Chapter 4

Methodological Considerations

In the preceding chapters we have argued that, although interrelated, beliefs, attitudes, and intentions are conceptually distinct concepts which must be independently assessed. In the first part of this chapter we will try to show that available techniques for measuring each of these variables are reliable and have both convergent and discriminant validity. This discussion will be followed by a consideration of more general methodological problems of experimentation in the attitude area.

RELIABILITY AND VALIDITY OF MEASUREMENT

The concepts of reliability and validity concern the degree to which the measuring instrument is free of measurement error. In Chapter 3 we mentioned that an observed score can be viewed as composed of a "true" score and some measurement error. The error component can be further divided into two parts: variable or random error and constant error. This model is presented in Eq. 4.1,

$$x_i = t_i + e_v + e_c, \tag{4.1}$$

where x_i is the observed score; t_i is the "true" score; e_v is the variable error; and e_c is the constant error.

Reliability refers to the degree to which a measure is free of variable error. Thus, if we assume that the "true" score remains constant (e.g., that the person's "true" attitude has not changed), a perfectly reliable instrument will yield the same results on different occasions. Variable factors, such as the person's mood, the temperature or other weather factors, the testing situation, etc., may have different effects on responses on different occasions, thereby reducing the instrument's

reliability. Needless to say, the lower the reliability of some measuring instrument, the less useful it is.

Validity refers to the degree to which an instrument measures the "true" score it was designed to measure—in the present context, the degree to which it measures a given belief, attitude, or intention rather than some other variable. Clearly, the presence of constant error will reduce a measure's validity since the observed score will be consistently contaminated by some irrelevant factor.

One potential source of constant errors is response biases and response sets (Cronbach, 1946, 1950; Guilford, 1954; Scott, 1968). For example, subjects may differ in the extent to which they tend to agree with statements (acquiesce), to give socially desirable responses, and to use extreme or moderate response categories. It should be obvious that when a subject's responses do not reflect his personal beliefs (i.e., the "true" score) but rather reflect his perception of what is socially desirable, the observed score will not be an adequate measure of the true score. Recall that one purpose of most standard scaling methods is to eliminate items to which subjects respond on the basis of variables other than their attitudes. However, these standard techniques may not always succeed in arriving at an instrument that is free of response biases. Guilford (1954), A. L. Edwards (1957), and Scott (1968) have discussed various means by which it is possible to reduce the effects of such biases.

Reliability

There is abundant evidence that standard attitude scales are highly reliable, yielding comparable results when administered on different occasions. Shaw and Wright (1967) and Robinson and Shaver (1969) have reported reliability coefficients for a large variety of Likert, Thurstone, Guttman, and semantic differential scales. Reliabilities are generally very high.

That disguised techniques and physiological measures have tended to be much less reliable (cf. Kidder and Campbell, 1970) may account in part for the fact that investigators have hesitated to employ such techniques in their research.

The reliability of single-response measures varies considerably, depending on the particular judgment required of the subject. Osgood, Suci, and Tannenbaum (1957) have reported relatively high reliabilities for single seven-point bipolar scales used in the semantic differential. Thus responses to probabilistic scales of the semantic-differential type, such as *probable–improbable, likely–unlikely* tend to yield highly reliable measures of the strength of beliefs or intentions. For example, Davidson (1973) reported test-retest reliabilities greater than .95 for the *likely–unlikely* scale.

Clearly, it is possible to locate subjects on evaluative and probabilistic dimensions with a high degree of reliability. The question of reliability, therefore, does not pose a major problem for the measurement of beliefs, attitudes, and intentions when appropriate instruments are employed.

Convergent and Discriminant Validity

Attempts to assess the validity of a measuring instrument can take several forms. If an instrument is a valid measure of attitude toward some object, it should correlate highly with another valid measure of attitude toward the same object; that is, the two measures should exhibit *convergent* validity. The same should be true for two measures of a given belief or two measures of a given intention.

Although many investigators are content with establishing convergent validity, Campbell and Fiske (1959) have argued that an instrument should also have *discriminant* validity. Clearly, two measures of the same concept may be highly correlated (i.e., have convergent validity), but the basis for the correlation may be the same constant error rather than the "true" score. To eliminate this possibility it must be shown that when the same method or instrument (e.g., the semantic differential or the Likert procedure) is used to measure different variables (e.g., attitudes toward different objects), different results are obtained.[1]

Convergent validity. One question to be raised is the degree to which different measures within a given conceptual category can be expected to yield equivalent results. As noted in Chapter 3, there is one and only one attitude toward a given object. In contrast, a multitude of beliefs and intentions can be specified with respect to the same object. Each belief about a given object links the object to a different concept and thus constitutes a different probability dimension. Similarly, each intention with respect to a given object also constitutes a different probability dimension. These probability dimensions link the individual to different actions with respect to the object. In terms of convergent validity, different measures of a given single dimension (evaluative or probabilistic) should yield comparable results. In contrast, convergent validity cannot be expected when two different probability dimensions or two different evaluative dimensions are assessed.

