CHAPTER 25

# Designing online experiments

Michael H. Birnbaum

## Theory

A psychological theory is a set of statements that satisfies five philosophical criteria. The first criterion is that the proposed explanation can be used to deduce the behavioural phenomena to be explained. For example, suppose someone asks, 'Why is bread good to eat?' We use operational definitions of 'bread' and 'good to eat' in order to link these concepts to the world of observations. Define a substance as 'good to eat' if more than 75 per cent of 1-year-old rats given access to water and that substance under standard lab conditions survive on this diet for six months. Define 'bread' as what comes from the market in a package marked 'Weber's bread'.

The following argument illustrates deduction:

P1: Bread is made of cyanide

P2: Everything made of cyanide is good to eat

C: Bread is good to eat.

If both premises are true, then the conclusion follows by logic. This example illustrates deduction. It also shows that one can deduce a true conclusion from false premises. Thus, a true conclusion does not 'prove' the premises to be true. However, if the conclusion is false, at least one of the premises must be false. For this reason, we speak of *testing* theories rather than 'proving' them.

The second criterion of a theory is that it should not be meaningless. The empirical meaning of a statement is equivalent to the set of specifiable, testable implications. If a statement has no testable implications, then it is devoid of empirical meaning. Unfortunately, many so-called theories, like psychoanalysis, are built on entities such as subconscious conflicts that cannot be observed empirically. By definition, the contents of a mind are private, which means that only one mind can observe the contents of that mind. By definition, the contents of the subconscious mind cannot be observed by that mind itself. Unless concepts are linked to events and objects that can be observed, measured, or tested, they fall outside the world of science and into the pages of poetry. So, if we theorize that all rats have subconscious conflicts and whenever an organism has a subconscious conflict, bread is good to eat, we have a deductive theory, but one that is meaningless, except for its conclusion.

The third criterion of a theory is that it is predictive. In principle, if one knew the theory in advance, one could have predicted the behaviour or events to be explained. A system that is deductive, meaningful and predictive is a called a *predictive system*. An example of a predictive system is Kepler's laws of astronomy. Kepler postulated that planets travel in elliptical orbits around the sun with the sun at one focus and that they sweep out equal areas in the ellipse in equal time. He also assumed that the squares of the periods of revolution are proportional to the cubes of the average distance from the sun.

From these three assumptions and geometry, one can make many predictions of the future. For example, one can accurately predict the positions of Sun, Moon, Venus, Mars, Jupiter and Saturn as seen against the stars for any date in the next thousand years, predict when these planets will go into retrograde motion, and predict eclipses of both sun and moon.

Although a predictive system can be valuable, we usually want more from an explanation than just prediction. The fourth criterion of an explanation is that it should contain a causal argument. That means that the explanation provides a way, in principle, to control the behaviour to be explained. In other words, we can predict the results of manipulation of variables. Whereas correlational relationships allow one to predict ongoing behaviour, it is causal statements that allow us to predict what would happen if we introduced changes in the system. Kepler's astronomy does not predict what would happen if we could change the mass of the sun, for example, but Newton's laws do allow such 'what if' calculations.

The difference between correlation and causation is the difference between prediction and control. Both are useful concepts, but they lead to different uses and they can appear to be opposites. For example, a correlational survey would find that people who received antibiotics last year are more likely to be dead this year than people who received no antibiotics last year. So we can use antibiotics to predict death. However, by means of an experiment, we can randomly assign people with infections to two groups, one that receives antibiotics and the other receives a placebo. The results of such studies show that antibiotics cause a reduction in the death rate. So, we find that receiving antibiotics is positively correlated with death in surveys and receiving antibiotics is negatively correlated with death in an experiment. Although paradoxical, there is no contradiction.

Both correlational and causal relations are interesting and useful, even when they seem to say the opposite things. Suppose you have a life insurance company; you sell insurance that pays out when a person dies. Before you sell someone insurance, you could ask if they have been taking antibiotics. If yes, you do not want to sell them insurance because they are likely to die. However, if you already sold a policy to a client and that person becomes sick, you would like them to take antibiotics because it causes a reduction in the death rate.

Correlation has been called the 'instrument of the Devil' when evidence of a correlation is used to argue for a causal conclusion. For example, it has been shown that students in small classes do worse in high school than students in large classes. What class in a high school is the smallest? It is the class for 'special education' students, students who have behaviour problems or are mentally retarded. So, small classes size is correlated with poor performance. Those who misunderstand this correlation argue that all classes should be large, because larger classes get better performance.

The fifth criterion of an explanation is that it is general. A general explanation for one phenomenon can also be used to explain other phenomena. Put another way, the premises of an explanation have the characteristics of scientific laws, statements that hold in general. This means that good explanations lead to new testable implications. Although we can't change the mass of our sun in practice, we can bring objects of different masses together on earth and measure the forces between them. Newton's laws are considered very general because they can be used to make many predictions for objects in space and on earth including falling bodies, trajectories of cannon balls, collisions and thousands of other calculations useful in mechanical and structural engineering.

## Experiments test among theories

Psychology is the study of alternative explanations of behaviour. The purpose of an experiment is to test between alternative theories. Students sometimes talk about trying to 'prove' a theory, as if they could somehow show that a theory is 'true'. Such thinking leads to bad research. For example, consider a person who thinks that bread is good to eat because it is made of cyanide, and everything made of cyanide is good to eat. To 'prove' the theory, the person eats bread and argues that this 'proves' the theory true, since if the theory is true, bread should be good to eat. I assume the reader can think of some different experiments that would refute these premises.

