8 Between- or Within-Subjects Design: A Methodological Dilemma

Gideon Keren
Free University of Amsterdam

A common question, frequently faced by researchers, concerns the use of a between- or within-subjects experimental design. In the between design each subject is exposed to a single treatment, whereas in the within (or repeated measures) design, subjects are exposed to several or all the treatments that are included in the study. Faced with the choice of manipulating the independent variable in a between- or within-subjects design, which one should the investigator use? The purpose of this chapter is to discuss the considerations that should guide researchers in making their choice.

A brief glance at the experimental psychological literature suggests that studies in perception, psychophysics, memory, and learning virtually all employ a within-subjects design. In contrast, studies in areas such as social psychology, personality, and decision making tend to use a between-subjects design. The adoption (in different research areas) of one or the other design, implies certain implicit assumptions that may be crucial in interpreting empirical results. Moreover, there is evidence that the two experimental designs do not always yield the same pattern of results (see Erlebacher, 1977, for some illuminating examples). It is therefore imperative that researchers will use the appropriate considerations in making their choice of design, and be aware of the assumptions underlying the chosen design in the process of interpreting experimental results.

The different relevant considerations, although intertwined, can be classified into three groups. First are *statistical considerations* reflecting the different analysis associated with each design. Although we do not elaborate here on subtle statistical issues associated with between- or within-subjects design,[1] several

[1]A detailed exposition of the statistical characteristics of within-subjects design is presented in chap. 3, Vol. 2. Meyers (1979, chap. 7) also offers a lucid discussion on the topic.

notes are made in the following section, and potential misconceptions are pointed out.

A second class of considerations are *methodological issues*. The concern here is with potential (and usually unwanted) side effects that are solely due to the choice of design. Much of the debate regarding the appropriate choice of design has focused on the type of issues presented in the following section.

Finally, the third type of consideration that should be taken into account are *theoretical issues*. These relate to the particular questions the researcher wants to answer. Although this aspect has received relatively little attention, a later section purports that the theory and particular hypotheses one wants to test should be an essential determinant in the choice of design.

As mentioned, the empirical results obtained from within- and between-subjects designs are often incompatible. For theoretical or methodological reasons, the researcher may be interested in testing the difference in the patterns of results obtained from the two different designs, in which case the design type explicitly plays the role of an independent variable. The appropriate methods for analyzing such experiments (e.g., Erlebacher, 1977; Rosenthal & Rubin, 1980) and some representative empirical results are presented in a later section. The major issues to be taken into account in the process of choosing a within- or between-subjects design are summarized in the final section.

STATISTICAL ASPECTS

It is often claimed that the advantage of the within-subjects design is based on the fact that differences observed among conditions are not confounded with individual differences. Consequently, the claim goes on, the exclusion of individual differences results in a higher degree of sensitivity to treatment effects or, in other words, yields a substantial increase in power as compared with a between-subjects design. This statement is somewhat simplistic and may lead to misunderstanding. To make our point clear, consider a simple experiment with i subjects ($i = 1, \ldots, n$) and two different treatments $j = 1, 2$,[2] which can be modeled by

$$y_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ij} \tag{1}$$

where α_i represent subject differences, β_j stands for the treatment effect, and $\alpha\beta_{ij}$ is the interaction between subjects and treatment. To estimate the difference between the two treatments in a *within-subjects* design, we calculate

$$\begin{aligned} y_{i1} - y_{i2} &= \mu + \alpha_i + \beta_1 + \alpha\beta_{i1} + e_{i1} - (\mu + \alpha_i + \beta_2 + \alpha\beta_{i2} \\ &\quad + e_{i2}) \\ &= (\beta_1 - \beta_2) + (\alpha\beta_{i1} - \alpha\beta_{i2}) + (e_{i1} - e_{i2}) \end{aligned} \tag{2}$$

[2]Although the discussion here is limited to two treatment groups, the generalization to K treatment groups is straightforward.

and averaging across subjects we obtain

$$\bar{y}_{.1} - \bar{y}_{.2} = (\beta_1 - \beta_2) + (\overline{\alpha\beta}_{.1} - \overline{\alpha\beta}_{.2}) + (\bar{e}_{.1} - \bar{e}_{.2}). \tag{3}$$

