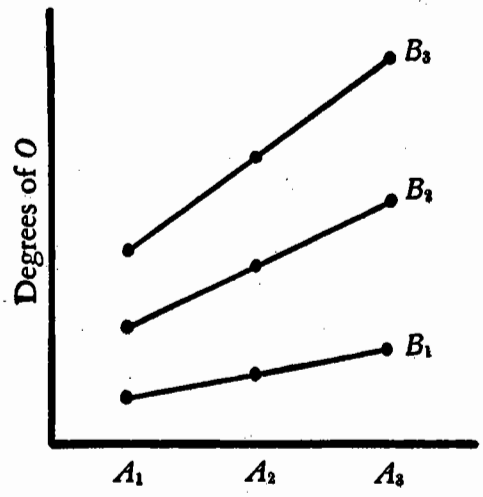

Fig. 2*b*.

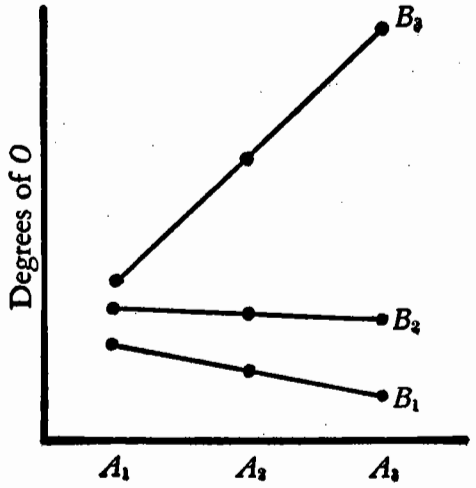

Fig. 2*c*.

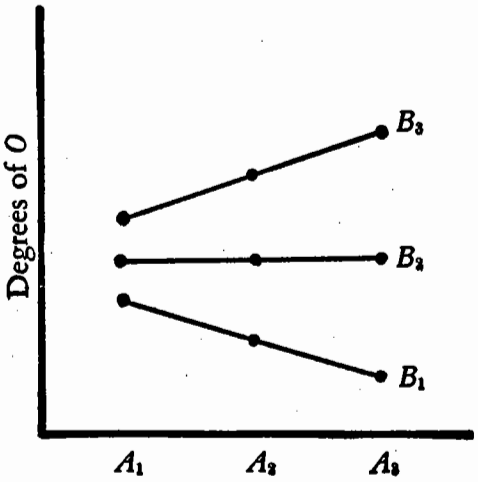

Fig. 2*d*.

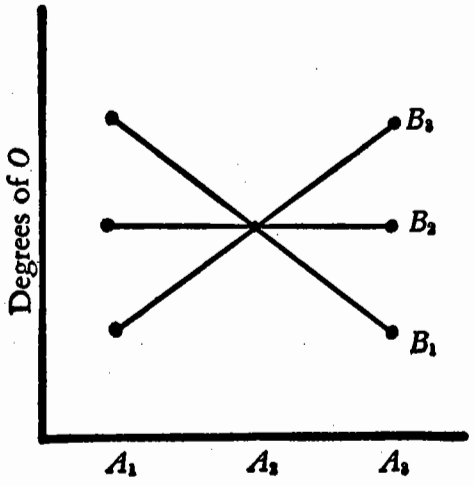

Fig. 2*e*.

Fig. 2. Some Possible Outcomes of a 3 x 3 Factorial Design.

reader would find it comprehensible. As before, our emphasis here is upon the implications for generalizability. Let us consider in graphic form, in Fig. 2, five possible outcomes of a design having three levels each of X_a and X_b, to be called here A and B. (Since three dimensions [A, B, and O] are to be graphed in two dimensions, there are several alternative presentations, only one of which is used here.) In Fig. 2*a* there is a significant main effect for both A and B, but no interaction. (There is, of course, a summation of effects—A_3, B_3 being strongest—but no interaction, as the effects are additive.) In all of the others, there are significant interactions in addition to, or instead of, the main effects of A and B. That is, the law as to the effect of A changes depending upon the specific value of B. In this sense, interaction effects are specificity-of-effect rules and are thus relevant to generalization efforts. The interaction effect in 2*d* is most clearly of this order. Here A does not have a main effect (i.e., if one averages the values of all three Bs for each A, a horizontal line results). But when B is held at level 1, increases in A have a decremental effect, whereas when B is held at level 3, A has an incremental effect. Note that had the experimenter varied A only and held B constant at level 1, the results, while internally valid, would have led to erroneous generalizations for B_2 and B_3. The multiple-factorial feature of the design has thus led to valuable explorations of the generalizability or external validity of any summary statement about the main effect of A. Limitations upon generalizability, or specificity of effects, appear in the statistical analysis as significant interactions.