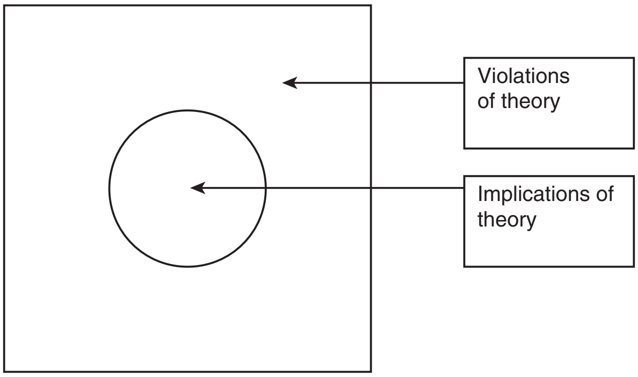

**Figure 25.1** Set representation of results of studies to test a theory. It is argued that people should devise experiments to look outside the circle of implications of the theory.

Consider the diagram of Figure 25.1. Suppose the box represents the universe of all possible results of experimental tests, and the elements inside the circle represent data that would be consistent with the theory. Many researchers conduct experiments to look inside the circle; that is, they look for confirmations of their theory. For example, a person might eat bread to 'prove' their theory. Instead, what researchers should do is devise experiments to test the theory by looking outside the circle; that is, they should for results that would refute the theory. For example, test if bread is really made of cyanide and test if cyanide is really good to eat.

The fallacy that one should try to 'prove' one's theory leads to bad research. To avoid this problem, I suggest that researchers think in terms of comparing at least two theories, and searching for implications to test that would be consistent with one theory and that would refute the other. If both theories qualify as theories of the behaviour in question, one should be able to devise a test that would refute at least one of them.

## Example: expected value theory

Suppose a researcher wanted to test the theory that people evaluate gambles by their expected values. In particular, let $G = (x_1, p_1; x_2, p_2; \mathsf{K}\,; x_n, p_n)$ represent a gamble with $n$ mutually exclusive, exhaustive outcomes, which pay cash prizes of $x_\mathrm{i}$ with probabilities, $p_\mathrm{i}$, where $\Sigma_{i=1}^{n} p_i = 1$. Let $G \succ F$ represent systematic preference for gamble $G$ over gamble $F$. The concept of systematic preference is given the following operational definition: we can reject the hypothesis that the probability of choosing $F$ over $G$ is less than or equal to 1/2 in favour of the hypothesis that the probability of choosing $G$ over $F$ exceeds 1/2.

Now suppose we have the following theory:

$$G \succ F \Leftrightarrow U(G) > U(F) \Leftrightarrow EV(G) > EV(F), \quad (1)$$

$$EV\left(G\right) = \sum_{i=1}^{n} p_i x_i \qquad (2)$$

According to this theory, increasing the expected value of a gamble should improve it. So, consider the following two gambles: $G$ = (\$100,0.5; \$0,0.5), a 'fifty-fifty' gamble to win either \$100 or \$0, and $F$ = (\$100,0.2; \$0,0.8), a 20 per cent chance to win \$100 otherwise \$0. The expected values are $EV(G)$ = \$50 and $EV(F)$ = \$20, so people should prefer $G$ over $F$. Indeed, few people would not prefer $G$ to $F$. Similarly, people should prefer $G$ = (\$100,0.5; \$0,0.5)over $F'$ = (\$50,0.5; \$0,0.5) because $EV(G')$ = \$50 and $EV(F')$ = \$25. We could continue doing such 'confirming' experiments for years and continue to find evidence consistent with ('for') EV. However, such research is not very informative – it is like eating bread to prove that bread is made of cyanide. To test a theory, we should think of how it can be refuted rather than on how it might succeed.

According to EV theory, two gambles with the same EV should be equally attractive. Consider $G$ = (\$100,0.5; \$0,0.5) and $F''$ = (\$55,0.5; \$45,0.5) which both have EV = \$50. When these gambles are presented for comparison, most people prefer

$F'' \succ G$, so this result is not consistent with EV. Indeed, most people prefer \$45 for sure to gamble *G*, which has a higher EV of \$50. By looking for exceptions to the theory, we find that EV is systematically violated. Not only do people prefer 'safe' gambles with lower EV over 'risky' ones with higher EV, it has been known for about 300 years that people even prefer a small amount of cash to certain gambles with infinite expected value.

The St Petersburg paradox involves a gamble with infinite expected value. Suppose we toss a coin and if it is heads you win \$2, but if it is tails we toss again, and now if heads occurs you win \$4, but if tails, we toss again. If the coin shows heads on the third toss, the prize is \$8, but if tails, we toss again. The prize for heads doubles each time tails occurs. When heads shows the prize is given and the game ends. This gamble has infinite expected value because

$$\mathrm{EV} = \sum_{\mathrm{i}=1}^{\infty} p_i x_i = \frac{1}{2}\$2 + \frac{1}{4}\$4 + \frac{1}{8}\$8 + \mathrm{K}$$
$$= 1 + 1 + 1 + \mathrm{K} = \infty \qquad (3)$$

However, most people say they would prefer \$10 for sure over a chance to play the game once. This preference was known as a 'paradox' because mathematicians who accepted expected value as the 'fair' price of a gamble also thought it was reasonable to prefer a small amount of cash over this gamble.

