ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 41, 233–247 (1988)

On Between-Subjects versus Within-Subjects Comparisons in Testing Utility Theory

GIDEON B. KEREN AND JEROEN G. W. RAAIJMAKERS

TNO Institute for Perception, Soesterberg, The Netherlands

Empirical studies of expected utility theory often employ a between-subjects design. This practice has been recently criticized by J. C. Hershey and P. J. H. Schoemaker (1980, *Organizational Behavior and Human Performance*, **25**, 395–418). The present paper provides a critical analysis of the controversial issues concerning the use of between-subjects vs within-subjects comparisons. It is claimed that the choice of experimental design should be determined, among other things, by theoretical aspects (e.g., the interpretation of utility theory) and the nature of the scientific problem. Following, we present relevant psychological considerations and conclude that, in the context of testing utility theory, a between-subjects design will often be more desirable. We then describe three different hypotheses that a researcher may be interested in testing, and identify the appropriate design for testing each of these hypotheses. The relationships between the different hypotheses are discussed. We apply our framework to reanalyze the reflection effect and compare it with the analysis proposed by Hershey and Schoemaker (1980). Methodological implications for future research are briefly discussed.

A common question faced by researchers in the behavioral sciences concerns the use of a between-subjects or within-subjects design. Greenwald (1976) proposed that the decision as to which type of design should be employed depends on two types of considerations: psychological and statistical. Notwithstanding, we suggest an additional consideration, namely, the nature of the theory that is being tested and the particular scientific problems that are supposed to be answered by the experiment (Grice, 1966).

The impetus for the present article stems from a recent paper by Hershey and Schoemaker (1980) that criticized the practice of comparing preferences across (between) rather than within individuals. In particular, these authors claimed that the "reflection" effect, proposed by Kahneman and Tversky (1979), can be tested meaningfully only in a within-subjects design. The purpose of the present article is to evaluate the relative merits of between- vs within-subjects designs. Although most parts

This work represents equal share of both authors. Requests for reprints should be addressed to Gideon Keren, TNO Institute for Perception, PO Box 23, 3769 ZG Soesterberg, The Netherlands.

0749-5978/88 $3.00

of the discussion may have broader implications, we limit ourselves only to studies that are designed to test utility theory. In particular, we focus on two recent studies that were designed to test the adequacy of utility theory: One study involves the demonstration of the so-called "reflection" effect (Kahneman & Tversky, 1979); the other study concerns the comparison of unique vs repeated gambles (Keren & Wagenaar, 1987).

It is important, especially in the present context, to distinguish between two types of a within-subjects design. 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 (exactly) the same stimulus more than once. Major reasons to employ such a design are the gain in statistical power and economy in the use of subjects.

In the other type of design, the same subject may experience the same (or very similar) stimulus on several trials. Such a design may be used when the researcher's explicit goal is to investigate learning effects due to repeated trials. If practice and learning effects are not desired, such a design can be used only if there are sufficient reasons to assume that, for all practical purposes, the trials are independent as, for example, in psychophysical experiments. Such an assumption of independence should be adopted with caution, and evidence favoring such an assumption should usually be provided explicitly. The major concern of the present article is with this latter class of designs.[1] In the next section we describe two experiments that were both conducted originally in a between-subjects design, and (following Hershey & Schoemaker, 1980) pose the question whether a within-subjects test would not be more appropriate. In the following section we provide a brief discussion on the fundamental nature of utility theory. We claim that different experimental comparisons (i.e., between or within) are not just distinct from a methodological viewpoint, but should also be interpreted differently within a given theoretical framework. Subsequent sections deal with psychological considerations, and with methodological and statistical issues concerning the use of the two experimental designs. The final section provides a brief summary and implications for future research, some of which may be extended beyond the present context of utility theory.

ON INTERPRETING EXPERIMENTAL TESTS OF UTILITY THEORY

Kahneman and Tversky (1979) presented a set of experiments that were designed to point out the inadequacy of expected utility theory.

[1] The difference between the two types of designs is continuous in nature depending on the degree to which repeated stimuli are similar. The major focus in the present article is on repeated stimuli that are either identical or extremely similar.

