Psychological Bulletin
1976, Vol. 83, No. 2, 314-320

Within-Subjects Designs: To Use or Not To Use?

Anthony G. Greenwald
Ohio State University

This article considers the several factors pertinent to deciding whether a within- or between-subjects design should be employed for a research application. A general principle favoring within-subjects designs is the statistical efficiency afforded by removing subject variance from error terms used to test treatment effects. Within-subjects designs, however, are often faulted for being subject to context effects of practice, sensitization, and carry-over that may limit interpretation of results. At the same time, between-subjects designs are not devoid of context effects, but rather have the context that a single treatment affords itself. Since ecological validity of results depends on the correspondence of the research context to the generalization context, within-subjects designs may be preferred when the generalization context includes the equivalent of several concurrent treatments. The discussion focuses additionally on (a) procedures to minimize practice, sensitization and carry-over effects in within-subjects designs when they are not desired, and (b) means of using these effects to advantage in research.

Frequently an investigator faces the choice of whether to examine the effects of two or more experimental treatments by exposing each subject to (a) only a single treatment (between-subjects design) or (b) several or all of the treatments (within-subjects or repeated-measures design). Grice (1966) has pointed out that the pattern of treatment effects obtained may vary considerably between the two types of designs. However, only rarely does an investigator make a choice of type of design after consideration of the appropriateness of each type to the problem being investigated. I attempt to assemble here several considerations that may often be appropriate to the decision between a within- or between-subjects design.

Although they are mentioned briefly, statistical considerations relating to choice of design are not of primary interest here. These statistical matters are well handled in standard statistical texts, as referenced below. My aim, rather, is to detail the *psychological* considerations that are critical to the choice of design. Some of these points are also covered in statistical texts, particularly insofar as they may affect the choice of statistical procedures. I have added only a few novel points to these earlier treatments and have aimed more at (a) putting the several points together in a single place and (b) observing that the prevailing cautions against the use of within-subjects designs need to be moderated without, however, being abandoned.

Poulton (1973, 1974; see also Rothstein, 1974) has recently issued a general warning against within-subjects designs, pointing out that the context provided by exposure to other treatments ("range effect") may often alter the effect of a given treatment. This point is certainly valid and is acknowledged here by considering (a) how procedures may serve to minimize or maximize such context effects and (b) when it may or may not be appropriate to allow the occurrence of context effects. The context effects that may be generated by a within-subjects design are discussed under three headings: Practice, Sensitization, and Carry-Over.

Context Effects in the Between-Subjects Design

Poulton (1973) concluded that since context or range effects are to be expected in within-subjects designs, these designs should ordinarily be avoided or, if used, bolstered by

Preparation of this report was facilitated by support to the author from National Science Foundation Grant GS-42981 and National Institute of Mental Health Grant MH-20527.

Requests for reprints should be sent to Anthony G. Greenwald, Department of Psychology, Ohio State University, 404C West 17th Avenue, Columbus, Ohio 43210.

between-subjects design results. Implicit in this conclusion is the principle that the between-subjects design provides a standard of validity against which results of a within-subjects design must be evaluated. This may be questioned on three grounds. First, as Poulton (1973) noted, "The influence of range of stimuli cannot always be prevented by restricting each man to a single stimulus" (p. 115). This may be because extralaboratory experience leaves some residue of context. Second, even if the extralaboratory context can be safely ignored, the presentation of a single treatment to each subject does not really achieve the *absence* of context, but rather the presence of the context provided by the single treatment. An example makes this clearer.

Example 1: Researcher 1 uses two designs to study the effect of foreperiod duration on simple reaction time. In a between-subjects design, each subject is assigned to a single foreperiod treatment: 0, 200, 500, or 1,000 msec. In a parallel within-subjects design, each subject receives a series of trials in which the four treatments are randomly sequenced.

It is known that Researcher 1's results will be different for the two types of design. The within-subjects design may produce either an increasing or a decreasing function relating reaction time to foreperiod duration (see Poulton, 1973, Table 1). Which function will be obtained depends on whether the procedures are arranged to produce increasing or decreasing expectation of the response signal as the foreperiod grows. Thus, it may be said that the within-subjects design introduces an expectancy or readiness process that is affected by the context of other treatments (foreperiods).