Bernoulli (1738) proposed expected utility (EU) as an explanation for the St Petersburg paradox and showed how this theory could explain why people might buy and sell gambles and insurance. Expected utility of gamble *G*= $(x_1, p_1; x_2, p_2; \mathrm{K}; x_n, p_n)$ can be written as follows:

$$EU(G) = \sum_{i=1}^{n} p_i u(x_i) \qquad (4)$$

where u($x$) is the utility of a cash prize of $x$. Whereas $x$ is the objective cash value, Bernoulli assumed that utility of money is not linearly related to money. In particular, Bernoulli suggested that utility is a logarithmic function of wealth. If so, the St Petersburg gamble has finite expected utility (equivalent to the utility of \$4) even though it has infinite EV. Bernoulli showed how EU implies a poor person would not be ill-advised to sell a 50–50 chance to win 20,000 ducats to a rich man for less than its expected value, and how a rich person should be happy to buy it for that price.

Expected utility theory is a theory that is more general than EV in the sense that EV is a special case of EU in which $u(x) = x$. Therefore, evidence consistent with EV is also consistent with EU, but EU can predict phenomena that cannot be explained by EV. This situation is shown in Figure 25.2. In a sense, it seems almost 'unfair' in that there is no observation that can refute EU in favour of EV but there are results that can refute EV in favour of EU.

## Allais' paradoxes refute EU

Allais (1953) proposed two paradoxes that violated EU. These were combinations of choices that cannot be reconciled with either EU or EV. They are known as the 'constant ratio' paradox and the 'constant consequence' paradox. The constant ratio paradox can be illustrated by the following two choices (Birnbaum, 2001):

| | | |
|---|---|---|
| *A*: 0.5 to<br>win \$100<br>0.5 to win \$0 | *B*: sure to<br>win \$50 | 89.5%<br>choose *B* |
| *C*: 0.01 to<br>win \$100<br>0.99 to win \$0 | *D*: 0.02<br>to win \$50<br>0.98 to win \$0 | 34.9%<br>choose *D* |

According to expected utility, $A \prec B \Leftrightarrow C \prec D$ From EU,

$A \prec B \Leftrightarrow EU(A)<EU(B)$. $EU(A) = 0.5u(\$100) + 0.5u(\$0)$; $EU(B) = u(\$50)$. Because most people choose *B* over *A*, we have, $0.5u(\$100)+0.5u(\$0) < u(\$50)$. Multiplying both sides of the inequality by 0.02, we have $0.01u(\$100)<0.02u(\$50)$; subtracting $0.01u(\$0)$ from both sides, we have, $0.01u(\$100)<0.02(\$50)–0.01u(\$0)$; adding $0.99u(\$0)$ to both sides, we have, $0.01u(\$100)+ 0.99u(\$0) < 0.02u(\$50)+0.98u(\$0)$, which holds if and only if $C \prec D$. Consistent with Allais' paradox, which has been replicated many times (Kahneman and Tversky 1979), Birnbaum (2001) found that most people violate EU. Indeed, of the 743 participants

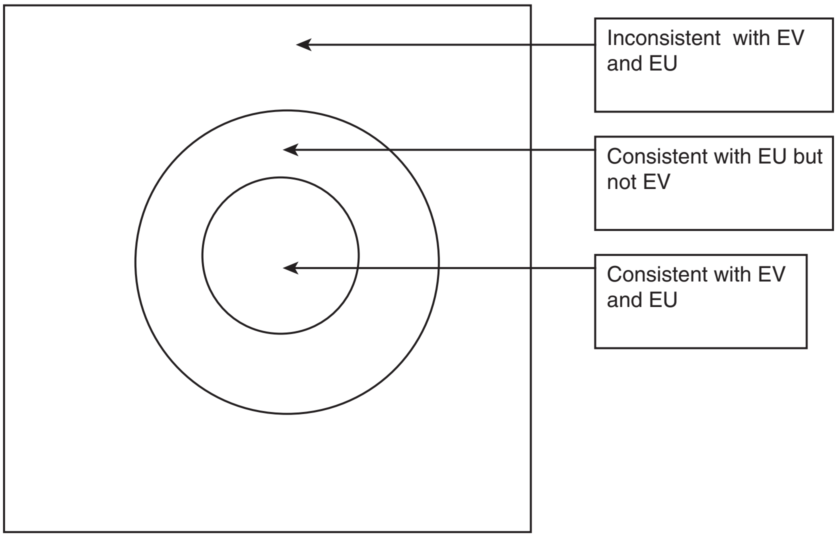
Inconsistent with EV and EU
Consistent with EU but not EV
Consistent with EV and EU

**Figure 25.2** Relationship between EV and EU theories. EV is a special case of EU.

who made both choices, 426 (57 per cent) chose $A \prec B$ and $C \succ D$ and only 20 (2.7 per cent) had the opposite pattern of preferences.

The constant consequence paradox can be illustrated with the following two choices:

| | | |
|---|---|---|
| *E:* 15 to win \$500,000<br>85 to win \$11 | *F*: 10 to win \$1,000,000<br>90 to win \$11 | 70% chose *F* |
| *G*: sure to win \$500,000 | *H*: 10 to win \$1,000,000<br>85 to win \$500,000<br>05 to win \$11 | 29% chose *H* |

Birnbaum (in press-b) tested 200 participants who made each choice twice. According to EU theory, $E \prec F \Leftrightarrow G \prec H$. This implication follows from EU because $E \prec F \Leftrightarrow EU(E) < EU(F)$. $EU(E) = 0.15u(\$0.5M) + 0.85u(\$11) < EU(F) = 0.1u(\$1M) + 0.9u(\$11)$. We can subtract $0.85u(\$11)$ from both sides and add $0.85u(\$0.5M)$, so, $u(\$0.5M) < 0.1u(\$1M) + 0.85u(0.5M) + 0.05u(\$11)$, which holds if and only if $G \prec H$. Therefore, the fact that significantly more than half the participants preferred *F* over *E* and significantly more than half preferred *G* over *H* is inconsistent with EU theory.