Among the pervasive effects demonstrated by Kahneman and Tversky were the so-called certainty effect (i.e., people overweigh certain outcomes relative to probable outcomes) and the possibility effect (i.e., when winning is possible but not probable, people prefer the gamble that offers the larger gain). Both effects were demonstrated for positive and negative outcomes. The results reported by Kahneman and Tversky (1979) showed that preference between gambles with negative outcomes was the mirror image of the preferences between gambles with positive outcomes. This pattern of results was labeled the reflection effect.

Recently, Hershey and Schoemaker (1980) have questioned the generality of the reflection hypothesis. They assert that the between-subjects comparison used by Kahneman and Tversky (1979) constitutes only a very crude approximation to individual reversals, and that the proper design for testing reflection should be within subjects. To further support their claim they reanalyzed the data reported by Kahneman and Tversky (1979) and alleged that they are inconclusive with regard to reflection. In addition, they conducted three separate studies using a within-subjects design. According to Hershey and Schoemaker (1980), the results obtained from these three experiments seriously question the generality of the reflection effect. To what extent are the claims by Hershey and Schoemaker justified? In this section we examine their claims in the light of theoretical considerations concerning the interpretation of utility theory.

Expected utility theory is formally a theory of individual choice behavior. Originally, it was developed as a prescriptive theory, justified only by normative considerations (Coombs, Dawes, & Tversky, 1970). Consequently, even at later stages when the theory was modified and extended to serve also as a descriptive theory, it maintained a deterministic flavor. Empirical evidence, however, unequivocally suggests that the theory fails as a deterministic descriptive theory. Observed inconsistencies in preferences are assumed to be randomly distributed (Coombs, *et al.*, 1970; Luce & Raiffa, 1957) and led Luce (Luce & Raiffa, 1957) to propose a probabilistic theory of utility. This theory adopts in essence the fundamental assumptions of utility theory as formulated by von Neuman and Morgenstern (1947) except for some modifications that convert the theory from deterministic to probabilistic. In any event, as suggested by Schoemaker (1982), despite the attractiveness of expected utility theory its structural validity at the individual level is highly questionable.

In light of the above comment we now examine the exact meaning of the reflection hypothesis. A careful reading of Kahneman and Tversky (1979) suggests that unlike the certainty and possibility effects, both of which should be considered as psychological phenomena (constructs), the reflection effect is simply a label that conveniently describes a certain

pattern of results. At no point do Kahneman and Tversky (1979) claim the existence of a reflection mechanism for each subject, in which preferences in the negative domain are obtained by reflecting the preferences in the positive domain.[2] All that the reflection effect implies is that risk aversion in the positive domain is accompanied by risk seeking in the negative domain. In accordance with the above discussion this relationship should be interpreted in a probabilistic manner. Hence, risk aversion in the positive domain does not necessarily imply risk seeking in the negative domain. In other words, deviation from the modal behavior in the positive domain does not necessarily have to lead to deviation from the modal in the negative domain. Our fundamental disagreement with Hershey and Schoemaker (1982) is that they assume that reflection is a deterministic phenomenon and thus expect that every subject who is risk averse in the positive domain should be risk prone in the negative domain and vice versa. We, in contrast, assert a probabilistic relationship: The prediction we derive from the reflection phenomenon is that the modal choice will reverse when changing from gambles with positive to negative outcomes. The difference in the meaning attached to the reflection effect leads to different methodological requirements: The interpretation that is supposedly offered by Hershey and Schoemaker (1982) entails that the existence of the reflection effect can be meaningfully tested only by a within-subjects comparison. In contrast, the interpretation offered by us as derived from Kahneman and Tversky (1979) implies that a between-subjects comparison is equally appropriate. Unlike Hershey and Schoemaker, we conclude that a between-subjects design is perfectly acceptable for testing the reflection effect. There are, of course, other considerations that should enter the decision of whether to employ a between- or within-subjects comparison, and these are presented in later sections.