Figure 2*e* represents a still more extreme form of interaction, in which neither A nor B has any main effect (no general rules emerge as to which level of either is better) but in which the interactions are strong and definite. Consider a hypothetical outcome of this sort. Let us suppose that three types of teachers are all, in general, equally effective (e.g., the spontaneous extemporizers, the conscientious preparers, and the close supervisors of student work). Similarly, three teaching methods in general turn out to be equally effective (e.g., group discussion, formal lecture, and tutorial). In such a case, even in the absence of "main effects" for either teacher-type or teaching method, teaching methods could plausibly interact strongly with types, the spontaneous extemporizer doing best with group discussion and poorest with tutorial, and the close supervisor doing best with tutorial and poorest with group discussion methods.

From this point of view, we should want to distinguish between the kinds of significant interactions found. Perhaps some such concept as "monotonic interactions" might do. Note that in 2*b*, as in 2*a*, there is a main effect of both A and B, and that A has the same directional effect in every separate panel of B values. Thus we feel much more confident in generalizing the expectation of increase in O with increments in A to novel settings than we do in case 2*c*, which likewise might have significant main effects for A and B, and likewise a significant A-B interaction. We might, in fact, be nearly as confident of the generality of A's main effect in a case like 2*b* as in the interaction-free 2*a*. Certainly, in interpreting effects for generalization purposes, we should plot them and examine them in detail. Some "monotonic" or single-directional interactions produce little or no specificity limitations. (See Lubin, 1961, for an extended discussion of this problem.)

Nested Classifications

In the illustrations which we have given up to this point, all of the classification criteria (the As and the Bs) have "crossed" all other classification criteria. That is, all levels of A have occurred with all levels of B. Analysis of variance is not limited to this situation, however.

So far, we have used, as illustrations, classification criteria which were "experimental treatments." Other types of classification criteria, such as sex and age of pupils, could be

introduced into many experiments as fully crossed classifications. But to introduce the most usual uses of "nested" classifications, we must present the possibility of less obvious classification criteria. One of these is "teachers." Operating at the fully crossed level, one might do an experiment in a high school in which each of 10 teachers used each of two methods of teaching a given subject, to different experimental classes. In this case, teachers would be a fully crossed classification criterion, each teacher being a different "level." The "main effect" of "teachers" would be evidence that some teachers are better than others no matter which method they are using. (Students or classes must have been assigned at random; otherwise teacher idiosyncrasies and selection differences are confounded.) A significant interaction between teachers and methods would mean that the method which worked better depended upon the particular teacher being considered.

Suppose now, in following up such an interaction, one were interested in whether or not a given technique was, in general, better for men teachers than women. If we now divide our 10 teachers into 5 men and 5 women, a "nesting" classification occurs in that the teacher classification, while still useful, does not cross sexes; i.e., the same teacher does not appear in both sexes, while each teacher and each sex do cross methods. This nesting requires a somewhat different analysis than does the case where all classifications cross all others. (For illustrative analyses, see Green and Tukey, 1960, and Stanley, 1961a.) In addition, certain interactions of the nested variables are ruled out. Thus the teachers-sex and teachers-sex-method interactions are not computable, and, indeed, make no sense conceptually.

"Teachers" might also become a nested classification if the above experiment were extended into several schools, so that schools became a classification criterion (for which the main effects might reflect learning-rate differences on the part of pupils of the several schools). In such a case, teachers would usually be "nested" within schools, in that one teacher would usually teach classes within just one school. While in this instance a teacher-school interaction is conceivable, one could not be computed unless all teachers taught in both schools, in which case teachers and schools would be "crossed" rather than "nested."

Pupils, or subjects in an experiment, can also be treated as a classification criterion. In a fully crossed usage each pupil gets each treatment, but in many cases the pupil enters into several treatments, but not all; i.e., nesting occurs. One frequent instance is the study of trial-by-trial data in learning. In this case, one might have learning curves for each pupil, with pupils split between two methods of learning. Pupils would cross trials but not methods. Trial-method interactions and pupil-trial interactions could be studied, but not pupil-method interactions. Similarly, if pupils are classified by sex, nesting occurs.