In the illustration of Figure 25.2, the Allais paradoxes fall outside the circles representing EU and EV theory. These results refute both EV and EU. A number of theories were proposed to account for the Allais paradoxes and new paradoxes have been devised to test among these new theories. The point of this example is to illustrate how the Allais paradoxes led to refutation of EU theory.

To pick up the modern thread of this story, which includes the development of prospect theories to account for the Allais paradoxes (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), see Birnbaum (2004a, b, 2005a, b; in press a, b). Although the prospect theories can explain the classic Allais paradoxes, Birnbaum showed that a series of 'new' paradoxes refute both versions of prospect theory. The new paradoxes remain consistent with a model by Birnbaum, which awaits the invention of new paradoxes to refute it.

## Between versus within-subjects designs

Suppose we want to manipulate a variable to determine its causal effect. For example, to test EV theory, we might want to compare gambles with equal EV that have different variances.

This could be done by randomly assigning participants to groups that receive different levels of the variable or by manipulating the variable for each person. When we randomly assign people to different groups, it is called a *between-subjects design*, and when the same person receives two or more levels of the independent variable, it is called a *within-subjects design*.

## Confounded contexts

It is important to realize that when the dependent variable of an experiment is a judgement, within and between-subjects designs often yield opposite conclusions. For example, Birnbaum (1999) randomly assigned participants to two groups, each of which was instructed to judge 'how big' a number was. One group judged the size of the number 9, and the other group judged the number 221. It was found that 9 is significantly 'bigger' than 221. Of course, no single participant ever said that 9 is greater than 221, but by the rules of between-subjects designs, an experimenter would conclude that 9 is 'subjectively' bigger than 221. This conclusion should seem silly, but some investigators studied less obvious examples to draw odd inferences.

The problem with between-subjects research is that the context for judgement and the stimulus are completely confounded. Although people have seen numbers both larger and smaller than 9 and 221 before participating, they do not have a context for comparison so they must supply it themselves. Apparently, the number 221 brings to mind a context that includes larger numbers (among which 221 seems 'small') compared to the context evoked by the number 9. This example was devised to create a situation in which few would argue that the conclusion is that 9 'really is' subjectively bigger than 221.

There are many different areas of psychology in which between-subjects designs have been used to reach conclusions that are reversed in within-subjects designs. For example, Jones and Aronson (1973) found that respectable victims are rated more at fault for their own rape than less respectable victims. In particular, women described as a 'virgin' or 'married' were rated more blameworthy than those described as 'divorced'. Jones and Aronson theorized that in order to believe in a 'just world', a respectable victim would not deserve to be raped and therefore she must have done something to deserve it. However, this conclusion is reversed in a within-subjects design. When participants rate both victims or even when they rate both victim and perpetrator, the divorced woman is judged more at fault than the virgin or married woman (Birnbaum 1982).

In the area of human judgement, it has been argued that people 'neglect' base rate, based on the small effects observed when this variable is manipulated between-subjects (Kahneman and Tversky 1979). Similarly, people supposedly fail to distinguish sources of information that differ in validity when making predictions. However, when these variables are manipulated within-subjects, people do attend to base rate and to source credibility (Birnbaum 1976; Birnbaum and Mellers 1983).

Kahneman has argued that the world is 'more like' a between-subjects design than a within-subjects design. So, when results from these designs conflict, he prefers the between-subjects design. If you believe that 9 is subjectively 'bigger' than 221, then you might accept these arguments for between-subjects designs, but if you think otherwise, you should be sceptical of results until they are confirmed or reversed by within-subjects tests.

## Dropouts in between-subjects designs

Many investigators are attracted to web research because of the possibility of testing large numbers of subjects. Between-subjects designs require large numbers of participants, so web-based research might seem a good way to do such research. For example, if a person has a simple 2 × 2 × 2 design with 50 participants in each condition, it requires 400 participants. This requirement might exceed a semester's quota an experimenter might be able to receive from the 'subject pool' of many universities. So, experimenters designing between-subjects studies are attracted to the idea of recruiting large numbers of participants via the Web.

However, web-based experiments have higher dropout rates than lab studies. In lab studies, a participant would have to tell someone they are leaving, so there is some social pressure to stick it out to the end. Via the Web, however, people feel no qualms about just clicking another button to leave a boring task (Birnbaum 2004b, c; Birnbaum and Reips, 2005). For studies of who quits, when and why, see Frick *et al.* (2001). See also Reips (2000, 2002a, b; Reips and Stieger 2004) for suggestions about how to minimize dropout and how to analyse the causes of dropout. When there are dropouts, even when dropout rates are equal in all groups, the data can give a misleading picture of the actual effects of a variable.

For example, Birnbaum and Mellers (1989) showed that a treatment intended to improve test scores (e.g., workshops intended to prepare people to take the Graduate Record Exam [GRE]) could appear to be beneficial by simply including a sample test at the end of the treatment. Suppose that the treatment actually lowers test scores, but people who do poorly on the sample test are less likely to take the GRE at the next test date. Because those who would score low decide not to take the test, a harmful treatment could appear to produce a beneficial gain compared to the control group, when all it did was increase the correlation between preparation and the decision to retake the test. See Birnbaum and Mellers for a detailed numerical example of how this can happen.

Dropouts are still a problem but a less serious problem for within-subjects studies because dropping out is not confounded with the manipulated independent variable – everyone who completes the study provides a separate test of the two conditions. To test for the effects of a GRE workshop, each participant receives both the treatment condition and the control condition, with half the participants receiving the two treatments in each of the possible orders. Participants in both groups will receive the sample test. This mixed design allows both between-subjects comparisons for the effects of treatment orders and within-subject comparisons for the effects of the treatment. The test also provides two dependent variables, the test score after the first treatment and the test score following the second treatment.