Consistent with these expectations, whenever investigators have obtained more than one measure of attitude toward the same object, the results were almost always identical. Evidence for the convergent validity of standard attitude scales has been available for some time. For example, Edwards and Kenney (1946) constructed Thurstone and Likert scales to measure attitudes toward the church. These different measures of the affective dimension were found to be highly correlated. Similarly, Osgood, Suci, and Tannenbaum (1957) compared their semantic differential measure of attitude toward crop rotation with a Guttman scale designed to measure attitudes toward the same object. Again, these two techniques yielded comparable results. In several studies, Byrne's (1966, 1971) two-item measure of interpersonal attraction was found to be related to standard attitude scales, such as the evaluative dimension of the semantic differential. Moreover, this index was also found to correlate with the seven-point *good–bad* scale. Ostrom

1. Of course, different results can be obtained only to the extent that the different "true" scores are unrelated or independent.

(1969) and Fishbein and Ajzen (1974) found that single self-report scales of attitude toward religiosity (e.g., "My attitude toward being religious is *extremely favorable–extremely unfavorable*") correlated highly with four traditional attitude scales (Thurstone, Guttman, Likert, and semantic differential scales).

There is abundant evidence, then, for the equivalence of different measures of attitude toward the same object. Recall that measures of beliefs and intentions locate an individual on dimensions of probability. The question must be raised whether different measures of a given probability dimension will also produce comparable results. Unfortunately, relatively few studies have obtained more than one measure of a given belief or a given intention. One reason for the relative lack of evidence for convergent validity of different measures of beliefs or intentions is that few attempts have been made to develop standard scaling procedures for these dimensions. Some findings of relevance come from studies by Fishbein and his associates. For example, Fishbein and Raven (1962) used semantic differential scales such as *probable–improbable, true–false, likely–unlikely, agree–disagree,* and *possible–impossible* to measure belief strength. Although most investigations have been concerned with the sum across such scales as a general index of belief, these scales have also proved to be highly intercorrelated. Thus ratings of a statement such as "Orientals are intelligent" on a *true–false* scale yields approximately the same results as ratings of the same statement on a *probable–improbable* scale.

Ajzen (1971a) obtained estimates of belief strength by asking such questions as "The chances are ____ in 100 that Student B is in favor of ending the nuclear arms race." In addition, such statements as "Student B is in favor of ending the nuclear arms race" were rated on four of Fishbein and Raven's (1962) probability scales mentioned above. The sum over the four scales was taken as one measure of belief and the quantitative estimate as a second. The two measures were found to correlate highly and to yield comparable results.

There is thus some evidence that different measures of the same belief are comparable. Since intentions also deal with probability dimensions, it is to be expected that this conclusion also holds for different measures of the same intention. Empirical support for the convergent validity of intentional measures was reported by Davidson (1973), who used *true–false* and *likely–unlikely* scales to assess a variety of family-planning intentions. For example, women's intentions to have two children in their completed families were measured first by one scale and later in the questionnaire by the second scale. The two measures yielded comparable results; convergent validity coefficients were approximately .90.

One may therefore conclude that different measures of a given dimension, whether it is evaluative or probabilistic, will tend to produce comparable results. However, different results may be obtained for measures of different dimensions. This is the problem of discriminant validity.

Discriminant validity. It seems hardly necessary to show that measures of *different* dimensions will tend to yield different results. As noted in Chapter 1, many studies have measured different beliefs, attitudes, and/or intentions and have obtained

markedly different results for each measure. Responding in part to these inconsistent findings, we have suggested the distinctions between beliefs, attitudes, intentions, and behaviors. From our point of view, there is no reason to expect that measures of different dimensions, whether they are beliefs, attitudes, or intentions, will yield comparable results. To be sure, one dimension may be related to some other dimension, and thus similar results may sometimes be obtained. However, the relation between two different dimensions is an empirical question. It follows that one should usually be able to demonstrate a given instrument's discriminant validity by showing that it yields different results when applied to two or more different dimensions. For example, the Likert technique could be used to measure attitudes toward the *church* and toward *supersonic transports*. On the assumption that the "true" attitudes toward these two objects are unrelated, the instrument would have discriminant validity if the observed attitude scores were also found to be unrelated. However, different acquiescence tendencies (i.e., tendencies to agree with any statement) and other factors may result in a spurious correlation between the two attitudes.