Note that the first term in Equation 3 represents fixed effects, the second term contains interactions that are correlated (with $\overline{\alpha\beta}_{.1} = -\overline{\alpha\beta}_{.2}$, and thus the second term can be expressed as $2\alpha\beta_{.1}$), and the third term is comprised of two uncorrelated error terms. The variance associated with $\bar{y}_{.1} - \bar{y}_{.2}$, which is the variance of the estimate in a within-subjects design with n subjects is

$$\begin{aligned}\sigma^2_{WS} &= \text{var}[\Sigma(y_{i1} - y_{i2})/n] \\ &= 4\ \text{var}(\alpha\beta_{i1})/n + [\text{var}(e_{i1}) + \text{var}(e_{i2})]/n.\end{aligned} \tag{4}$$

The comparable estimate of the difference between the two treatments in a *between-subjects* design is

$$\begin{aligned}y_{i1} - y_{i'2} = \mu + \alpha_i + \beta_1 + \alpha\beta_{i1} + e_{i1} - (\mu + \alpha_{i'} + \beta_2 + \\ \alpha\beta_{i'2} + e_{i'2}) = (\alpha_i - \alpha_{i'}) + (\beta_1 - \beta_2) + (\alpha\beta_{i1} - \alpha\beta_{i'2}) \\ + (e_{i1} - e_{i'2})\end{aligned} \tag{5}$$

(where the prime $'$ is used to distinguish between the two groups). Averaging across subjects in each group we obtain

$$\begin{aligned}\bar{y}_{.1} - \bar{y}_{.'2} = (\bar{\alpha}_{.} - \bar{\alpha}_{.'}) + (\beta_1 - \beta_2) + (\overline{\alpha\beta}_{.1} - \overline{\alpha\beta}_{.'2}) \\ + (\bar{e}_{.1} - \bar{e}_{.'2}).\end{aligned} \tag{6}$$

Note that in contrast to Equation 3, $\bar{\alpha}_{.}$ and $\bar{\alpha}_{.'}$ are uncorrelated because they represent different subjects, so that we now have three sources of uncorrelated errors. The variance of the estimate of the treatment for a between-subjects design (with n subjects for each treatment) is

$$\begin{aligned}\sigma^2_{BS} = \text{var}[\Sigma y_{i1}/n - \Sigma y_{i'2}/n] = 2\text{var}(\alpha_i)/n + 2\text{var}(\alpha\beta_{i1})/n + \\ [\text{var}(e_{i1}) + \text{var}(e_{i'2})]/n.\end{aligned} \tag{7}$$

There are several important conclusions to be derived from the previous analysis. First, as far as the comparison between treatments is concerned, both the within- and between-subjects designs provide unbiased estimates (Equations 3 and 6 respectively) of the treatment effect $(\beta_1 - \beta_2)$. The difference between the two designs lies in the precision by which the treatment effect is estimated, namely by the variance of the estimate (Equations 4 and 7). Moreover, it is not necessarily the case that the within-subjects design will always result in a better precision. To make this point clear we can compare the variance of estimate for the two designs and obtain

$$\sigma^2_{WS} - \sigma^2_{BS} = (2/n)^*[\text{var}(\alpha\beta_{ij}) - \text{var}(\alpha_i)] \tag{8a}$$

or alternatively

$$\sigma^2_{BS} - \sigma^2_{WS} = (2/n)^*\text{cov}(y_{i1}, y_{i2}). \tag{8b}$$

The conclusion from Equation 8b is that a within-subjects design is more powerful whenever $\rho_{12} > 0$, that is there is a positive correlation between the treat-

ments. In practice this correlation is usually (but not necessarily always!) positive, which accounts for the fact that the error term for a within-subjects design is typically smaller than the comparable error term of a between-subjects design.[3] In summary, the within-subjects design usually entails a larger power but not necessarily under all circumstances.

The within-subjects design contains two additional appealing characteristics: One concerns the obvious advantage derived from the larger number of degrees of freedom associated with the within-subjects design. The other is related to the saving in the number of subjects required to complete the experiment. Indeed, this economical aspect often serves as the major appeal of a within-subjects design.

METHODOLOGICAL ISSUES

It is important to distinguish between two types of within-subjects designs. In one type, the same subject is exposed to different conditions and there is a substantial difference between stimuli employed in different experimental conditions. Thus, in this sort of design the subject is never exposed to the same (or a similar) stimulus more than once. A major reason to employ a within-subjects design under such circumstances is the claim that subjects serve as their own control, which enables a direct and unconfounded comparison between the different conditions. Additional advantages are the gain in statistical power and economy in use of subjects.