Is this expectancy process absent from the between-subjects design? No—rather, readiness occurs and is focused at the end of the (single) expected foreperiod. Thus, the single treatment in the between-subjects design provides a very real context that influences performance. This context effect could be avoided by presenting each subject with only a single trial at the selected foreperiod duration, but this would be an impractical way of collecting data on the problem. Further, the researcher may well wish to *ignore* the first (or first several) trials, since these involve warm-up processes (effects due to lack of context!) that are not of interest.

These considerations raise the third basis for questioning the notion that between-subjects designs provide a standard of validity against which to evaluate within-subjects designs. In fact, the ecological or external validity (Campbell & Stanley, 1966) of a piece of research depends on the extent to which the research context approximates the context existing in the domain to which the researcher wishes to generalize the results. This point is considered further in the concluding section of this article.

Statistical Considerations

No attempt is made here to detail the technical problems involved in statistics used to analyze the within- or between-subjects designs. However, a few general principles of a statistical nature must be considered as background. A more complete discussion of these points may be found in standard sources such as Myers (1972, especially chapter 7) and Winer (1971, especially chapter 4).

Power. When each subject provides data for two or more treatments, the subject may be said to serve "as his own control" in comparisons among treatment effects (i.e., treatment differences are not confounded with subject differences). To the extent that the *subjects* classification in the ensuing analysis of variance constitutes a substantial source of variance, this feature of the within-subjects design results in substantially more sensitivity to treatment effects (power) than would characterize a between-subjects design employing the same number of observations. Since a k-treatment between-subjects design would employ k times the number of subjects used in a within-subjects design with the same number of observations, it is apparent that a within-subjects design might often reach a desired level of power while using fewer than $1/k$ times the number of subjects in an equally powerful between-subjects design. The within-subjects design can therefore represent an immense experimental economy, particularly when per-subject costs are

considerable in relation to per-treatment costs.

Violation of assumptions. The standard analyses of within-subjects designs depend on an assumption of equality of the variances of differences between pairs of treatments (Winer, 1971, p. 282). It has been noted by statisticians that real data often violate this assumption and that the standard F ratio tests may be biased considerably by such violations. For this reason, within-subjects designs must be treated with a certain amount of special statistical care. Nonetheless, the techniques for dealing with violations of assumptions seem well enough established so that such violations can be tolerated but not ignored. The appropriate procedures include tests for extent of departure from assumptions, adjustments in degrees of freedom to correct for such departures, and the use of alternative statistical tests such as the multivariate analysis of variance (see Poor, 1973), which make less restrictive assumptions.

Context Effects in Within-Subjects Designs

Practice

Example 2: Researcher 2 is interested in assessing the effects of performance at a rotary pursuit task under three levels of distracting white noise: 75 db (A), 90 db (B), and 105 db (C). Should the effects of the three treatments be compared in a within- or between-subjects design?

A within-subjects design in which subjects were given Treatments A, B, C on Days 1, 2, 3 would suffer the obvious problem that the effects of treatments would be confounded with days. To the extent that performance on the motor skill task improves with practice, as is quite likely, this particular within-subjects procedure would yield seriously misleading results. There is a sometimes satisfactory remedy of *counterbalancing* the assignment of treatments to days in either (a) all possible combinations (six in this example) or (b) a balanced subset of combinations, as in a 3×3 Latin square design with *days* as the column factor, *groups of subjects* as the row factor, and *noise treatments* as the cell entries. This solution may not be satisfactory because the several treatments may be differently effective at different levels of practice. As a result, the observed treatment effects may be mixed inseparably with treatment–practice interactions.

At this point Researcher 2 should consider the relative interest of (a) treatment effects under minimum practice, (b) treatment effects under extensive practice, or (c) treatment effects across a range of practice levels (i.e., treatment–practice interaction). If the researcher is interested in treatment effects under minimum practice, the within-subjects design is inappropriate because subjects are providing data for two of the three treatments (more generally, $k - 1$ of k treatments) under more than minimum practice. A between-subjects design would be obligatory. If interest is in the treatment effect on the highly practiced skill, then a completely within-subjects design is possible, employing extensive practice to achieve a performance asymptote prior to administration of treatments in counterbalanced order. Finally, if interest is in the treatment effects across levels of practice, it may be best to use a combined between- and within-subjects design in which each subject provides data for performance at several levels of practice under only a single treatment condition.