Issues very similar to those discussed above are also applicable to a recent study by Keren and Wagenaar (1987), in which they investigated subjects' preferences in unique and repeated gambles. Specifically, these researchers used gambles similar to those employed by Kahneman and Tversky (1979) to demonstrate the certainty and possibility effects. Keren and Wagenaar wanted to test whether the violations of utility theory, reported by Kahneman and Tversky, would generalize to repeated conditions. They employed a between-subjects design and presented one group of subjects (unique condition) with a choice between two alternative gambles. The chosen gamble was to be played a single

[2] In describing the empirical results, Kahneman and Tversky (1979) explicitly state: "We label this pattern the *reflection effect*" (p. 268). Later, when they develop prospect theory, they use this empirical finding in order to describe the nature of the value function. They never propose reflection as a deterministic mechanism. Rather, like other characteristics of the value function, reflection should be interpreted in a probabilistic manner.

time. A second group of subjects (repeated condition) was presented with exactly the same alternatives, except that they were told that the chosen gamble would be played 10 times in succession. The pattern of results for unique and repeated gambles was quite different. Whereas the majority of the subjects in the unique conditions made choices that were incongruent with utility theory, thus replicating Kahneman and Tversky (1979), the majority of the subjects in the repeated conditions did not exhibit such violations. Consequently, Keren and Wagenaar suggested that utility theory may not be violated under repeated conditions.

Following the line of reasoning offered by Hershey and Schoemaker (1980), one could argue that the conclusion proposed by Keren and Wagenaar (1987) is premature without using a within-subjects comparison. Specifically, according to this argument, if a within-subjects design had been employed, then it would be possible to compute the probability that a subject would violate expected utility under the repeated condition, given that the expected utility was violated under the unique condition. As in the case of the reflection effect, there is an underlying assumption here of an individual reflection mechanism, that translates choices in the unique condition to choices in the repeated condition. Keren and Wagenaar never proposed such a mechanism. Their only goal was to demonstrate that unique and repeated gambles yield different patterns of results, and consequently warned researchers to be careful in making generalizations from unique to repeated gambles. For this purpose, we claim, the use of a etween-subjects comparison is perfectly adequate.

METHODOLOGICAL CONSIDERATIOINS REGARDING PSYCHOLOGICAL ASPECTS

Generally, within-subjects designs are often appealing because of their high statistical power and economy in number of subjects. However, as was noted by several researchers (e.g., Greenwald, 1976; Grice, 1966; Poulton, 1973), they may often be vulnerable to several unwanted psychological effects. Below we present a brief discussion of such possible effects, in particular with regard to the present context concerning experimental tests of utility theory.

Within-subjects designs may often lead to carry-over effects (Greenwald, 1976; Poulton, 1973), which are unwanted unless the researcher wishes to investigate the effect of practice. These effects are especially salient when repeated stimuli are identical or highly similar. A likely effect, in the context of the present paper, of such similar repeated stimuli is to evoke the subjects' awareness to provide consistent responses (even if they do not reflect the true preferences). For instance, consider the experiments conducted by Keren and Wagenaar (1987) to investigate

subjects' preferences concerning unique and repeated gambles. Subjects were required to choose between two gambles and were assigned (according to a between-subjects design) to either the unique or repeated condition. With the exception of one dimension (number of times the gamble is played), the stimuli in the two conditions are in fact identical. The similarity between the unique and repeated options is so transparent that if a within-subjects design had been used, whatever subjects' preferences were, they could make sure that their choices in the two conditions were not contradictory. In other words, the preferences exhibited in such an experiment using a within-subjects design will be highly correlated.

Such a problem is, of course, not unique to investigations of utility theory. For example, consider the experiments of Milburn (1978). Milburn had subjects estimate the likelihood of (among others) positive events occurring in each of four successive decades. Using a within-subjects design where each subject judged the likelihood of the *same* events for each of the four decades, it was shown that the positive events were predicted as increasingly likely over time. However, using a between-subjects design, this pattern of results could not be replicated, and the likelihood estimates were shown to decrease.

A common way to guard against possible carry-over effects is to counterbalance the different conditions. However, in the more limited context of preferences the possible confounding, due to the need to be consistent, cannot be eliminated by such counterbalancing. A more efficient way would be to present the two conditions (e.g., unique vs repeated) at different points in time, or to embed the two different choice conditions among other different experimental tasks. Adopting such procedures may be tedious, but they will eventually reduce the unwarranted effects, though not necessarily eliminate them completely.