In the other type of design, the same subject may experience the same (or a very similar) stimulus on several trials. Such a design is certainly appropriate when the researcher's explicit goal is to investigate learning effects due to repeated trials (where trials, or blocks of trials are treated as an independent variable). If practice and learning effects are not desired, such a design can be used only if there are sufficient reasons to assume that (at least) for all practical purposes, the different trials are independent (in the psychological sense[4]).

A major concern with regard to the adoption of a within-subjects design is the lack of independence between different trials or different treatments administered

---

[3]An analysis similar to the one already described is offered by Hays (1973) in the context of paired observations. Hays showed that for groups matched by pairs, the variance of a difference between means is

$$\sigma^2_{\text{diff}} = \sigma^2_{M_1} + \sigma^2_{M_2} - 2\text{cov}(M_1, M_2)$$

where M_1 and M_2 are the means of the two groups. Hays noted that "in general, for groups matched by pairs, this covariance is a positive number, and thus the variance and standard error of a difference between means will usually be *less* for matched than for unmatched groups.

[4]One should distinguish between *statistical* independence and the *psychological* independence. The latter implies that the behavior of a subject on trial n is not influenced by the behavior of the same subject on trial $n - 1$. Although the two concepts are related, they are not identical.

to the same subject. Such dependencies among trials or treatments may introduce undesired interactions between the particular treatment under study and unwanted exogenous influences that cannot be separated. The potential presence of contaminated effects can appear in different forms.

Consider first the case, as frequently used in psychophysics or reaction time studies, in which the same or very similar stimuli are presented on a large number of trials. Literally, every event can occur only once and thus, from a strict point of view, one can argue that even two trials in which exactly the same stimulus is presented are not the same. Adopting a slightly looser approach, there are often sufficiently strong reasons to assume (for all practical purposes) independence among the different trials. A typical example is psychophysics, in which the same stimulus is presented a large number of times. Even in this case, doubts may be raised about the independence assumption, in particular the effect of practice (discussed later) and the buildup of potential expectations in the course of presenting repeated trials cannot be ignored.

Potential dependencies may also exist in a design in which a subject is exposed to a single trial in each of several conditions or treatments. Memory of previous trials cannot be erased nor can we prevent subjects from forming any hypotheses about the nature of the experiment (e.g., Rothenthal, 1976), thus any inferred relations among the different treatments may result in subjects' responses that are not necessarily independent. Responses to trials in previous conditions may affect, directly or indirectly, the subject's response in the present condition.

The concern about possible dependencies that are not always transparent is thus a major argument usually raised against the use of within-subjects designs. It should be noticed that the extent to which potential dependencies may exist could be judged on different dimensions: Similarity between stimuli, potential inferred relations between different items, the range of potential responses, and others. Poulton (1973) referred to such dependencies as *range effects* and provided a long list of experimental examples[5] in which subjects' responses are influenced by the range of stimuli, by the range of responses or by both. According to Poulton, such effects are always present when stimuli or responses can be ordered in a consistent manner, and where such an ordering implies explicit or implicit (derived) dependencies among different items and different experimental conditions. One customary remedy against such possible dependencies is to randomize items and conditions in order to avert the construction of a consistent structure. Randomization, however, may not always be effective: Subjects may search for dependencies even when stimuli are presented randomly, because they are unable to distinguish between random and nonran-

[5]Rothstein (1974) suggested that most of the examples drawn by Poulton concern magnitude estimation and motor performance studies (which indeed are particularly vulnerable to range effects), and questioned the generality of Poulton's sample.

dom series (see Rosenthal, chap. 20, this vol.). This is illustrated by the probability matching phenomenon (e.g., Estes, 1964), in which subjects attempt to match their response probabilities to experimental probabilities even though stimuli are presented randomly.

Greenwald (1976) referred to the possible contaminations arising from the use of a within-subjects design as *context* effects, and classified them in three categories that he termed practice, sensitization, and carry-over effects.

1. *Practice effects* are frequently present in psychophysical, motor performance, and attention studies that require a large number of trials. If the different experimental treatments are administered on different days, or even in blocks of trials in the same experimental session, the possibility that treatment effects are confounded with practice cannot be ruled out. One possibility to minimize such effects is *counterbalancing,* namely, equal use of all the possible combinations by which the treatments can be ordered (alternatively, if the number of combinations is too large, one may use a balanced subset of combinations such as employing a latin square design). However, as Greenwald pointed out, such a solution is not always satisfactory, because various treatments may be differently effective at different levels of practice. Remember that the possible interactions of treatment and practice in a within-subjects design cannot be removed: The only effect of counterbalancing is to spread the unwanted variance arising from such interactions among the different treatments, with the hope that it is equally spread. An alternative, and perhaps better, method for preventing the contamination of practice effects is to provide extensive training prior to the experimental tests, assuming that performance has reached asymptote by the end of the training period.