The last design described above should be recognized as one of the most common instances of within-subjects designs—the learning experiment. Many psychologists would not think of studying practice effects with anything but a repeated-measurement assessment. Nonetheless, the decision to use a within-subjects design in a learning experiment should be made only after some thought. For a design with k different treatments and m levels of practice, it is possible to use km groups, each group being given a test for learning only once, after completion of the appropriate amount of practice. This might be advisable if the test for learning involves experiences that when applied repeatedly, might themselves affect performance. For example, paired-associate learning by passive exposure to word pairs could be tested after each passive exposure by presenting the first word of each pair and asking the subject to produce the second word that had been paired with each. For a variety of reasons, this type

of test might affect later performance independently of what was learned during the exposure period. On the other hand, if an anticipation method is being used (learning trials consist of first-word presentation, after which the subject tries to produce the second before being shown it), then the researcher is able to obtain information on performance at various stages of practice without interfering in any way with the practice procedure. Here it would be folly to employ anything but a repeated-measurement procedure for the study of acquisition.

Summary. A within-subjects design should be avoided in studying effects of several treatments when the researcher is interested in the effects of the treatments in the absence of practice and practice is likely to affect performance (either a main effect of successive tests or an interaction of successive tests with treatments). For the purpose of using a within-subjects design, undesired practice effects may sometimes be controlled by counterbalancing the sequence of treatments, or may be avoided by providing extensive practice prior to administering any treatments. Choice among within- or between-subject procedures here should depend on the level(s) of practice at which it is appropriate to examine the treatment effects. Finally, the practice effect is often intended to be the direct object of study itself—in learning experiments. Here, within-subjects designs will often be appropriate, but only when performance information can be obtained (as it frequently can) without having an impact on the acquisition process.

Sensitization

Example 3: Researcher 3 wishes to determine the effect of room illumination on worker productivity. Each subject is put to work on a well-learned task in a room in which the illumination is altered at periodic intervals in counterbalanced order across subjects, and the investigator determines the work rate under each illumination condition.

Researcher 3 should be concerned here with the possibility that the subject can readily discriminate the illumination differences and may thus be more sensitive and responsive to illumination than if there were exposure to only one of the several illumination level treatments. This sensitization to treatment variations may result in the subject's forming hypotheses about the treatment effects and responding to those perceived hypotheses rather than or in addition to the treatments themselves.

A variety of camouflaging strategies may be used to minimize the sensitization problem. The researcher in Example 3 may alter illumination from one treatment level to another so gradually that the subject will not notice it. In other circumstances, the experimenter may systematically alter several variables extraneous to the research design in order to draw attention away from a critical treatment variable (while, of course, not confounding the treatment with the extraneous variables).

The fact that perceptions of differences among treatments may be enhanced by their juxtaposition in a within-subjects design may be used to advantage in research when the experimenter is interested in observing the subject's capacity to discriminate such differences. Psychophysical studies constitute a large category of experiments in which the sensitization effect may be put to work for the researcher. In a brightness-judging experiment, for example, the experimenter is interested in the perceiver's sensitivity to brightness differences and wishes to optimize the conditions for observing such discrimination ability. By juxtaposing different treatments (brightnesses) in a within-subjects design, the limits of discrimination capacity can be assessed much more readily than in a between-subjects design.

Summary. A within-subjects design should be avoided when juxtaposition of treatments enhances perception of treatment variations *if* such perceptions can interfere with the processes the researcher desires to study. With ingenuity, it may often be possible to camouflage treatments so that this problem can be avoided. In quite a few experimental situations, particularly studies of perceptual discrimination, sensitization as a consequence of juxtaposing treatments (stimuli) in a within-subjects design will greatly facilitate the research.

Carry-over

Example 4: Researcher 4 is interested in the effects of Drugs A, B, and C on performance on a simple reaction time task. In order to employ a within-subjects design, Researcher 4 gives each subject four performance tests separated by 20 minutes, each test being preceded by the administration of a standard dosage injection of one of the three drugs or a placebo control and with the sequence of treatments being counterbalanced across subjects.