## Representative design

Brunswik (1956) argued that between-subjects designs should be avoided because they create situations that are not representative of the environment to which generalization is desired. He also argued against systematic designs (such as one factor designs and factorial designs), in which the independent variables are made to have zero correlations with other variables and with each other. If people use the distribution of the variables including the variance and covariance of the independent variables, then systematic research creates situations in which important variables influencing judgement will have been fixed to unrealistic levels.

Brusnwik (1956) argued that the only basis for generalization from experiments is the theory of statistical sampling. If we wish to know the mean in a population, for example, and if we have random samples, we can use statistical theory of random sampling to estimate and make inferences about that population mean. Similarly, if we want to know the effect of an independent variable, we should sample that variable randomly as well. If the effect of a variable depends on its levels and correlation with other variables in the textured environment, we need to sample randomly from that environment if we hope to generalize our results to that environment.

Brunswik went on to argue that if for practical reasons we cannot sample randomly, the next best approach is to sample representatively.

For example, suppose we wished to predict the outcome of a district election, and we know that republicans and democrats favour different candidates. Suppose 55 per cent of those who vote in this district are democrats. It would certainly be unrepresentative if our sample included 90 per cent republicans. To achieve a more representative sample, we can make sure that the percentage of democrats in our sample matches the percentage of democrats among those we think likely to vote. The same could be done for age, gender, and other variables that we think might affect the outcome. Representativeness is not a very precise concept; indeed, random samples are often not representative. Nevertheless, some scientists are content to treat samples that they believe are representative as if they were random and apply the same statistics.

Brunswik theorized that people are sensitive to the ecological validities of cues in perception. The ecological validity of a cue is the correlation between that cue and the distal state of nature that the perceiver is trying to infer. For example, in order to know how large an object is, one must not only use its proximal size (the size of the retinal image), one needs to know its distance.

But there are many cues to distance; among them are binocular disparity (relative separations in the retinal positions of objects in the two eyes), height in the visual field, geometric perspective and many others. Usually, objects that are higher in the visual field are farther away than objects lower in the field. For example, the horizon is both farther away and higher in the visual field than one's foot on the ground, so this cue has ecological validity in predicting distance.

Suppose an experimenter sets up a study to make height in the visual field independent of all the other variables. If so, that experimenter has made the ecological validity of that cue zero, because in a systematic study of one variable, height in the visual field no longer correlates with actual distance or anything else. If people are sensitive to the ecological validity of a cue, and cue intercorrelations, they should stop using height in the visual field in this experiment, because it no longer has validity as a predictor of distance. In other words, when an experimenter makes this variable independent of all other variables, the experimenter has changed the situation to one from which one cannot generalize. To Brunswik, trying to use this experiment to predict the effect of height in the visual field would be like trying to predict an election with a sample of all republicans.

Figure 25.3 shows a diagram of a factorial design in which each level of variable X is paired with each of the five levels of independent variable Y. This makes X and Y uncorrelated. If this correlation is itself an important determinant of behaviour, this experiment sets this variable to a level that may not be representative of the natural ecology of the person tested.

Brunswik proposed using representative design, in which variables were to be studied in the natural environment. Statistical analyses would then be required to tease apart the effects of confounded variables. Unfortunately, these ideas led to the use of multiple linear regression as both a data analysis device and substantive theory of human judgement. Multiple linear regression is

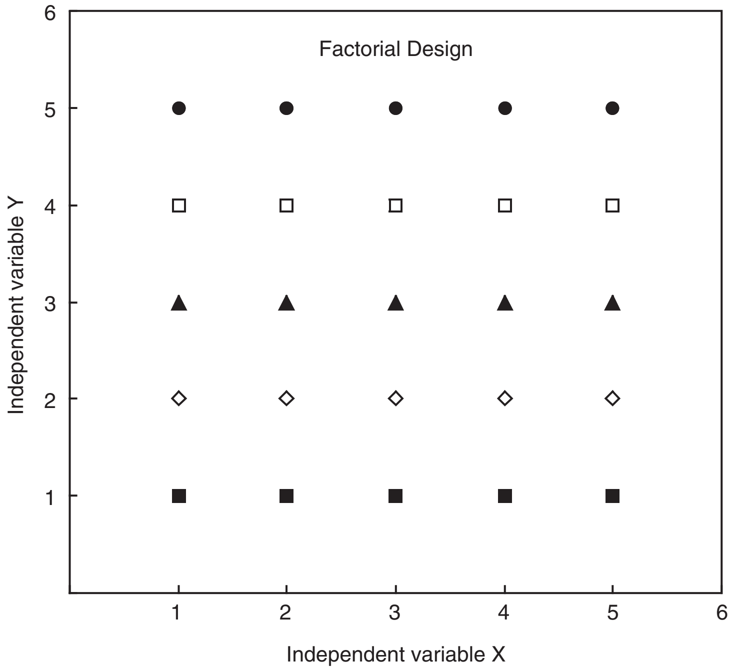

**Figure 25.3** Factorial design. Variables X and Y are independent, producing a zero correlation, but suppose the correlation is not zero in the natural environment and suppose it affects behaviour; if so, then this design sets this factor to an unrealistic value.

known to have many flaws, which are exacerbated when variables are correlated. As a data analysis device, it is not well-understood by those who use it, who often draw inappropriate conclusions from its calculations.