2. *Sensitization* refers to the possibility of perceived dependencies (regardless of whether or not they are justified) between trials or treatments that may lead subjects to form hypotheses about the treatment effect and respond accordingly.

For example, consider the experiments conducted by Keren and Wagenaar (1987), which were designed to investigate whether subjects' preferences concerning choice among gambles when played a single time (unique condition) or many times (the repeated condition) are the same. For purposes of comparison the gambles in the two conditions had to be the same and there was one difference: In the unique condition subjects were told that the gamble of their choice will be played only once, whereas subjects in the repeated condition were told that the chosen gamble will be played 10 times. The similarity between the unique and repeated conditions was so transparent (because the stimuli in the two conditions were in fact identical), that if a within-subjects design had been used, whatever subjects' preferences were, they would probably make sure their

choices in the two conditions were not contradictory. In other words, the preferences exhibited in such a design would have been highly correlated.[6]

3. *Carry-over effects* result when the effect of a specific treatment persists in one way or the other, and thus contaminates the measurements at the time that the effect of other treatments is tested. Practice effects are an instance of carry-over effects; drug treatments that may leave a trace for a period of time serve as another example.

As with practice effects, the use of counterbalancing provides only a partially adequate remedy. A more effective way according to Greenwald is to separate the relevant treatments by sufficiently long time intervals, though this may not always be feasible.

Though the possible contaminations of within-subjects designs due to context effects are certainly real and can at best be only partially controlled, the question is whether a between-subjects design provides an adequate safeguard against context effects. Even Poulton (1973), who is certainly the most outspoken researcher regarding the deficiencies of within-subjects designs, admitted that potential context effects cannot always be prevented by reverting to a between-subjects design. Mainly, even in a between-subjects design context effects may be present, though they may be of a different nature compared with those arising from a within-subject design.

The claim that context effects will also influence a between-subjects design, is based on two (plausible) assumptions (Birnbaum, 1982; Greenwald, 1976) namely that (a) context is also provided by a single treatment, and (b) that subjects do not enter the experimental laboratory as tabula rasa but rather with extralaboratory experience that may leave some residue of context. It is the interaction of these two sources that may lead to effects that the researcher cannot completely control (by either experimental or statistical methods). Demonstrations of potential context effects in between-subjects design are reported in Birnbaum (1982) and Birnbaum and Mellers (1983).

Effects of stimulus range, stimulus spacing, and frequency, all of which are potential sources for contextual effects in perception, have been studied by Parducci and incorporated in his range-frequency theory (e.g., Parducci, 1965, 1983). Some methods, which may potentially help to identify the locus of such context effects in judgments, have been proposed by Wedell (1990). If both within- and between-subjects designs are vulnerable to context effects (even if of somewhat different natures), how should a researcher proceed in making his selection of design? Notwithstanding the different considerations discussed up to

[6]One could ask whether in a between-subjects design (as indeed employed by Keren and Wagenaar), the comparison of choices made by different subjects is a meaningful one. This question is addressed in a later section.

now, the decision of which design to adopt should also be determined by the theoretical framework in which the research is conducted and the particular hypotheses one wants to test.

EXTERNAL VALIDITY AND THEORETICAL FRAMEWORK

The methodological considerations discussed up to now refer mainly to what may be termed *internal validity,* that is, the extent to which valid inferences can be derived from the particular experiment at hand. Two additional major concerns are (a) whether the particular design chosen also has external validity, and (b) whether the inferences that can indeed be logically drawn by using a particular design, are also the most relevant for enhancing the broader research program in which the study is being conducted.

External Validity

Greenwald (1976) correctly pointed out that the *external validity* of an experiment, namely the extent to which the results can be generalized beyond the specific experiment that is being conducted, is an important aspect in deciding which type of design one should use. Considerations of external validity are to a large extent determined by contextual effects. As mentioned earlier, contextual effects may be present in either a between- or a within-subjects design, so it is important to consider and weigh such potential effects before making a final decision on the design to be employed. Potential contextual effects should also not be ignored in the process of interpreting experimental results.