A second, and related, aspect concerns the sensitization to treatment variations that may lead subjects to form hypotheses about the treatment effect. The larger the similarity betwen treatments, the more likely it is that such a sensitization may take place. For instance, to test the reflection hypothesis (Kahneman & Tversky, 1979) in a within-subjects design, one has to present subjects with identical problems that differ only in that one is expressed in gain and the other in loss prospects. It is quite likely that such a design would result in subjects forming hypotheses concerning the Experimenters' intentions (Rosenthal, 1976). Independent of how subjects resolve the problem, it is clear that their responses may not reliably reflect their natural preferences. Again, as in the previous case, the unwanted effects can be reduced by a careful, judicious design and experimental procedures. However, such precautions will be less effective the higher the similarity between the repeated stimuli, as in the experimental procedures for testing the reflection effect, and unique vs repeated gambles.

Another consideration concerns the ecological or external validity of a study (Campbell & Stanley, 1966; Greenwald, 1976). Ideally, one would like to employ a representative design (Brunswik, 1956) that would most resemble the natural environment. For example, consider again the study by Keren and Wagenaar (1987) that was aimed at exploring possible differences between unique and repeated gambles. Under real-life circumstances people are faced with decisions concerning unique (e.g., should I undergo an operation? Should I resign from my current job?) as well as repeated (e.g., should I invest in bonds or stocks? Should I insure my house again?) events. People, however, are not faced with both unique and repeated decisions regarding the same situations at the same time (or even during a period of time). In reality, being faced with unique and repeated gambles at the same time is a logical impossibility. Thus, the use of a within-subjects design in such a case would be artificial and nonrepresentative.

In summary, we have pointed out several psychological drawbacks involved in the use of within-subjects designs. Several procedures exist to reduce these undesired effects, but the success of such precautions may be limited, in particular for highly similar repeated stimuli. Surprisingly, Hershey and Schoemaker (1982) did not use *any* such safeguards, which may account for the fact that their results were somewhat different from those obtained by Kahneman and Tversky (1979). Indeed, Budescu and Weiss (1985) employed a within-subjects design, and using appropriate precautionary measures, obtained results similar to those reported by Kahneman and Tversky (1979).

METHODOLOGICAL AND STATISTICAL ISSUES

Given the limitations of within-subjects designs, the question still remains as to what valid conclusions can be drawn from a between-subjects experiment. In this section an attempt is made to answer this question in the limited context of choice behavior.

Following Kahneman and Tversky (1979), a prospect $(x_1p_1; \ldots; x_np_n)$ is defined as a contract that yields outcome x_i with probability p_i where $\Sigma_{i=1}^{n} p_i = 1$. Further, to simplify notation, the prospect of obtaining x_i with probability p_i and y_i with probability $1 - p_i$ is denoted by (x_i,p_i,y_i) where the complementary probability is simply omitted. Now consider experimental situations in which subjects are presented with a pair of prospects A_i and B_j of the form

$$A_i = (x_i,p_i,y_i)$$

and (1)

$$B_j = (x_j,p_j,y_j).$$

The subjects' task is to indicate which of the two prospects they prefer.

Assume an experimenter who wants to study two different experimental conditions, each represented by a pair of prospects, say A1, B1 in Condition 1 and A2, B2 in Condition 2. The prospects in the two conditions are constructed such that for any subject who follows the dictates of expected utility, if she or he prefers prospect A in one condition, he/she will also prefer prospect A in the second condition. Similarly, subjects who prefer prospect B in one condition will also prefer prospect B in the second condition. Put differently, the choice of prospect A in one condition, and prospect B in the other condition would be inconsistent with utility theory.

The two experimental conditions may stand for prospects with positive and negative outcomes, i.e., testing the "reflection" hypothesis (Kahneman & Tversky, 1979), or they may constitute unique vs repeated gambles as in Keren and Wagenaar (1987), or they may represent certain against uncertain outcomes for testing the "certainty" effect (Kahneman & Tversky, 1979).

Following Hershey and Schoemaker (1980), we present the four possible preference combinations people may exhibit (ignoring indifference cases) in such a hypothetical experiment in a 2 × 2 design as shown in Table 1. Note that in practice, for most of the experiments employing the design outlined in Table 1, the two conditions are usually unrelated. This is because the prospects are designed such that A1 and A2 are highly similar as are B1 and B2 (e.g., Kahneman & Tversky, 1979; Keren & Wagenaar, 1987). It is because of this built-in similarity that the choice of A1 and B2, or of A2 and B1, are interpreted as exhibiting inconsistency.