It is unfortunate that some people took Brusnwik's term, ecological validity, and changed its definition as if it referred to a characteristic of an experiment, a theory, a task, or of results. In Brunswik's terminology, ecological validity is an objective concept and it can be calculated. When this term is misued, it usually refers to someone's mushy intuitive judgement of how 'good' or 'bad' someone's study is with respect to how well it psychologically resembles somebody's idea of some 'real world'. When the term 'ecological validity' is used in reference to an experiment, one can simply rewrite the sentence, 'the study has low ecological validity', as follows: 'this study does not appeal to me'. For another view of representative design and the misuse of terms, see Hammond (1998): http://www.albany.edu/cpr/brunswik/notes/essay2.html.

Brunswik also discussed the use of a hybrid design in which a factorial design is modified by leaving out certain combinations that are unusual, creating correlations among variables. But this approach assumes that presentation of these combinations would affect the responses to other combinations and allows no way to test that proposition. Figure 25.4 shows an example of hybrid design. In this case, it was assumed that the environmental correlation between *X* and *Y* is positive, so the experimenter has removed some of the cells of the factorial design that are rare in nature to preserve this correlation in the experiment.

## Systextual design

A criticism of representative design is that it assumes an empirical theory to determine a methodological approach that prevents testing of the empirical theory upon which the method is founded (Birnbaum and Veit 1973, 1974; Birnbaum 1975). An alternative approach is to

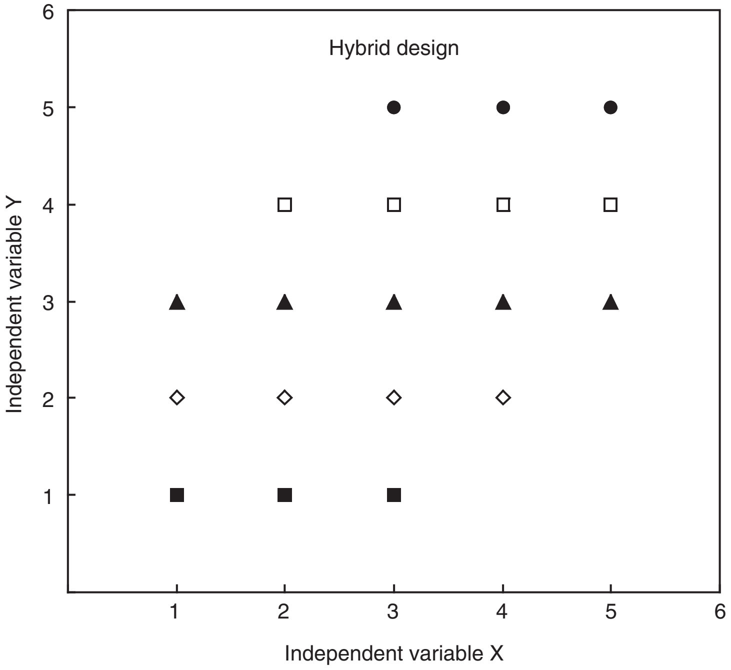

**Figure 25.4** Brunswik deleted some cells in order to remove combinations of variables that were not representative of the natural environment, calling the remaining design a 'hybrid' design. This design still holds the correlation between variables fixed.

use systextual design rather than representative design. Systextual design requires systematic manipulation of the research design itself, including all aspects of the environment that are thought important to behaviour. This method allows us to determine if these variables are important and it allows us to use theory supported by evidence to generalize to new environments, including ones not found in nature.

The theory that people are influenced by cue correlations and intercorrelations is a theory that can be tested by systextual design, which involves systematic manipulation of the context in an experiment. Birnbaum (1975) shows how one can use contextual stimuli to manipulate the correlation between two variables while using a factorial design nested within the overall correlation to analyse the effects of the variables, including the correlation. Several studies reported that the effect of a variable can be altered and even reversed by manipulation of its correlation with a third variable.

Figure 25.5 illustrates an example of systextual design that allows an experimenter to manipulate cue intercorrelation and still use a factorial design to analyze the data. In this case, the factorial combinations form a 5 × 5, *X* by *Y* factorial design. In order to manipulate the correlation between *X* and *Y*, some additional cue combinations are added to create a correlation. To create a positive correlation between *X* and *Y*, the experimenter could present stimuli shown in the figure as '+' symbols. To create a negative correlation, the experimenter can present those combinations indicated by '–', and to create a zero correlation while keeping the range and marginal distributions of *X* and *Y* the same, the experimenter can alternate presentations of the cue-combinations labeled '–' and '+'.