Considerations Prescribed by the Theoretical Framework

What design to use should in many cases be determined by the underlying theory and the particular hypotheses one would like to test. Unfortunately, this type of consideration has frequently been neglected by researchers, perhaps because no general guidelines can be outlined in this respect, and the relevant deliberations will be unique for each case. I believe, however, that this type of consideration has been often overlooked, so I try to highlight its importance by using an elaborated example from the decision-making literature.

Kahneman and Tversky (1979) recently offered an axiomatic alternative to traditional utility theory, which they termed *prospect theory.* The prospect $A = (x_i, p_i, y_i)$ is defined as a contract that yields outcome x_i with probability p_i, and an outcome y_i with probability $1 - p_i$. A specific hypothesis obtained from prospect theory is the so-called reflection effect: According to this hypothesis,

given the choice between two prospects people prefer the more risky prospect (smaller probability and higher gains) when negative outcomes are involved, and the less risky prospect when positive outcomes are involved. In other words, risk seeking in the negative domain is accompanied by risk aversion in the positive domain. Thus, preferences among prospects in the negative domain are a reflection (mirror image) of their preferences among the corresponding positive prospects.

Empirical support for the reflection hypothesis is obtained from experimental data reported by Kahneman and Tversky (1979). Employing a between-subjects design, they used five separate problems to demonstrate significant reversals of preference between gain and loss prospects. Hershey and Schoemaker (1980) presented an interesting challenge to the experimental evidence provided by Kahneman and Tversky, and argued that the evidence based on a between-subjects design was not conclusive for supporting the reflection hypothesis.

In order to understand the reasoning of Hershey and Schoemaker, consider Table 8.1, which represents the four (hypothetical) preference combinations subjects may exhibit in an experiment testing the reflection effect. The options A_1 and A_2 are the less risky ones compared with B_1 and B_2, respectively. The cell entries n_1 through n_4 denote the percentage of subjects corresponding to each preference combination, such that $\Sigma n_i = 100$. Using a between-subjects design, the column totals c and d represent subjects' choices among the positive prospects, and the row totals a and b represent subjects' choices among the negative prospects, such that $a + b = c + d = 100$.

In terms of Table 8.1 (using a between-subjects design) the reflection effect is exhibited if the row total "a" and the column total "d" (or alternatively "b" and "c" respectively) are each significantly larger than 50%. This is indeed the type of evidence provided by Kahneman and Tversky (1979). Hershey and Schoemaker (1980) however, questioned whether a between-subjects design provides an adequate test of the reflection effect. They argue that a given pattern of overall preferences (derived from a between-subjects design) could be consistent with a varying numbers of individual reversals.

Consider the experimental outcomes portrayed in the two matrices of Table 8.2. The marginals (i.e., row and column totals) are taken from Kahneman and Tversky (1979, problem 4) and show the reflection effect according to the criteria

TABLE 8.1
The Possible Preference Combinations in a 2 × 2 Design

		Positive prospects		
		A_1	B_1	
Negative prospects	A_2	n_1	n_2	a
	B_2	n_3	n_4	b
		c	d	100

TABLE 8.2
Two Different Patterns of Results for a Hypothetical Within-Subjects Experiment, with Identical Marginals of a Between-Subjects Design

		(A) *Positive prospects*				(B) *Positive prospects*		
		A_1	B_1			A_1	B_1	
Negative prospects	A_2	42	0	42	A_2	7	35	42
	B_2	23	35	58	B_2	58	0	58
		65	35	100		65	35	100

previously mentioned. Suppose, however, that the results were obtained from a within-subjects design. Hershey and Schoemaker pointed out that several values of n_1, n_2, n_3, n_4 (obtained from a within-subjects design) could underlie a given set of marginals a, b, c, d. Note, that the number of individuals who exhibit reflection in a within-subjects design is given by $n_2 + n_3$. According to this criterion, using the numbers in the left table (A) of Table 8.2 only 23% (0 + 23) of the individual subjects exhibit reflection. In contrast, 93% (58 + 35) of the individuals exhibit reflection in Table 8.2B (note that the margins in both tables A and B are the same). Thus, the results from a between-subjects design can be compatible with a mere 23% reflections, and at the time compatible with as much as 93% reflections.[7] Consequently, according to Hershey and Schoemaker (1980), the results from a between-subjects design are inconclusive.

Notwithstanding the claim of Hershey and Schoemaker, there are two important problems that remain open. First, the use of a within-subjects design for testing the reflection effect would be seriously hindered by unwanted contaminations discussed in the previous section. In particular, because the stimuli for the negative and the positive prospects are so similar, they may evoke the subjects' awareness to provide consistent responses (even if they do not reflect the true preferences). The use of a within-subjects design should be seriously questioned, because of the high likelihood of sensitization and carry-over effects in this case.