Without loss of generality, one may consider three different hypotheses concerning the possible outcomes of such a hypothetical experiment. To simplify the notation, let $P(A1)$ denote the probability of choosing $A1$ given a choice between $A1$ and $B1$. Similarly for $P(B1)$, etc.

$$\begin{aligned} &\text{H1:} && P(A1) > 1/2 \text{ and } P(B2) > 1/2 \\ &\quad \text{or alternatively} && P(B1) > 1/2 \text{ and } P(A2) > 1/2 && (2) \\ &\text{H2:} && P(A2|A1) < P(A2|B1) \\ &\quad \text{or equivalently} && P(B2|A1) > P(B2|B1) && (3) \\ &\text{H3:} && P(A1 \cap B2) + P(A2 \cap B1) > 1/2. && (4) \end{aligned}$$

Note that these three hypotheses represent the major examples of three different classes of hypotheses, H1 being formulated in terms of marginals or average preferences, H2 in terms of statistical dependencies, and H3 in terms of the number of individual reversals. Other hypotheses are, of course, also possible: for instance, that more than 50% of those subjects that chose A1, chose B1. We believe, however, that the three hypotheses described above are the most important ones.

TABLE 1
THE POSSIBLE PREFERENCE COMBINATIONS IN A 2 × 2 DESIGN

	Condition 1		
	*A*1	*B*1	
Condition 2			
*A*2	n_1	n_2	*a*
*B*2	n_3	n_4	*b*
	c	*d*	*N*

Despite the apparent similarity among the three hypotheses, in particular that they are all in violation of expected utility theory, they are nevertheless distinctly different. Moreover, no hypothesis necessarily implies the truth of either of the other two. A proof of this assertion is given in the appendix. It is therefore essential, in our opinion, that researchers of choice behavior who adopt a design similar to that shown in Table 1 should specify explicitly which particular hypothesis they are testing or referring to. The first hypothesis refers to (average) *group* preferences. It claims that the pattern of preferences of subjects in Conditions 1 and 2 is incongruent with expected utility theory. In terms of the design presented in Table 1, it is a hypothesis about marginal probabilities and, as such, may naturally be tested in a between-subjects design.

The second hypothesis clearly refers to *individuals,* and can be tested only in a within-subjects design. This hypothesis is formulated in terms of conditional probabilities. It is important to note that this hypothesis compares the choices of people who initially have different preferences (i.e., in the first condition). An alternative interpretation of H2 is in terms of deviations from independence. Note, however, that deviation from independence in Table 1 can be established in two ways: One is by demonstrating that $P(A2|A1) < P(A2|B1)$, that is, by finding support for H2. A second is by demonstrating that $P(A2|A1) > P(A2|B1)$, that is, by affirming the opposite hypothesis. In that respect, H2 is a directional hypothesis.

The third hypothesis is explicitly formulated in terms of reversals (that are incongruent with utility theory). It predicts that the majority of *individuals* in a given population will exhibit reversals. Strictly speaking, as in the case of H2, H3 can only be tested in a within-subjects design.

Despite their commonalities, H2 and H3 are quite different. In particular, conditions may exist in which H3 is not satisfied but H2 is, and in fact testing H3 would not be very sensible. Conditions like that exist when there is an initial strong preference for one of the two prospects (A or B) in both conditions, as reflected in the marginal probabilities. An example is provided in Table 2. In that example there is an overwhelming

TABLE 2
POSSIBLE PREFERENCE COMBINATIONS SATISFYING H2 BUT NOT H3

	Condition 1		
	*A*1		*B*1
Condition 2			
*A*2	5	95	100
*B*2	95	805	900
	100	900	1000

preference for prospect B in both Conditions 1 and 2, which immediately rules out the possibility that either H1 or H3 will be satisfied. Nevertheless, simple calculations indicate that H2 is satisfied.

What are the relationships between the three hypotheses under discussion? Earlier, we proposed that in the context of choice behavior, and in particular for testing utility theory, within-subjects designs would most often be methodologically contaminated and should therefore be avoided. The most interesting question, therefore, is what inferences can be drawn from H1 (tested in a between-subjects design) to H2 or H3 (that are supposed to be tested in a within-subjects design)?