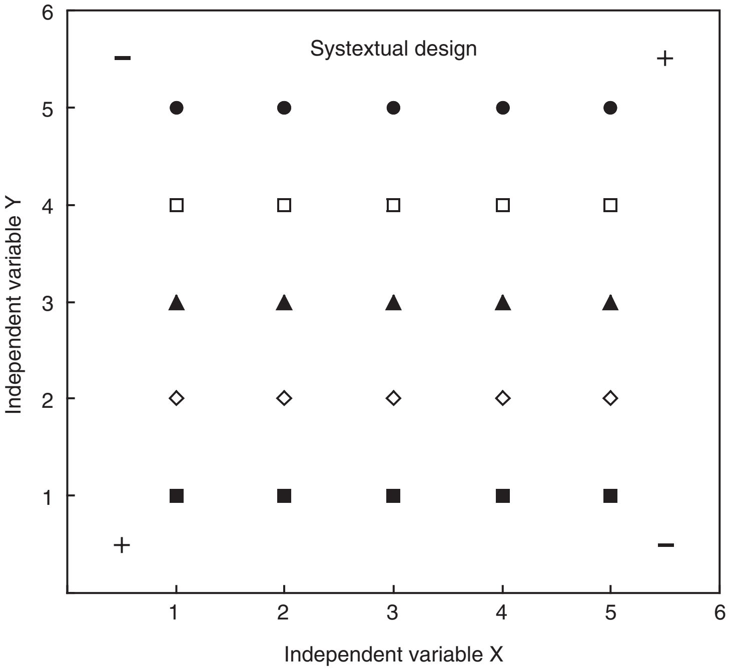

**Figure 25.5** Systextual design. In systextual design, additional cells are added to manipulate the correlation between the independent variables. In this case, adding the combinations denoted plus (+) creates a positive correlation between the variables; adding only the combinations marked with a minus (–) creates a negative correlation, and by presenting both sets of contextual combinations, one maintains a zero correlation. Thus, a factorial design can be nested within an overall correlation.

Using this approach, Birnbaum and Veit (1973) showed that the effect of variable *X* on judgments can be reversed by manipulating its correlation with *Y*. They manipulated the correlation between the number of dots on a card and the size of the card. Participants were instructed to judge the subjective numerosity of the dots on the cards. Participants saw the size of each card before it was turned over to show the number of dots. When there was no correlation between size and actual number, ratings of numerosity were lower on larger backgrounds, as if people expected more dots on larger cards. When the correlation was positive, this effect was magnified. However, when the correlation was negative, judgements were increased in larger areas, showing contrast with a reversed expectancy.

So Brunswik's empirical assumption is not unfounded: people do react to correlations among variables. However, by means of systextual design we can test theories of the effect of these correlations. It is not sampling but theory that is the basis for generalization. By knowing how correlation affects behaviour, we can develop theories that enable us to generalize to any new environment in which the correlations might be different. Brunswik was right to be concerned that cue intercorrelation can influence behaviour, but his advice to use representative design does not allow the scientist to test that proposition or to generalize to a new situation with any correlation, which is possible via theory combined with systextual design.

Parducci (1995) summarizes a research programme investigating the effects of the spacing and probability distribution of independent variables. He has shown that the relationship between physical stimuli and judgments of those stimuli depends on the stimulus distribution, and that this effect is not linear but follows instead a range-frequency compromise. Indeed, this theory was used by Birnbaum (1982) in his analysis of between-subjects studies.

Another literature that involves a systematic manipulation of context is reviewed and summarized by Rieskamp *et al.* (2006). They review studies in which the probability of choosing *A* over *B* depends on the other alternatives in the choice set.

Ordóñez (1998) manipulated the correlation between price and quality by means of a systextual design. It is usually the case that goods of higher quality come at higher prices, but not always, and when the price is fixed, the components of quality are often negatively correlated. For cameras of the same price, one digital camera might have a higher powered optical zoom and another might have a higher resolution. If these correlations were perfect, one would need only to decide how much to pay for an item and one would automatically find the best quality for that price. But these correlations are not perfect, so buyers must compare products.

Fasolo *et al.* (2005) studied how consumers use websites that allow comparisons of goods like digital cameras. They investigated consumer behaviour by means of systextual design and found that people indeed adapt to and use the correlation between cues when they use websites to make decisions about consumer goods. This web-based study examined ways that products can be described on the Web, the effects of decision-aiding tools intended to help people make decisions, and the effect of quality intercorrelations on the way people search for information about the goods. They found that people are sensitive to the correlation structure when they search for information about products. Two decision aids were compared: a compensatory model that aggregated the attribute information and an elimination by aspects model that set thresholds for quality on the attributes. In the negative correlation condition, with the compensatory aid, people clicked on more options. With the elimination by aspects aid, people made more attribute clicks with the negative than positive correlation.

Thus, how people search for information to compare products should not be thought of as a fixed process. It is not the case that people look at attributes for each option or compare the options on a given attribute. Instead, the manner in which people search for information depends on the structure of the environment. In particular, the correlation among the attributes as well as the decision aids available influence how people search for information.

## Conclusions

There is a famous question that is asked at nearly all doctoral oral examinations. This question is some variation of the following: 'I see that your results agree with the theory that you proposed. What would it have been like if your theory were wrong?' As one might expect, the student who passes the final orals should be able to describe what would have convinced them that their theory was wrong and who can show that the study was capable of disproving that theory. Unless one attends to devising a test of at least one theory (and preferably at least two), the effort may be viewed as another case of a person eating bread.

The results of psychological experiments can and do depend on the experimental designs used to establish causal effects. For example, the effect of a variable can be opposite in within- as opposed to between-subjects designs; indeed, in between-subjects, we find that the number 9 is 'bigger' than 221, whereas within-subjects, everyone says 221 is bigger than 9. Between subjects designs are also vulnerable to experimental dropouts, so use of these methods should be avoided if possible in web research, where participants find it easy to drop in and drop out. If a between-subjects design is absolutely required, it is suggested that the experimenter do everything possible to reduce the dropout rate to an absolute minimum.

The effect of a variable can also be reversed when the correlations among independent variables are manipulated via systextual design. These findings mean that the conclusions one draws need to be restricted to the type of experimental design used until one has established the effects of experimental designs themselves. Whereas the representative design uses the theory of sampling to generalize to a particular context, this article advocates the use of psychological theory to generalize to any context.

## Acknowledgements

Support was received from National Science Foundation Grants, SBR-9410572, SES 99–86436, and BCS-0129453.

## References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine. *Econometrica 21*, 503–546.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropoliannae 5*, 175–192. Translated by L. S. (1954). Exposition of a new theory on the measurement of risk. *Econometrica 22*, 23–36.