There is a second important consideration, related to the underlying theory and the particular hypotheses one would like to test. Without loss of generality, one may consider three different hypotheses concerning the outcomes of a hypo-

[7]Hershey and Schoemaker showed that given a pattern of results obtained from a between-subjects design (i.e., the values of a, b, c, d), the percentage of Individual Reversals (IR) that is compatible with a within subjects-design can be as small as

$$IR_{min} = 100 - \min\{a, c\} - \min\{b, d\}$$

and as large as

$$IR_{max} = \min\{a, d\} + \min\{b, c\}.$$

thetical experiment as illustrated in Table 8.1. To simplify the notation, let $P(A_1)$ denote the probability of choosing A_1 given the choice between A_1 and B_1. Similarly, for $P(B_1)$, and so forth. The three hypotheses one may contemplate are:

$$
\begin{array}{lll}
H1: & P(A1) > 1/2 \quad \text{and} \quad P(B2) > 1/2 \\
\quad \text{or alternatively} & P(B1) > 1/2 \quad \text{and} \quad P(A2) > 1/2 \\
H2: & P(A2|A1) < P(A2|B1) \\
\quad \text{or equivalently} & P(B2|A1) > P(B2|B1) \\
H3: & P(A1 \cap B2) + P(A2 \cap B1) > 1/2
\end{array}
$$

It is important to emphasize that $H1$, $H2$, and $H3$ are examples of three different classes of hypotheses: $H1$ is formulated in terms of marginals or average group preferences and, as such, may naturally be tested in a between-subjects design. $H2$ is formulated in terms of conditional probabilities (testing statistical dependencies). It clearly refers to individuals and consequently should be examined by a within-subjects design. Finally, $H3$ is explicitly phrased in terms of the number of individual reversals, and strictly speaking can accurately be tested only with a within-subjects design. It should be realized that despite the similarity among the three hypotheses, they are distinct and no one hypothesis implies either of the other two (a proof is given in Keren & Raaijmakers, 1988).

The previous analysis suggests that the design to use would depend on the particular hypothesis the researcher wants to test. If hypothesis H_1 is of interest, then a between-subjects design would be the most natural one to be used, whereas for testing hypotheses H_2 or H_3, a within-subjects design should be preferred. Which of the three hypotheses then is the most appropriate one to describe the reflection effect, would depend on how one interprets the reflection hypothesis and prospect theory. For instance, if prospect theory would allege the existence of a cognitive hypothetical construct that we may term the *reflection mechanism*, according to which the preferences of an individual in the negative domain are obtained by reflecting preferences in the positive domain, then certainly $H3$ would be the most natural hypothesis to test. On the other hand, if the reflection effect is only a label that conveniently describes a certain *pattern* of results (as is suggested by a careful reading of Kahnaman & Tversky, 1979), then the most appropriate hypothesis to test would be $H1$.

A lengthy exploration of possible interpretations of the reflection hypothesis is beyond the scope of this chapter (for more details, see Keren & Raaijmakers, 1988). The intention of the earlier discussion was only to demonstrate that the theoretical considerations and the exact formulation of the hypothesis one wants to test are crucial elements in the decision of what experimental design one wants to adopt.

DIRECT COMPARISONS OF BETWEEN- VERSUS WITHIN-SUBJECTS DESIGNS

Between- and within-subjects designs may often yield different patterns of results, so the researcher may sometimes be interested in comparing the results obtained from the two different designs. Such an undertaking may be expensive and time consuming, and the researcher should have good reasons for conducting such a test.

When a comparison between the two designs is desired, the design type becomes an explicit independent variable, and the question is whether design type interacts with any of the other substantive independent variables. Rosenthal and Rubin (1980) provided a lucid presentation of how to conduct such tests, and the following discussion borrows heavily from their article. As pointed out by Rosenthal and Rubin, there are at least two kinds of comparisons to be considered: comparison of the variabilities and comparison of means.

Consider an experiment in which subjects are tested on general knowledge items and are asked on each item to choose between two alternatives (e.g., The capital of Turkey is: 1. Istanbul. 2. Ankara.), and then provide a confidence rating (between 50% and 100%) that their chosen answer is indeed the correct one. Furthermore, suppose that items are divided into two categories, easy items with a high percent of correct answers, and difficult items for which the percent of correct responses is relatively low. The major question is whether subjects can distinguish between the difficult and easy items as reflected in their confidence ratings.