Most variables of interest in educational experimentation can cross other variables and need not be nested. Notable exceptions, in addition to those mentioned above, are chronological age, mental age, school grade (first, second, etc.), and socioeconomic level. The perceptive reader may have noted that independent variables, or classification criteria, are of several sorts: (1) manipulated variables, such as teaching method, assignable at will by the experimenter; (2) potentially manipulable aspects, such as school subject studied, that the experimenter might assign in some random way to the pupils he is using, but rarely does; (3) relatively fixed aspects of the environment, such as community or school or socioeconomic level, not under the direct control of the experimenter but serving as explicit bases for stratification in the experiment; (4) "organismic" characteristics of pupils, such as age, height, weight, and sex; and (5) response characteristics of pupils, such as scores on various tests. Usually the manipulated independent variables of Class 1 are of primary interest, while the unmanipulated independent variables of Classes 3, 4, and sometimes 5, serve to in-

crease precision and reveal how generalizable the effects of manipulated variables are. The variables of Class 5 usually appear as covariates or dependent variates. Another way to look at independent variables is to consider them as intrinsically ordered (school grade, socioeconomic level, height, trials, etc.) or unordered (teaching method, school subject, teacher, sex, etc.). Effects of ordered variables may often be analyzed further to see whether the trend is linear, quadratic, cubic, or higher (Grant, 1956; Myers, 1959).

Finite, Random, Fixed, and Mixed Models

Recently, stimulated by Tukey's unpublished manuscript of 1949, several mathematical statisticians have devised "finite" models for the analysis of variance that apply to the sampling of "levels" of experimental factors (independent variables) the principles well worked out previously for sampling from finite populations. Scheffé (1956) provided a historical survey of this clarifying development. Expected mean squares, which help determine appropriate "error terms," are available (Stanley, 1956) for the completely randomized three-classification factorial design. Finite models are particularly useful because they may be generalized readily to situations where one or more of the factors are random or fixed. A simple explanation of these extensions was given by Ferguson (1959).

Rather than present formulas, we shall use a verbal illustration to show how finite, random, and fixed selection of levels of a factor differ. Suppose that "teachers" constitute one of several bases for classification (i.e., independent variables) in an experiment. If 50 teachers are available, we might draw 5 of these *randomly* and use them in the study. Then a factor-sampling coefficient (1—5/50), or 0.9, would appear in some of our formulas. If all 50 teachers were employed, then teachers would be a "fixed" effect and the coefficient would become (1—50/50) = 0. If, on the other hand, a virtually infinite population of teachers existed, 50 selected randomly from this population would be an infinitesimal percentage, so the coefficient would approach 1 for each "random" effect. The above coefficients modify the formulas for expected mean squares, and hence for "error" terms. Further details appear in Brownlee (1960), Cornfield and Tukey (1956), Ferguson (1959), Wilk and Kempthorne (1956), and Winer (1962).

Other Dimensions of Extension

Before leaving the "true" experiments for the quasi-experimental designs, we wish to explore some other extensions from this simple core, extensions appropriate to all of the designs to be discussed.

Testing for Effects Extended in Time

In the area of persuasion, an area somewhat akin to that of educating and teaching, Hovland and his associates have repeatedly found that long-term effects are not only quantitatively different, but also qualitatively different. Long-range effects are greater than immediate effects for general attitudes, although weaker for specific attitudes (Hovland, Lumsdaine, & Sheffield, 1949). A discredited speaker has no persuasive effect immediately, but may have a significant effect a month later, unless listeners are reminded of the source (Hovland, Janis, & Kelley, 1953). Such findings warn us against pinning all of our experimental evaluation of teaching methods on immediate posttests or measures at any single point in time. In spite of the immensely greater problems of execution (and the inconvenience to the nine-month schedule for a Ph.D. dissertation), we can but recommend that posttest periods such as one month, six months, and one year be included in research planning.

When the posttest measures are grades and examination scores that are going to be collected anyway, such a study is nothing but a

bookkeeping (and mortality) problem. But where the *O*s are introduced by the experimenter, most writers feel that repeated posttest measures on the same students would be more misleading than the pretest would be. This has certainly been found to be true in research on memory (e.g., Underwood, 1957a). While Hovland's group has typically used a pretest (Design 4), they have set up separate experimental and control groups for each time delay for the posttest, e.g.:

R	*O*	*X*	*O*	
R	*O*		*O*	
R	*O*	*X*		*O*
R	*O*			*O*

A similar duplication of groups would be required for Designs 5 or 6. Note that this design lacks perfect control for its purpose of comparing differences in effect as a function of elapsed time, in that the differences could also be due to an interaction between *X* and the specific historical events occurring between the short-term posttest and the long-term one. Full control of this possibility leads to still more elaborate designs. In view of the great expense of such studies except where the *O*s are secured routinely, it would seem incumbent upon those making studies using institutionalized *O*s repeatedly available to make use of the special advantages of their settings by following up the effects over many points in time.