Since H1 is a hypothesis about marginals, and H2 is formulated in terms of conditional probabilities, no direct inferences can be drawn from H1 to H2. In particular, in terms of Table 1, the first alternative of H1 can be rewritten as

$$n_1 + n_3 > n_2 + n_4$$

and (5)

$$n_3 + n_4 > n_1 + n_2.$$

In a similar manner, H2 can be written as

$$n_1 \cdot n_4 < n_2 \cdot n_3 \quad (6)$$

It is immediately evident that the set of requirements in (5) and (6) are unrelated.

A closer relationship exists between H1 and H3, although they are not perfectly related. In the following section we focus the discussion on the reflection hypothesis and examine the analysis performed by Hershey and Schoemaker (1980). In this context we also discuss the possible inferences one could make from H1 to H3.

THE REFLECTION HYPOTHESIS: A REANALYSIS

In the first section of this article we claimed that a careful reading of Kahneman and Tversky (1979) suggests that the most meaningful inter-

pretation of what they labeled as the "reflection effect" is in terms of what we called here Hypothesis 1. It is, therefore, not surprising that these authors have used a between-subjects design.

Notwithstanding previous considerations, we may still wonder what would be the case if the reflection effect is to be interpreted in terms of Hypothesis 2 or 3?

Hershey and Schoemaker (1980) criticized the evidence from the between-subjects studies of Kahneman and Tversky (1979) by noting that a given pattern of overall preferences could be consistent with varying numbers of individual reversals. In particular, they showed that the number of individual reversals could be much lower than 50%, assuming within-subject data with the same overall preferences. In terms of Table 1, for fixed values of the marginals a, b, c, and d, there are a number of combinations of n_1, n_2, n_3, and n_4 that are consistent with these marginals.

For example, suppose that in Table 1, $a = 60$, $b = 40$, $c = 40$, and $d = 60$. Such a pattern of results would (in a between-subjects study) be regarded as supporting the reflection hypothesis. However, they could have been obtained from $n_1 = 35$, $n_2 = 25$, $n_3 = 5$, and $n_4 = 35$ (i.e., 30% reversals) but also from $n_1 = 5$, $n_2 = 55$, $n_3 = 35$, and $n_4 = 5$ (i.e., 90% reversals). Assuming, as Hershey and Schoemaker do, a criterion of at least 50% reversals, the first situation would not be supportive for the reflection hypothesis while the second would.

Clearly, if H3 is the hypothesis to be tested (i.e., 50% or more reversals), this is an unsatisfactory situation. Hershey and Schoemaker (1980) therefore propose to use a criterion based on the *minimum* number of reversals that could have been obtained in a within-subjects design with the same marginal results. To support the hypothesis of within-subject reflectivity, they suggest that the minimum number of reversals should be significantly larger than 50%.

The analysis proposed by Hershey and Schoemaker may be questioned on several grounds. In particular, the analysis does not take into account the probability of actually obtaining the minimum number of reversals given the marginals. That is, although the minimum number may be smaller than 50%, the probability for such an extreme result may be very small indeed. A more acceptable approach would be to calculate the probability of obtaining less than 50% reversals given the observed marginal probabilities. The hypothesis of within-subject reflectivity would be supported if this probability turns out to be smaller than, for instance, conventional significance levels (i.e., $\leq .05$).

In order to derive this probability, we will make use of a somewhat simplified model that is consistent with the Hershey and Schoemaker approach. We will assume that each subject has a probability r of reversing

his preferences. Without loss of generality, it may be assumed that $P(A1) > .5$. The probability of choosing $A2$, $P(A2)$, is then equal of

$$P(A2) = (1 - r)P(A1) + r[1 - P(A1)].$$

It is easy to see that $P(A2) < .5$ if and only if $r > .5$. Hence, to test the hypothesis $r > .5$ (50% or more reversals), it suffices to test whether $P(A2) < .5$.

In our numerical example (which is in fact taken from the data of Kahneman and Tversky, 1979) the minimum value of $n_2 + n_3$ is 28. Since this is smaller than 50%, this would not be regarded as supporting the within-subject reflectivity hypothesis by Hershey and Schoemaker (1980). However, using the above analysis, it can be calculated that a value of $c = 26$ would be significant at the 5% level. Hence, since the observed value of c is much smaller, we would arrive at a different conclusion with respect to the hypothesis of reflectivity. In fact, Hershey and Schoemaker concluded that only two out of the five examples of Kahneman and Tversky supported this hypothesis. Using our statistical analysis, and a significance level of .05, only one example (i.e., Problem 4 of Kahneman & Tversky, 1979) failed to support the reflection hypothesis.