The multitrait-multimethod matrix. Since the degree of reliability sets the upper limit for convergent validity, and since apparent discriminant validity may be merely the result of unreliability, Campbell and Fiske (1959) have suggested a procedure that allows simultaneous examination of reliability, convergent validity, and discriminant validity. This procedure is known as the *multitrait-multimethod matrix* analysis since it involves the measurement of at least two traits (or attitudes) by at least two methods. This analysis is based on the intercorrelation matrix of the different traits assessed by the different methods, as well as of the same traits measured by the same methods (i.e., reliabilities). A hypothetical multitrait-multimethod matrix is presented in Table 4.1, where perfect reliabilities and validities are assumed. The solid parallel lines are called the reliability diagonal since they contain correlations between the same measure of the same attitude taken on two different occasions.[2] The broken parallel lines are called the convergent validity diagonal; they contain correlations between two different measures of the same attitude. The solid triangles contain correlations between different attitudes measured by the same method, and the dotted triangles contain correlations between different attitudes measured by different methods.

Lack of perfect reliability would be indicated by coefficients below unity in the reliability diagonal. Similarly, lack of perfect convergent validity is shown by coefficients below unity in the convergent validity diagonal. Lack of perfect discriminant validity is indicated by correlations above 0 in the triangles of Table 4.1.[3]

---

2. Split-half or equivalent-forms reliability coefficients could also be entered in this diagonal.

3. The degree of discriminant validity has to be assessed on the basis of multiple comparisons between coefficients in the triangles and the diagonals. Interested readers are referred to Campbell and Fiske (1959).

Table 4.1 Hypothetical Multitrait-Multimethod Matrix with Perfect Reliabilities and Validities

		Thurstone Scale		Likert Scale	
		Church	SST	Church	SST
Thurstone Scale	Church	1.0			
	SST	0	1.0		
Likert Scale	Church	1.0	0	1.0	
	SST	0	1.0	0	1.0

Note: Cell entries represent correlation coefficients.

Davidson (1973) has reported multitrait-multimethod analyses for measures of attitude and measures of intention. He demonstrated high reliabilities as well as considerable convergent and discriminant validities for his measures. Table 4.2 shows Davidson's multitrait-multimethod matrix for two measures of attitude, each assessing attitudes toward two different objects. The two methods were (1) a sum over three evaluative semantic differential scales and (2) a Guilford self-

Table 4.2 Multitrait-Multimethod Matrix for Attitude Measures (Adapted from Davidson, 1973)

		Semantic differential		Guilford self-rating	
		Next 2 years	Having 2 children	Next 2 years	Having 2 children
Semantic differential	Next 2 years	0.92			
	Having 2 children	−0.13	0.92		
Guilford self-rating	Next 2 years	0.89	0.14	0.86	
	Having 2 children	−0.08	0.91	−0.08	0.84

rating scale. The attitude objects were "having a child in the next two years" and "having two children in my completed family." A comparison of Table 4.2 with Table 4.1 shows the high reliabilities, convergent validities, and discriminant validities of the two attitude measures employed.[4]

Relation between Reliability and Validity

Equation 4.1 above indicates that a measure cannot be valid unless it is reliable; that is, in the presence of variable error, the observed score cannot be equivalent to the "true" score. The relationship between reliability and validity is expressed in Eq. 4.2, *Spearman's correction for attenuation,*

$$r_{xy} = r'_{xy} \sqrt{r_{xx'}} \sqrt{r_{yy'}}, \tag{4.2}$$

where r_{xy} is the observed convergent validity of measures x and y; r'_{xy} is the "true" convergent validity of measures x and y, assuming that these measures are perfectly reliable; $r_{xx'}$ is the observed reliability of measure x; and $r_{yy'}$ is the observed reliability of measure y.

With perfect reliabilities ($r_{xx'} = r_{yy'} = 1$), the observed convergent validity is equal to the "true" convergent validity ($r_{xy} = r'_{xy}$). As the reliability of either measure decreases, so does the observed convergent validity.

