D.T. Campbell, J.C. Stanley

of observation,
cal, are far from
s retained is de-
by pure chance.
his point as the
elevance of the
process. Experi-
self viewed as a
contradictory to
ather a refining
the probably val-
practice. Advo-
nce of education
ng a position in-
wisdom.
uspicion that the
lutionary scheme
ecifically human
n Doe, when con-
ty for deciding
l textbook or re-
longer, probably
anty knowledge.
s sheer efficiency
ter his mind. The
o ways: keep the
l as or better than
the revised book
unrevised edition.
ng in two ways:
e new one is bet-
ok when it is no

ls might be esti-
the two erroneous
d energy-expendi-
principal in com-
ents, and school-
to teachers, pupils,
oorer instruction.
oney, energy, con-
nd personal threat
he probability that
he probability that
ected. If the prin-
without suitable
ing Cost 3 (poorer
to overemphasize
s seem stacked in

favor of a conservative approach—that is, retaining the old book for another year. We can, however, try to cast an experiment with the two books into a decision-theory mold (Chernoff & Moses, 1959) and reach a decision that takes the various costs and probabilities into consideration explicitly. How nearly the careful deliberations of an excellent educational administrator approximate this decision-theory model is an important problem which should be studied.

Factors Jeopardizing Internal and External Validity

In the next few sections of this chapter we spell out 12 factors jeopardizing the validity of various experimental designs.[3] Each factor will receive its main exposition in the context of those designs for which it is a particular problem, and 10 of the 16 designs will be presented before the list is complete. For purposes of perspective, however, it seems well to provide a list of these factors and a general guide to Tables 1, 2, and 3, which partially summarize the discussion. Fundamental to this listing is a distinction between *internal validity* and *external validity*. *Internal validity* is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? *External validity* asks the question of *generalizability*: To what populations, settings, treatment variables, and measurement variables can this effect be generalized? Both types of criteria are obviously important, even though they are frequently at odds in that features increasing one may jeopardize the other. While *internal validity* is the *sine qua non*, and while the question of *external validity*, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal. This is particularly the case for research on teaching, in which generalization to applied settings of known character is the desideratum. Both the distinctions and the relations between these two classes of validity considerations will be made more explicit as they are illustrated in the discussions of specific designs.

Relevant to *internal validity*, eight different classes of extraneous variables will be presented; these variables, if not controlled in the experimental design, might produce effects confounded with the effect of the experimental stimulus. They represent the effects of:

1. *History*, the specific events occurring between the first and second measurement in addition to the experimental variable.
2. *Maturation*, processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. *Testing*, the effects of taking a test upon the scores of a second testing. ?
4. *Instrumentation*, in which changes in the calibration of a measuring instrument or changes in the observers or scorers used may produce changes in the obtained measurements.
5. *Statistical regression*, operating where groups have been selected on the basis of their extreme scores. ?
6. Biases resulting in differential *selection* of respondents for the comparison groups.
7. *Experimental mortality*, or differential loss of respondents from the comparison groups.
8. *Selection-maturation interaction*, etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

The factors jeopardizing *external validity* or *representativeness* which will be discussed are:

9. The *reactive* or *interaction effect of testing*, in which a pretest might increase or

[3] Much of this presentation is based upon Campbell (1957). Specific citations to this source will, in general, not be made.

decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of the effects of the experimental variable for the unpretested universe from which the experimental respondents were selected.

10. The *interaction* effects of *selection* biases and the *experimental variable.*

11. *Reactive effects of experimental arrangements,* which would preclude generalization about the effect of the experimental variable upon persons being exposed to it in nonexperimental settings.

12. *Multiple-treatment interference,* likely to occur whenever multiple treatments are applied to the same respondents, because the effects of prior treatments are not usually erasable. This is a particular problem for one-group designs of type 8 or 9.

In presenting the experimental designs, a uniform code and graphic presentation will be employed to epitomize most, if not all, of their distinctive features. An *X* will represent the exposure of a group to an experimental variable or event, the effects of which are to be measured; *O* will refer to some process of observation or measurement; the *X*s and *O*s in a given row are applied to the same specific persons. The left-to-right dimension indicates the temporal order, and *X*s and *O*s vertical to one another are simultaneous. To make certain important distinctions, as between Designs 2 and 6, or between Designs 4 and 10, a symbol *R*, indicating random assignment to separate treatment groups, is necessary. This randomization is conceived to be a process occurring at a specific time, and is the all-purpose procedure for achieving pretreatment equality of groups, within known statistical limits. Along with this goes another graphic convention, in that parallel rows unseparated by dashes represent comparison groups equated by randomization, while those separated by a dashed line represent comparison groups not equated by random assignment. A symbol for matching as a process for the pretreatment equating of comparison groups has not been used, because the value of this process has been greatly oversold and it is more often a source of mistaken inference than a help to valid inference. (See discussion of Design 10, and the final section on correlational designs, below.) A symbol *M* for materials has been used in a specific way in Design 9.

THREE PRE-EXPERIMENTAL DESIGNS

1. The One-Shot Case Study

Much research in education today conforms to a design in which a single group is studied only once, subsequent to some agent or treatment presumed to cause change. Such studies might be diagramed as follows:

X *O*

As has been pointed out (e.g., Boring, 1954; Stouffer, 1949) such studies have such a total absence of control as to be of almost no scientific value. The design is introduced here as a minimum reference point. Yet because of the continued investment in such studies and the drawing of causal inferences from them, some comment is required. Basic to scientific evidence (and to all knowledge-diagnostic processes including the retina of the eye) is the process of comparison, of recording differences, or of contrast. Any appearance of absolute knowledge, or intrinsic knowledge about singular isolated objects, is found to be illusory upon analysis. Securing scientific evidence involves making at least one comparison. For such a comparison to be useful, both sides of the comparison should be made with similar care and precision.

In the case studies of Design 1, a carefully studied single instance is implicitly compared with other events casually observed and remembered. The inferences are based upon general expectations of what the data would have been had the *X* not occurred