Table 8.3 presents hypothetical results of such an experiment. The first two columns contain the mean confidence ratings for 20 subjects who were tested on the easy items and 20 different subjects who were tested on the difficult items (a between-subjects design). Columns 3 and 4 contain the scores from an experiment in which subjects were exposed to both the easy and the difficult items (a within-subjects design), where easy and difficult items were randomly mixed. We start by comparing variabilities with each design separately. For the between-subjects design, the independently estimated variances are compared by employing an F test (e.g., Glass & Stanley, 1970):

$$F(19, 19) = S_D^2/S_E^2 = 45.40/38.37 = 1.18 \tag{9}$$

where S_D^2 and S_E^2 are the variance estimates for the difficult and easy conditions, respectively. The ratio is obviously not significant. For the within-subjects design in which the variances are not independent, the following test statistic (e.g., Glass & Stanley, 1970) is employed:

$$t(18) = \frac{S_D^2 - S_E^2}{4S_D^2S_E^2(1 - r_{DE}^2)/(N - 2)} = 1.098 \tag{10}$$

TABLE 8.3
Hypothetical Results for Two Experiments on Confidence Ratings Using a Between- and Within-Subjects Design

Subject Number	Mean Confidence Ratings			
	Between Design		Within Design	
	Easy Items	Difficult Items	Easy Items	Difficult Items
1	86	72	84	80
2	80	64	68	68
3	89	80	76	72
4	76	66	82	80
5	70	68	76	70
6	84	76	74	73
7	82	64	82	80
8	90	73	72	75
9	70	67	76	70
10	84	76	80	78
11	76	81	76	69
12	82	75	84	78
13	82	76	68	75
14	73	61	86	84
15	84	72	81	76
16	80	76	79	80
17	76	76	70	76
18	82	62	71	67
19	74	85	87	83
20	69	78	78	80
$\bar{X}$ (MEAN)	79.45	72.40	77.50	75.70
S_{N-1} (*SD*)	6.19	6.74	5.80	5.09

where r_{DE} is the correlation coefficient between the mean confidence ratings for difficult and easy items calculated on the 20 paired means. The difference is not significant.

We now proceed to comparisons among means. For the between-subjects design we use the t test for independent samples

$$t(38) = (79.45 - 72.40)/2.046 = 3.44,$$

which is highly significant ($p < .005$). The difference between means for the within-subjects design is tested by means of a t test for correlated samples, which yields

$$t(19) = (77.5 - 75.7)/.884 = 2.03$$

suggesting a significant difference ($.025 < p < .05$). An eyeball inspection suggests that the difference between the easy and the difficult items is much

larger in the between design as compared with the within design (the latter difference was nevertheless significant due, as mentioned below, to the controlled between-subjects variance).

Finally, we want to compare the variance of estimates across the two designs, that is the precision of the estimates of the two designs. The purpose of such a comparison is related to the question of whether the design had an effect. Using the same F test for independently estimated variances used in (10), we obtain

$$F(38, 19) = (2.046/.884)^2 = 5.36$$

which is significant at the .05 level. The reason that the precision of the within-subjects design is significantly larger (i.e., the standard error of the within design is smaller) is due to the fact that the scores on the easy and difficult items are positively correlated, thus suggesting that under those circumstances (see Equation 8) individual differences in the within design are better controlled.

The difference between the easy and the difficult items obtained by the between- and the within-subjects designs (i.e., the interaction between item difficulty and design) is tested by a t test

$$t = (7.05 - 1.8)/\sqrt{2.046^2 + .884^2}$$

where the denominator consists of the pooled standard errors of the two designs. Unfortunately, as noted by Rosenthal and Rubin (1982), this statistic is distributed neither normally nor exactly as a t distribution, though it is quite similar to the latter. When the *df* for both studies are sufficiently large, we may use the standard normal distribution that will provide an adequate estimate for all practical purposes.

When the samples for both studies are rather small, the appropriate number of *df* will lie somewhere between the *df* of the smaller of the two studies and the *df* of the two studies added together. Zwick (chap. 2, vol. 2) shows how the number of degrees of freedom in such cases can be approximated, and cites several statistical packages in which such an approximation procedure is built in.

Returning to our example and using Equations 9 and 10 in the chapter by Zwick (in vol. 2), yields an approximation of 19 *df*, and thus the t statistic computed earlier ($t = 2.35$) suggests that the difference between easy and difficult items was significantly larger ($p = .015$) in the between-subjects design compared with the within-subjects design.