One factor that influences the reliability of a given attitude scale is the number of items on the scale. We mentioned in Chapter 3 that when the length of a scale is increased, random or variable errors will tend to cancel out across different items. This relationship between the number of items on a scale and its reliability is described by the *Spearman-Brown prophecy formula,* given in Eq. 4.3,

$$r'_{xx'} = \frac{m r_{xx'}}{1 + (m-1) r_{xx'}}, \tag{4.3}$$

where $r'_{xx'}$ is the estimate of the reliability of a scale m times as long as the original scale, and $r_{xx'}$ is the observed reliability of the original scale.

For example, consider a 20-item Likert scale with an observed reliability of .80. According to Eq. 4.3, if only one of these items was used as the measure of attitude, its reliability could be only .17 and the 20-item scale would still have a

4. Our discussion of convergent and discriminant validity has been concerned only with correlations between two or more measures of a given concept. Investigators have also attempted to validate attitude measures through known-group comparisons or by looking at the degree to which the measure is predictive of overt behavior. The known-group comparison method was described in Chapter 3. The problems involved in using a measure of attitude to predict overt behavior will be discussed in Chapter 8; at this point it is sufficient to note that both convergent and discriminant validities have to be assessed even when overt behavior is used as the criterion against which an attitude measure is to be validated.

reliability of .80.[5] Thus a single-item measure may be highly unreliable although a test made up of many such items may have considerable reliability.

This single-item measure of x could be correlated with some criterion y for purposes of validation. Assuming perfect "true" convergent validity (and perfect reliability of the criterion measure y), Eq. 4.2 indicates that the observed convergent validity would be equal to $\sqrt{r_{xx'}}$, the square root of the observed reliability of x. Since $\sqrt{.17} = .412$, the single-item measure would not be a good predictor of the criterion. In contrast, the 20-item scale would result in an observed convergent validity of $\sqrt{.80} = .894$.

These considerations point to some of the problems inherent in the widespread use of single-response measures in the attitude area, without independent evidence for their reliability. Clearly, many apparently conflicting findings may be due in part to the use of such unreliable and invalid measures. Conflicting findings, however, may be due not only to unreliable and invalid measurement but also to other methodological problems, including low validity of the experimental situation and improper data analyses. Some of these problems will be considered in the remainder of this chapter.

INTERNAL AND EXTERNAL VALIDITY OF RESEARCH DESIGNS

The basic purpose of an empirical investigation is to further our understanding of some phenomenon. Ideally, an investigator starts out with a theoretical network of interrelated constructs from which several hypotheses concerning a given phenomenon can be deduced. He then attempts to test these hypotheses empirically in order to support his theory. When a given hypothesis is repeatedly disconfirmed, he modifies his theory to take account of the data.

Two primary research methods can be distinguished: the *correlational* method and the *experimental* method (Cronbach, 1957). The former approach is largely descriptive; investigators examine the covariation of different variables, compare different groups with respect to one or more variables, look for dimensions underlying sets of responses, etc. For example, the correlational approach could be used to further our understanding of religious behavior. It would be possible to correlate church attendance with various demographic variables, such as age, sex, religious preference, and social status, or with attitudes toward the church, toward God, toward the Bible, etc. Similar information could be obtained by comparing persons who attend church with those who do not on each of these variables.

Alternatively, it would be possible to observe a large number of religious behaviors (e.g., church attendance, donation of money to a religious institution, tak-

5. This can be seen in the following computation:

$$r'_{xx'} = \frac{1/20(.80)}{1 + (1/20 - 1).80} = \frac{.04}{1 - .76} = .17.$$

ing Bible classes, etc.) and to perform a factor analysis in an attempt to determine the dimensions that underlie these behaviors.

The most characteristic feature of the correlational approach, therefore, is its reliance on existing variation. In contrast, the experimental approach creates variation by manipulating one or more variables and examines the effect of the manipulation (the independent variable) on some response measure (the dependent variable). In the correlational approach one merely obtains an index of the relationship between attitude toward the church and church attendance; the correlation between these variables is based on preexisting individual differences in attitudes and in church attendance. Clearly, a causal effect of attitude on church attendance could not be inferred since an observed relationship between these variables could also be the result of church attendance causing the attitude. Moreover, the relationship may also be a function of a third variable (e.g., belief in God) to which both attitude and church attendance are related.[6]

Internal Validity

Causal inferences, however, are the main objective in the experimental method. For example, one could manipulate attitude toward the church and observe the effects of this manipulation on church attendance. The most characteristic feature of the experimental method, therefore, is the experimenter's control over the independent variable or variables. However, the demonstration that a manipulation of attitude is accompanied by a change in church attendance is not sufficient evidence for inferring a causal relation. Care must be taken to ensure that the observed effects on church attendance are indeed attributable to the manipulation of attitude and not to some other uncontrolled variable. The term *internal validity* has been used for the degree to which an experimental design is free from such uncontrolled factors—that is, the degree to which changes in the dependent variable can be confidently attributed to the experimental manipulations and only to the experimental manipulations. High internal validity, therefore, eliminates possible alternative explanations of the observed effects (D. T. Campbell, 1957).