Generalizing to Other Xs: Variability in the Execution of X

The goal of science includes not only generalization to other populations and times but also to other nonidentical representations of the treatment, i.e., other representations which theoretically should be the same, but which are not identical in theoretically irrelevant specifics. This goal is contrary to an often felt extension of the demand for experimental control which leads to the desire for an *exact* replication of the *X* on each repetition. Thus, in studying the effect of an emotional versus a rational appeal, one might have the same speaker give all appeals to each type of group or, more extremely, record the talks so that all audiences of a given treatment heard "exactly the same" message. This might seem better than having several persons give each appeal just once, since in the latter case we "would not know exactly" what experimental stimulus each session got. But the reverse is actually the case, if by "know" we mean the ability to pick the proper abstract classification for the treatment and to convey the information effectively to new users. With the taped interview we have repeated each time many specific irrelevant features; for all we know, these details, not the intended features, created the effect. If, however, we have many independent exemplifications, the specific irrelevancies are not apt to be repeated each time, and our interpretation of the source of the effects is thus more apt to be correct.

For example, consider the Guetzkow, Kelly, and McKeachie (1954) comparison of recitation and discussion methods in teaching. Our "knowledge" of what the experimental treatments were, in the sense of being able to draw recommendations for other teachers, is better *because* eight teachers were used, each interpreting each method in his own way, than if only one teacher had been used, or than if the eight had memorized common details not included in the abstract description of the procedures under comparison. (This emphasis upon heterogeneous execution of *X* should if possible be accompanied, as in Guetzkow, et al., 1954, by having each treatment executed by each of the experimental teachers, so that no specific irrelevancies are confounded with a specific treatment. To estimate the significance of teacher-method interaction when intact classes have been employed, each teacher should execute each method twice.)

In a more obvious illustration, a study of the effect of sex of the teacher upon beginning instruction in arithmetic should use numerous examples of each sex, not just one

of each. While this is an obvious precaution, it has not always been followed, as Hammond (1954) has pointed out. The problem is an aspect of Brunswik's (1956) emphasis upon representative design. Underwood (1957b, pp. 281–287) has on similar grounds argued against the exact standardization or the exact replication of apparatus from one study to another, in a fashion not incompatible with his vigorous operationalism.

Generalizing to Other Xs: Sequential Refinement of X and Novel Control Groups

The actual *X* in any experiment is a complex package of what will eventually be conceptualized as several variables. Once a strong and clear-cut effect has been noted, the course of science consists of further experiments which refine the *X*, teasing out those aspects which are most essential to the effect. This refinement can occur through more specifically defined and represented treatments, or through developing novel control groups, which come to match the experimental group on more and more features of the treatment, reducing the differences to more specific features of the original complex *X*. The placebo control group and the sham-operation control group in medical research illustrate this. The prior experiments demonstrated an internally valid effect, which, however, could have been due to the patient's knowledge that he was being treated or to surgical shock, rather than to the specific details of the drug or to the removal of the brain tissue—hence the introduction of the special controls against these possibilities. The process of generalizing to other *X*s is an exploratory, theory-guided trial and error of extrapolations, in the process of which such refinement of *X*s is apt to play an important part.

Generalizing to Other Os

Just as a given *X* carries with it a baggage of theoretically irrelevant specificities which may turn out to cause the effect, so any given *O*, any given measuring instrument, is a complex in which the relevant content is necessarily embedded in a specific instrumental setting, the details of which are tangential to the theoretical purpose. Thus, when we use IBM pencils and machine-scored answer-sheets, it is usually for reasons of convenience and not because we wish to include in our scores variance due to clerical skills, test-form familiarity, ability to follow instructions, etc. Likewise, our examination of specific subject-matter competence by way of essay tests must be made through the vehicles of penmanship and vocabulary usage and hence must contain variance due to these sources often irrelevant to our purposes. Given this inherent complexity of any *O*, we are faced with a problem when we wish to generalize to other potential *O*s. To which aspect of our experimental *O* was this internally valid effect due? Since the goals of teaching are not solely those of preparing people for future essay and objective examinations, this problem of external validity or generalizability is one which must be continually borne in mind.

Again, conceptually, the solution is not to hope piously for "pure" measures with no irrelevant complexities, but rather to use multiple measures in which the specific vehicles, the specific irrelevant details, are as different as possible, while the common content of our concern is present in each. For *O*s, more of this can be done within a single experiment than for *X*s, for it is usually possible to get many measures of effect (i.e., dependent variables) in one experiment. In the study by Guetzkow, Kelly, and McKeachie (1954), effects were noted not only on course examinations and on special attitude tests introduced for this purpose, but also on such subsequent behaviors as choice of major and enrollment in advanced courses in the same topic. (These behaviors proved to be just as sensitive to treatment differences as were the test measures.) *Multiple Os should be an orthodox requirement in any study of teaching methods.* At the simplest

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.)