In general, two different sorts of explanations can account for cases where treatment effects of between- and within-subjects designs differ significantly (as has been the case in our example). One class of explanations is *methodological* in nature; Rosenthal and Rubin discussed three possibilities under this category:

- *Sampling effects:* Occur when subjects are not assigned randomly to each of the two designs. Under such circumstances subjects assigned to one design may differ from subjects in the other design on a relevant factor.

- *Laboratory effects:* These may occur if the two design studies have been conducted in different places or with different experimental procedures.
- *Sequence effects:* Can occur under within-subjects designs, as mentioned. Counterbalancing order of presentation is not necessarily a sufficient safeguard against such effects.

A second sort of explanation is more *substantive* in nature and relates to the particular hypothesis or theory being tested. For example, the results in our example can be interpreted to mean that subjects tend to anchor on a certain subset of confidence ratings, and each rating is assessed relative to this anchor. In the between-subjects design, according to this interpretation, there is a different anchor for the easy and difficult items, whereas in the within-subjects design where items are mixed, the anchor is based on a mixture of easy and difficult items. The explanation based on an alleged process of anchoring can be classified as a context effect. According to the claim, in this example, the context effect was different under the two designs yielding different patterns of results.

CONCLUSIONS

The decision of whether to use a within- or between-subjects design has many facets and as such constitutes a multiattribute decision. There is obviously no algorithm that will provide the researcher with a definitive answer. The purpose of this chapter was to briefly review the different considerations (some of which have often been neglected) that should be taken into account. Under certain circumstances one may seriously consider using both designs despite the additional costs associated with such a decision. It is important to remember that often within- and between-subjects designs are simply addressing a different question. The first step therefore should always be an unambiguous statement of the research questions to be answered. Subsequent decisions regarding the appropriate design should take into account the different dimensions discussed in this chapter, though the final judgment will often remain subjective.

REFERENCES

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegner (Ed.), *Social attitudes and psychological measurement* (pp. 401–485). Hillsdale, NJ: Lawrence Erlbaum Associates.

Birnbaum, M. H., & Mellers, B. (1983). Bayesian inference: Combining base rates with reports of sources. *Journal of Personality and Social Psychology, 45*, 792–804.

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within- and between-subjects manipulations of the independent variable. *Psychological Bulletin, 84*, 212–219.

Estes, W. K. (1964). Probability learning. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press.

Glass, G. V., & Stanley, J. C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs: Prentice-Hall.

Greenwald, A. G. (1976). Within-subjects design: To use or not to use. *Psychological Bulletin, 83*, 314–320.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston.

Hershey, J. C., & Schoemaker, P.J.H. (1980). Prospect theory's reflection hypothesis: A critical examination. *Organizational Behavior and Human Performance, 25*, 395–418.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.

Keren, G., & Raaijmakers, J.G.W. (1988). On between-subjects versus within-subjects comparisons in testing utility theory. *Organizational Behavior and Human Decision Processes, 41*, 233–247.

Keren, G., & Wagenaar, W. A. (1987). Violation of utility theory in unique and repeated gambles. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 387–391.

Meyers, J. L. (1979). *Foundations of experimental design* (3rd ed.). Boston: Allyn & Bacon.

Parducci, A. (1965). Category judgment: A range frequency model. *Psychological Review, 72*, 407–418.

Parducci, A. (1983). Category ratings and the relational character of judgment. In H. G. Geissler, H.F.J.M. Buffart, E.L.J. Leeuwenberg, & V. Saris (Eds.), *Modern issues in perception* (pp. 89–105). Berlin: VEB Deutche Verlag der Wissenschaften.

Poor, D.D.S. (1973). Analysis of variance for repeated measures designs: Two approaches. *Psychological Bulletin, 80*, 113–121.

Poulton, E. C. (1973). Unwanted range effects from using within-subjects experimental designs. *Psychological Bulletin, 81*, 201–203.

Rosenthal, R. (1976). *Experimenter effects in behavioral research*. New York: Appelton-Century-Crofts.

Rosenthal, R., & Rubin, D. (1980). Comparing within- and between-subjects studies. *Sociological Methods and Research, 9*, 127–136.

Rothstein, L. D. (1974). Reply to Poulton. *Psychological Bulletin, 81*, 199–200.

Wedell, D. (1990). Methods for determining the locus of context effects in judgments. In J. P. Caverni, J. M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases* (pp. 285–302). Amsterdam: North Holland.