Indeed, in the design of an experiment, emphasis has usually been placed on ensuring a high degree of internal validity in order to eliminate alternative explanations and to allow causal inferences to be made. Campbell (1957) and Campbell and Stanley (1963) have discussed a number of factors that may impair the internal validity of an experimental design and have suggested ways to avoid these difficulties.

Consider, for example, an experiment in which attitude toward some product x is measured five weeks before and immediately after exposure to a commercial advertising the product. Imagine that the mean attitude of subjects after exposure is significantly more positive than before exposure. Clearly, this effect may be

6. It is a well-known fact that the amount of damage observed at a fire increases with the number of firemen present. Clearly, this relationship is not a causal one; rather, it is the result of a third variable, the size of the fire.

attributable to factors other than the experimental manipulation (i.e., exposure to the commercial). One uncontrolled factor is the possible effect that taking the pretest may have on posttest responses; the pretest itself may have provided new information about product *x*, it may have made the subjects familiar with the brand name, etc. Another possibility is the occurrence of uncontrolled events in the interval between pretest and posttest. Some subjects may have used the product in the meantime (perhaps because the pretest made them aware of it) and they may have liked it; they may have been exposed to other ads in the interval or obtained additional information about the product, etc. Obviously, each of these factors constitutes a potential alternative to the explanation that the change in attitude was produced by the commercial.

The simplest way to deal with these uncontrolled factors is to add a group of subjects who are not exposed to the commercial but whose attitudes are measured at the same points in time. Subjects should be assigned at random into experimental and control groups. This procedure reduces the likelihood of systematic differences between the groups. The uncontrolled factors discussed above should therefore be as likely to occur in the control group as in the experimental group. One can conclude that the commercial had an effect on attitudes toward product *x* only if the change (from pretest to posttest) in the experimental group is significantly greater than the change in the control group.

Internal validity requires that an observed effect be attributed *solely* to a given manipulation. That is, it must also be shown that the effect cannot be attributed to any factor other than the manipulation. In the example above, use of a control group does not eliminate the possibility that the commercial had an effect only because subjects were pretested. Even though the experimental group may have changed significantly more than the control group, the pretest may have interacted with the commercial to produce the change in attitude, and if a pretest had not been given, the commercial might not have led to a more favorable attitude in the experimental group than in the control group. For example, the pretest may have sensitized subjects to the subsequent commercial, increased their attention, and thereby enabled the commercial to produce a significant change in attitudes.[7]

This pretest-manipulation interaction can be avoided by eliminating the pretest in the experimental and control groups; effects of the manipulation are then assessed by comparing the posttest scores of the two groups. Assuming initial equivalence in attitudes produced by random assignment, a significant difference between experimental and control groups is attributable to the manipulation.

If it is desirable to assess the degree to which a pretest-manipulation interaction is operative in a given experiment, the Solomon Four-Group Design can be used. In this design, all four groups mentioned above are employed, i.e., two experimental groups (one with a pretest and one without it) and two control

7. Campbell and Stanley (1963) discussed this effect in the context of external validity, which will be considered below.

groups (with and without pretest). The interaction effect is tested by considering the posttest scores of the four groups. For a more complete discussion of this and other designs, see Campbell and Stanley (1963).

The notion of pretest-manipulation interactions suggests that the experimental manipulation might not have produced the observed effect in the absence of a pretest. Similarly, it has been argued that the observed effects of manipulations in any experimental setting are partly a function of the experimental setting itself. That is, the manipulation may interact with the experimental setting in producing the observed effect. For example, knowledge of taking part in an experiment may increase attention to the commercial or any other manipulation. Furthermore, the subjects may form beliefs about the purpose of the experiment or about the experimenter's expectations; their responses may then be in part a function of these beliefs rather than of the manipulation itself.

Phenomena of this kind, called *reactive effects,* have attracted increasing attention in recent years.[8] Whereas the pretest-manipulation interaction can be controlled for by an appropriate experimental design, little can be done to reduce reactive effects in most experimental settings. Indeed, most research on reactivity has attempted to *demonstrate* the operation of these effects in various experimental settings (e.g., Orne, 1962; Rosenthal, 1966; Rosenthal and Rosnow, 1969; Page, 1969, 1970a) rather than to eliminate them.