In general, a carry-over effect occurs when the effect of one treatment persists in some fashion at the time of measurement of the effect of another. In Example 4, there are two types of potential carry-over. One is due to practice at the performance task and has been discussed separately above. The second is that traces of prior drug treatments may be present at the time of testing the effects of a later treatment. Counterbalancing provides an only partially adequate solution to this problem, since the interference effects may not be bidirectional and, further, they may obscure the treatment effects of the drugs taken individually.[1]

The chief means of reducing carry-over effects is to separate the treatments in time. This would likely be an effective means of applying a within-subjects design to the problem given in Example 4, assuming that practice effects are not also involved. In general, the strategy of separating treatments in time will be effective in reducing intertreatment carry-over only to the extent that the effects of any treatment are not permanent.

In addition to the study of learning, there are several other major areas of study in which the target of study is some process that can be interpreted as an intertreatment carry-over in the framework of a within-subjects design. Perceptual assimilation and contrast, incentive contrast, violation of expectation, transfer of training, primacy–recency in persuasion, resistance to extinction, and various types of adaptation are some of these. The fact that intertreatment carry-overs are likely to be a major source of serendipitous findings should not be overlooked as one of the virtues of employing within-subjects designs in which treatments that would otherwise not be examined in near temporal proximity are juxtaposed.

Procedures that permit the occurrence of carry-over effects present special problems for statistical analysis. Cochran and Cox (1957, pp. 133–142) discussed a variety of means of estimating separately the direct and carry-over effects of experimental treatments.

Summary. When treatments have persistent effects, a within-subjects design may be unsatisfactory because the effect of one treatment may still be in force at the time of measuring another's effect. However, the within-subjects design may be salvaged in this case by increasing the separation of the treatments in time. Effects dependent on carry-over or, more generally, upon the sequence in which treatments are administered and their temporal proximity are frequently of psychological interest in and of themselves.

External Validity

Several of the concerns already treated are appropriate to evaluating the internal validity of an experiment—that is, Does the within-subjects design permit the experimenter to test the hypothesis of interest, or will consequences of using the design in some way contaminate (by practice, sensitization, or carry-over) the hypothesis test? Now we take up a matter that may be at odds with some of these considerations and ask how the choice of design affects the external (or ecological) validity of the experiment (i.e., the ability of the researcher to account for the effects of treatment variations as they may occur in interesting nonresearch settings). (See Campbell and Stanley, 1966, for an exposition of internal and external validity.)

Example 5: Researcher 5 is interested in the effects of source credibility on persuasion, and is considering two possible designs. In one, a between-subjects design, communications on

[1] This inadequacy of counterbalancing involves the same considerations mentioned in discussing the possible inadequacy of counterbalancing in removing practice effects. Practice is certainly an instance of the general class of carry-over effects, but has been discussed separately because of the special status of learning effects in psychological research.

two topics are attributed, for some subjects, to a trustworthy and expert source whereas, for other subjects, the same communications are attributed to an untrustworthy and inexpert source. In an alternate within-subjects design, each subject is exposed to the same two communications, but one is attributed to the high-credible source, the other to the low-credible source, with source-communication assignments being counterbalanced across subjects. Which design is preferable? [2]

Persons familiar with persuasion research will be aware that the between-subjects design is most often chosen for the examination of source credibility effects (but not always —see Osgood & Tannenbaum, 1955). But this is perhaps the less justifiable choice if the researcher's primary interest is in predicting or characterizing source effects in the nonlaboratory environment. Consider that people tend to be exposed to persuasive communications in clusters in many mass communication settings, these communications frequently being identified with different sources (e.g., columns in a newspaper editorial section, political or product advertisements in magazines, or on radio or television). Therefore, the within-subjects design for studying the consequences of communicator credibility may have greater external validity than does the between-subjects design.[3]

Similar considerations would lead to a preference for the between-subjects design for other problems. For example, a researcher may be interested in studying the effect of reinforcement-based versus psychoanalytically based therapies for phobia symptoms. In such a situation exposure of the same subjects to several different treatments would create a situation rather lacking in external validity.