Birnbaum, M. H. (1975). Expectancy and judgment. In F. Restle, R. Shiffrin, N. J. Castellan, H. Lindman and D. Pisoni (eds), *Cognitive theory* (pp. 107–118). Hillsdale, NJ: Lawrence Erlbaum Associates.

Birnbaum, M. H. (1976). Intuitive numerical prediction. *American Journal of Psychology 89*, 417–429.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (eds), *Social attitudes and psychophysical measurement* (pp. 401–485). Hillsdale, NJ: Erlbaum.

Birnbaum, M. H. (1999). How to show that 9 > 221: collect judgments in a between-subjects design. *Psychological Methods 4*(3), 243–249.

Birnbaum, M. H. (ed.) (2000). *Psychological experiments on the internet*. San Diego, CA: Academic Press.

Birnbaum, M. H. (2001). A Web-based program of research on decision making. In U.-D. Reips and M. Bosnjak (eds), *Dimensions of Internet science* (pp. 23–55). Lengerich, Germany: Pabst Science Publishers.

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology 55*, 803–832.

Birnbaum, M. H. (2004a). Causes of Allais common consequence paradoxes: an experimental dissection. *Journal of Mathematical Psychology 48*(2), 87–106.

Birnbaum, M. H. (2004b). Methodological and ethical issues in conducting social psychology research via the Internet. In C. Sansone, C. C. Morf and A. T. Panter (eds), *Handbook of methods in social psychology* (pp. 359–382). Thousand Oaks, CA: Sage.

Birnbaum, M. H. (2004c). Tests of rank-dependent utility and cumulative prospect theory in gambles represented by natural frequencies: effects of format, event framing and branch splitting. *Organizational Behavior and Human Decision Processes 95*, 40–65.

Birnbaum, M. H. (2005a). A comparison of five models that predict violations of first-order stochastic dominance in risky decision making. *Journal of Risk and Uncertainty 31*, 263–287.

Birnbaum, M. H. (2005b). Three new tests of independence that differentiate models of risky decision making. *Management Science 51*, 1346–1358.

Birnbaum, M. H. (in press-a). Evidence against prospect theories in gambles with positive, negative and mixed consequences. *Journal of Economic Psychology*.

Birnbaum, M. H. (in press-b). Tests of branch splitting and branch-splitting independence in Allais paradoxes with positive and mixed consequences. *Organizational Behavior and Human Decision Processes*.

Birnbaum, M. H. and Bahra, J. (in press). Gain–loss separability and coalescing in risky decision making. *Management Science*.

Birnbaum, M. H., Kobernick, M. and Veit, C. T. (1974). Subjective correlation and the size-numerosity illusion. *Journal of Experimental Psychology 102*, 537–539.

Birnbaum, M. H. and Mellers, B. A. (1983). Bayesian inference: combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology 45*, 792–804.

Birnbaum, M. H. and Mellers, B. A. (1989). Mediated models for the analysis of confounded variables and self-selected samples. *Journal of Educational Statistics 14*, 146–158.

Birnbaum, M. H. and Reips, U.-D. (2005). Behavioral research and data collection via the Internet. In R. W. Proctor and K.-P. L. Vu (eds), *Handbook of human factors in web design* (pp. 471–491). Mahwah, NJ: Lawrence Erlbaum Associates.

Birnbaum, M. H. and Veit, C. T. (1973). Judgmental illusion produced by contrast with expectancy. *Perception and Psychophysics 13*, 149–152.

Birnbaum, M. H. and Veit, C. T. (1974). Scale-free tests of an averaging model for the size-weight illusion. *Perception and Psychophysics 16*, 276–282.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*, 2nd edn. Berkeley, CA: University of California Press.

Fasolo, B., McClelland, G. H. and Lange, K. A. (2005). The effect of site design and interattribute correlations on interactive web-based decisions. In C. P. Haugtvedt, K. Machleit and R. Yalch (eds), *Online consumer psychology: understanding and influencing behavior in the virtual world* (pp. 325–344). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Frick, A., Bächtiger, M. T. and Reips, U.-D. (2001). Financial incentives, personal information and drop-out in online studies. In U.-D. Reips and M. Bosnjak (eds), *Dimensions of internet science* (pp. 209–219). Lengerich: Pabst Science Publishers.

Hammond, K. R. (1998). Ecological validity: Then and now. Available at http://www.albany.edu/cpr/brunswik/notes/essay2.html, retrieved 12 ed June 2006.

Jones, C. and Aronson, E. (1973). Attribution of fault to a rape victim as a function of respectability of the victim. *Journal of Personality and Social Psychology 26*, 415–419.

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

Parducci, A. (1995). *Happiness, pleasure and judgment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Reips, U.-D. (2000). The web experiment method: advantages, disadvantages and solutions. In M. H. Birnbaum (eds), *Psychological experiments on the internet* (pp. 89–117). San Diego, CA: Academic Press.

Reips, U.-D. (2002a). Standards for internet experimenting. *Experimental Psychology 49*(4), 243–256.

Reips, U.-D. (2002b). Theory and techniques of conducting web experiments. In B. Batinic, U.-D. Reips and M. Bosnjak (eds), *Online Social Sciences* (pp. 219–249). Seattle, WA: Hogrefe and Huber.

Reips, U.-D. and Stieger, S. (2004). LogAnalyzer- analyze your logfile. Available at http://genpsylab-logcrunsh.unizh.ch/index.html: Retrieved 1 June 2004.

Rieskamp, J., Busemeyer, J. R. and Mellers, B. A. (2006). Extending the bounds of rationality: evidence and theories of preferential choice. *Working Manuscript*.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.