We shall consider these reactive effects in greater detail below. First, however, the reader should realize that the hypothetical experiment discussed earlier (even if internally valid) would permit only the conclusion that the commercial had an effect on attitudes. Little could be said as to which aspects of the commercial produced the change. Thus, the investigator might assume that the *information* about product x provided in the commercial was responsible for the effect, but in reality the change might have been brought about by *any* commercial irrespective of the kind or amount of information it contained. To test the hypothesis concerning effects of information about product x would require a different experiment involving additional manipulations. Specifically, the kind and amount of information about product x contained in the commercial would have to be systematically manipulated.

Laboratory versus field studies. It has often been argued that field studies provide mainly descriptive data and laboratory studies are more analytical and yield data about causal relationships. Generally speaking, this has indeed been true since most field studies have used the correlational method whereas the majority of laboratory studies have used the experimental method. However, there is no necessary relationship between a research setting and a research method. Thus the correlational method has frequently been employed in laboratory studies, and there is a growing interest in using the experimental method in field research.

8. Reactive effects were also treated in the context of external validity by Campbell and Stanley (1963).

Campbell and Stanley (1963) have discussed a number of "quasi-experimental" designs that ensure maximal internal validity in a variety of field settings.

It has also been argued that although field studies have lower internal validity than laboratory studies, their findings are more generalizable than findings obtained in the laboratory. Again, however, there is no necessary relation between generalizability and research setting. Campbell and Stanley have discussed the problem of generalizing experimental (as opposed to correlational) findings under the heading of external validity of experimental designs.

External Validity

Campbell (1957) has argued that an experiment must have not only internal validity but also generalizability or external validity. That is, "to what populations, settings, treatment variables, and measurement variables can this effect be generalized?" (Campbell and Stanley, 1963, p. 5). Clearly, these questions can be asked of any research, whether correlational or experimental, conducted in the field or in the laboratory.

Generalizing across populations. The prevalent use of college students in social-psychological research has always been a major focus of criticism on the grounds that college students are not representative of the general population. Similarly, it may be argued that research findings cannot be generalized from one culture or subculture to another.

Lack of generalizability constitutes a serious problem when the purpose of the research is primarily descriptive. Thus permissive attitudes toward use of marijuana found among college students may not be generalizable to other populations. However, when psychological processes, theories, or lawful relationships are under investigation, generalization across populations may be less problematic (cf. Kruglanski, 1973; in press). For example, although college students may have more favorable attitudes toward using marijuana than, say, law enforcement officers or union members, one may argue that the formation and change of these attitudes obey the same laws for all subject populations. Specifically, attitude toward marijuana use may always be found to be a function of beliefs about the consequences of using marijuana and the evaluation of those consequences.

Since the major purpose of most experimental investigations that use college students is to discover lawful processes (rather than to provide descriptive accounts of the research population), the concerns with problems of generalization across populations may have been exaggerated. This is not to say that the problem can be ignored in experimental investigations, since different psychological processes may be operating in different populations or cultures. Empirical evidence to date, however, indicates that most psychological processes are generalizable across different subject populations (Byrne, 1971; Triandis, Malpass, and Davidson, 1972).

Problems of generalizing to different populations also arise when manipulations interact with other aspects of the experimental situation. One possibility is

the pretest-manipulation interaction previously discussed under the heading of internal validity. Campbell (1957) has treated this interaction effect in the context of external validity since he interpreted it as preventing generalization of a finding from a pretested population to a nonpretested population.

Another possibility, also mentioned above, is reactive effects which may be interpreted as reducing the generalizability of findings from experimental subjects (aware that their responses are being observed) to nonexperimental populations. Reactive effects, however, are often viewed as impairing generalizability from an experimental *setting* to a nonexperimental setting (Campbell, 1957).

Generalizing across settings: Reactive effects. Specific research findings may not be generalizable from one setting to another. For example, Minard (1952) examined discriminatory behavior of white coal miners in Pocahontas County, Pennsylvania. He found that although 80 percent of the white miners were friendly toward blacks in the mine, only 20 percent were friendly in town. The conclusion that whites discriminate against blacks, therefore, cannot be generalized from the town to the mine.[9] Again, however, the processes underlying discriminatory or nondiscriminatory behavior may be the same in both settings, and hence the problem of generalizing across settings may be less severe in studies investigating psychological processes.