Considerations of external validity should not necessarily be uppermost in the researcher's mind. The between-subjects design may be preferred even in some situations for which the within-subjects design would have greater external validity, because the between-subjects design may allow cleaner tests of theoretical hypotheses. Example 6 presents such a case, in which internal validity is of more concern to the investigator than is external validity.

Example 6: Researcher 6 is interested in the effects of witnessing televised violence on subsequent aggressive behavior of children. A within-subjects design would involve exposure of each subject to several different program sequences of varying degrees of violence, each followed by the provision of some opportunity to act aggressively in a play situation with other children. Should this design be employed?

In this case, the within-subjects design might not be preferable because the carry-overs among treatments (subjects still being under the influence of Program A at the time of Test B) might weaken the researcher's hypothesis test. Accordingly, the between-subjects design might be chosen even though the within-subjects design clearly has greater external validity in its correspondence to the mixture of types of programs the child would normally see on television.

In many cases, a greater stress on internal validity than on external validity would lead to a choice of the within-subjects design. This might be particularly true in cases of basic research for which there is no readily apparent nonlaboratory setting for which the research is an analog. For example, a neuropsychologist studying functions of single cells in the central nervous system should almost certainly examine the consequences of the range of treatments in which he or she is interested on each of the research subjects.

Summary. Considerations of external or ecological validity may sometimes be at odds with considerations related to practice, sensitization, and carry-over effects. Thus, the

---

[2] Both of the designs mentioned in this example are within-subjects or repeated-measurement designs in that the effects of two communications are studied on each subject. However, the treatment variation of credibility is a between-subjects variation in the first design and a within-subjects variation in the second.

[3] Poulton's (1973) concerns about range effects are quite relevant here. The investigator who is interested in generalizing to nonlaboratory settings should be concerned to see that the range and distribution of treatment variations in the experiment correspond to their range and distribution in the appropriate nonlaboratory setting. Otherwise, the experimental treatment effects may misrepresent the effects of their nonlaboratory analogs.

within-subjects design may often have greater external validity because it contains these confounds, but these may also interfere with the researcher's ability to isolate the treatment effects.

Conclusions

A general force operating in the direction of selecting a within-subjects design is the statistical efficiency afforded by the removal of subject variance from error terms used to test treatment effects. However, context effects may often interfere with hypothesis tests and, therefore, should take precedence over considerations of statistical efficiency when choosing a design. Context effects may occur in either a between- or a within-subjects design, but the range of possible effects is much greater in the latter type of design and, correspondingly, the experimenter has greater potential control over them by selecting ranges of treatments to administer. In many situations a within-subjects design can be made more acceptable by appropriate counterbalancing of treatment sequences (to control practice effects), by camouflaging treatments (to reduce sensitization to the treatment dimensions), or by separating treatments in time (to reduce carry-over effects). In still other circumstances, the deliberate introduction of these context effects in a within-subjects design may have the desirable consequences of permitting the study of some interesting aspect of the context effect itself or of increasing the external (ecological) validity of the research.

REFERENCES

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1966.

Cochran, W. G., & Cox, G. *Experimental designs.* New York: Wiley, 1957.

Grice, G. R. Dependence of empirical laws upon the source of experimental variation. *Psychological Bulletin,* 1966, *66,* 488–498.

Myers, J. L. *Fundamentals of experimental design* (2nd ed.). Boston: Allyn & Bacon, 1972.

Osgood, C. E., & Tannenbaum, P. H. The principle of congruity in the prediction of attitude change. *Psychological Review,* 1955, *62,* 42–55.

Poor, D. D. S. Analysis of variance for repeated measures designs: Two approaches. *Psychological Bulletin,* 1973, *80,* 204–209.

Poulton, E. C. Unwanted range effects from using within-subjects experimental designs. *Pyschological Bulletin,* 1973, *80,* 113–121.

Poulton, E. C. Range effects are characteristic of a person serving in a within-subjects experimental design—A reply to Rothstein. *Psychological Bulletin,* 1974, *81,* 201–203.

Rothstein, L. D. Reply to Poulton. *Psychological Bulletin,* 1974, *81,* 199–201.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

(Received December 6, 1974)