It should be noted that the above discussion is predicated on the assumption that H3 is indeed the most appropriate hypothesis. Given that this is true (which we do not believe) and given (because of the psychological reasons discussed earlier) the difficulties involved in a within-subject design, our method may be used to infer what may be reasonably concluded from the results of a between-subjects study with respect to H3. However, in many cases, H3 may not be the most appropriate formulation.

CONCLUSIONS

We have discussed a number of issues concerning within- and between-subjects designs in testing expected utility theory, and the nature of the conclusions that can be drawn from studies using either type of design. Previous articles (e.g., Hershey & Schoemaker, 1980) have exclusively focused on statistical and methodological issues. Not undermining the importance of these considerations, we claim that psychological considerations are equally important. Especially, in the context of testing utility theory, one should be cautious in drawing conclusions from within-subjects designs in which the same stimuli (i.e., a choice between the same gambles) are presented more than once. Under certain circumstances, the use of such a within-subjects design would simply be inappropriate.

In addition, the choice of design should be related to the nature of the

hypothesis that one is interested in. We have discussed three types of hypotheses that might be relevant for choice studies. One of these (H1) does not require a within-subjects design. It is recommended that investigators explicitly specify which hypothesis best captures their intention.

The choice of hypothesis to be tested and the corresponding design to be used should also depend on the theoretical framework that is adapted. In the present context, it was proposed, two different interpretations of expected utility theory may lead to a different choice of experimental design. In particular, adapting a deterministic interpretation of utility theory (like Hershey & Schoemaker, 1980) entails that the only meaningful and valid test of the theory (independent of other considerations) is obtained by the use of within-subjects comparisons. In contrast, the view expressed by the present authors (and supposedly shared by Kahneman & Tversky) assumes a probabilistic framework. Under such assumptions testing H1, and accordingly employing a between-subjects design, would be a most natural and reasonable choice.

However, even if H1 is not the preferred hypothesis, a between-subjects design may still be the only one that is feasible due to the psychological factors discussed above. The question then arises what inferences might be drawn from such between-subjects data with respect to within-subject behavior. We have shown that no inferences can be drawn with respect to H2 since the marginal choice probabilities give no information with respect to the (in)dependence of the choices on the individual level. For H3 the situation is somewhat different. Although firm conclusions are a bit dangerous, a simple statistical analysis makes it possible to test whether the probability of a preference reversal is higher than 50%. Using this test, it was shown that the analysis presented by Hershey and Schoemaker (1980) is too conservative and does not do justice to the strength of the evidence that is available in the marginal probabilities.

APPENDIX

Proof of the assertion that none of the three hypotheses, H1, H2, or H3, implies the other two. We prove the assertion by presenting a number of counterexamples.

1.

	*A*1	*B*1	
*A*2	.3	.0	.3
*B*2	.3	.4	.7
	.6	.4	

This example shows that H1 does not imply H2 or H3.

2.

	*A*1	*B*1	
*A*2	.6	.2	.8
*B*2	.2	.0	.2
	.8	.2	

This shows that H2 does not imply H1 or H3.

3.

	*A*1	*B*1	
*A*2	.3	.3	.6
*B*2	.3	.1	.4
	.6	.4	

This shows that H3 does not imply H1.

4.

	*A*1	*B*1	
*A*2	.2	.0	.2
*B*2	.7	.1	.8
	.9	.1	

This shows that H3 does not imply H2.

REFERENCES

Brunswik, E. (1956). *Perception and the representative design of experiments*. Berkeley, CA: Univ. of California Press.

Budescu, D., & Weiss, W. (1985). Reflection of transitive and intransitive preference: A test of prospect theory. IPDM Report No 29, University of Haifa, Israel.

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

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology*. Englewood Cliffs, NJ: Prentice–Hall.

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

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

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 decisions under risk. *Econometrica*, **47**, 263–291.

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

Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. New York: Wiley.

Milburn, M. A. (1978). Sources of bias in the prediction of future events. *Organizational Behavior and Human Performance,* **21,** 17–26.

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* (2d ed.). New York: Wiley.

Schoemaker, P. J. H. (1982). The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature,* **20,** 529–563.

RECEIVED: June 4, 1986