A more important problem arises when the research setting interacts with the experimental manipulation to produce the observed effects, i.e., when reactive effects are present.[10] Reactive effects have been discussed under the labels of *demand characteristics* (Orne, 1962, 1969), *experimenter bias* (Rosenthal, 1966), and *evaluation apprehension* (M. J. Rosenberg, 1965b, 1969). Orne (1962, p. 778) described reactive effects as follows: "At some level [the subject] sees it as his task to ascertain the true purpose of the experiment and respond in a manner which will support the hypotheses being tested. Viewed in this light, the totality of cues which convey an experimental hypothesis to the subject becomes significant determinants of subjects' behavior." Orne called the sum total of these cues the "demand characteristics of the experimental situation."

For example, subjects may come to believe that one purpose of the experiment is to evaluate their emotional stability, intelligence, or mental health. Rosenberg (1965b) has argued that subjects who form such beliefs will try to behave in the experiment in a manner they think will win the experimenter's approval. He called this phenomenon "evaluation apprehension."

Cues that may allow the subject to form beliefs about the purpose of the ex-

9. The same problem exists when attempts are made to generalize specific findings over time; attitudes or behavioral patterns will tend to change over time.

10. Reactive effects are usually viewed as responses to the total experimental situation, including the experimental manipulation. Alternative explanations attributing the observed effects to aspects other than the experimental manipulation, however, can be ruled out by using an appropriate control group. The crucial reactive effects, therefore, are those produced by the manipulation-setting interaction.

periment, as well as beliefs about the experimenter's expectations, may be provided by the experimenter's behavior; Rosenthal (1966) has called this phenomenon "experimenter bias" effects. Other cues are provided by the situation itself. "For example, if a test is given twice with some intervening treatment, even the dullest college student is aware that some change is expected, particularly if the test is in some obvious way related to the treatment" (Orne, 1962, p. 779). The situational cues may be more subtle; Glinski, Glinski, and Slatin (1970), for example, found that a subject was more likely to believe that the experiment was concerned with persuasion when he was faced with unanimous opposition from a group than when at least one member of the group agreed with his position.

The social psychology of the psychological experiment. The various reactive effects are related to three broad processes: (a) The subject forms hypotheses to explain the sequence of experimental events that he observes. The hypotheses may be related to the purpose of the experiment or to contingencies between his own behavior and other events. (b) He forms beliefs that the experimenter expects or would like him to behave in certain ways. (c) He is or is not willing to meet the perceived expectations of the experimenter.

There seems to be little doubt that experimental manipulations may affect hypotheses that are being formed about the purpose of the experiment (Page, 1969, 1970; Silverman and Regula, 1968). For example, as mentioned above, variations in the unanimity of disagreeing majorities in a conformity situation were found to influence subjects' beliefs that the experiment was concerned with persuasion (Glinski, Glinski, and Slatin, 1970).

Evidence from verbal conditioning experiments indicates that subjects also form hypotheses about the contingencies between their own behavior and reinforcing events (Dulany, 1968; Page, 1969). Thus subjects may form the hypothesis that whenever they start a sentence with "I" or "We," a certain event occurs (e.g., a light appears or the experimenter says "good").

There is also abundant evidence for experimenter bias effects (e.g., Barber and Silver, 1968a, 1968b; Rosenthal, 1968; Rosenthal and Rosnow, 1969). In a series of experiments by Rosenthal and his associates, subjects were shown ten photographs of faces and were asked to rate the degree of success or failure shown in the face of each person. The task was administered by student experimenters, who were told either that subjects tend to rate the faces as expressing success or that subjects tend to rate them as expressing failure. The students were further told that the experiment's purpose was to "see how well they could duplicate experimental results which were already well established" (Rosenthal, 1969, pp. 223–225). Rosenthal found that experimenters tend to get results consistent with their expectations.

It can thus be argued that the experimenter's bias provides the subject with cues as to what is expected of him. A controversy has developed around the ease with which these biases can be communicated to the subject. Available evidence

indicates that they are not likely to occur unless there is an intentional influence attempt by the experimenter (Barber and Silver, 1968a, 1968b; Gallo and Dale, 1968). Such intentional influence attempts are likely to occur in Rosenthal's experimental paradigm described above.

Finally, evidence suggests that the experimental setting itself can influence a subject's willingness to perform what he perceives is expected of him. For example, when evaluation apprehension is produced by the experiment, subjects will tend to respond in a direction of favorable self-presentation, whether this is consistent or inconsistent with their perceptions of the experimenter's expectations (e.g., Silverman and Shulman, 1970).

A number of factors have been found to influence the effects above. Volunteering may increase the likelihood of belief formation and of motivation to meet the experimenter's perceived expectations (Horowitz, 1969; Rosnow and Suls, 1970). Differences in overall suspicion may affect reactions to the experiment (Stricker, Messick, and Jackson, 1969), as may use of a pretest (Rosnow and Suls, 1970). Similarly, variables that may be related to the subjects' sophistication (e.g., previous experience in experiments, knowledge of psychology, time of semester at which the experiment is conducted) have influenced obtained results (Holmes and Appelbaum, 1970). None of the factors discussed, however, has consistently produced significant results; contradictory findings are to be expected since different experimental settings may vary in the extent to which they allow beliefs and hypotheses to be formed. (Extensive reviews of this literature can be found in Rosenthal and Rosnow, 1969; Weber and Cook, 1972; A. G. Miller, 1972; and Kruglanski, in press.)

Demand characteristics, evaluation apprehension, experimenter bias, etc., have typically been viewed as variables that affect behavior only in research settings. It has therefore been argued that research findings cannot be generalized to other settings. However, we mentioned earlier that reactive effects involve the formation of beliefs about the situation, about the consequences of behavior, and about the expectations of other people, and further that they may be related to the subject's willingness or motivation to meet those expectations. Since the same processes are likely to operate in any situation, reactive effects in an experiment can be viewed as specific instances of the kinds of variables that influence behavior in general. In contrast to the notion that experimental settings are atypical, therefore, one could argue that they are actually quite representative of most real-life situations.

These considerations suggest that subjects, like other human beings, are conscious, active, intelligent organisms who react to the total situation in which they find themselves. It is therefore virtually impossible to study the effects of some manipulation in isolation from the beliefs and hypotheses that subjects form about the situation. To return to the question of internal validity, perfect internal validity appears to be an unobtainable goal in research on human subjects, since it is impossible to ascertain the pure or "true" effect of the experimental manipula-

tion on some dependent variable. The subject's perception and interpretation of the total situation may always interact with the manipulation and thus "contaminate" the effect of the latter.

Acceptance of the conclusion that "reactive effects" are an inevitable part of experimental settings and, for that matter, of any other setting suggests that an alternative research strategy is called for. Research to date has attempted either to demonstrate the operation of reactive effects (e.g., Orne, 1962; Page, 1969, 1970; Page and Scheidt, 1971; Rosenthal, 1966) or to minimize their "detrimental" impacts (Rosenberg, 1965b). It appears that little can be gained from this line of research on reactive effects. In order to increase our understanding of the ways in which an experimental manipulation affects some dependent variable, it seems necessary to consider the manipulation within a broader frame of reference. Specifically, research should be directed at investigating the formation of beliefs about behavioral contingencies, about the purpose of the experiment, and about expectations of other people, and at the determinants of the person's motivation to meet those expectations, etc. In the context of an experimental setting, these variables may be influenced by many factors, including the experimental manipulation; differences between experimental and control groups are expected only when the manipulation influences one or more of the variables.

Consistent with this reasoning, Page's (1969, 1970) attempts to demonstrate the operation of demand characteristics have shown that in many studies, experimental manipulations affect the dependent variables by influencing one or more of the following (intervening) variables: *contingency awareness, demand awareness,* and *motivation to comply.* The first variable is the subject's awareness of the relationships between his own behavior in the experiment and the occurrence of certain (reinforcing) events. This kind of awareness has been investigated primarily in studies of verbal conditioning. Demand awareness is the subject's awareness of the experimenter's hypothesis and expectations. The third variable is the subject's motivation to behave in accordance with the experimenter's hypothesis.

Although Page has not employed his analysis beyond attempts to demonstrate the operation of demand characteristics, Dulany (1961, 1968) had previously employed similar constructs to deal with questions of awareness in studies of verbal conditioning. More important, these variables were developed as part of a more general theory that can be used to predict behavior in experimental as well as nonexperimental settings. This theory and some of its implications for the attitude area will be discussed in Chapter 7.

In conclusion, although there is little doubt that various reactive effects are operating in experimental settings, it appears that these effects need not seriously impair generalizability of experimental findings to nonexperimental settings, since similar "reactive effects" influence behavior in the latter settings. So long as the investigation is concerned with psychological processes (rather than with descriptive accounts of phenomena), it appears possible to generalize from one setting to another.

Generalizing across measurement variables. There is little disagreement that the effect of a manipulation on a dependent measure x may not be generalizable to a totally different dependent measure y. For example, manipulation of physical effort expended during a 15-minute period may influence consumption of milk, but the effect would hardly be generalizable to consumption of a bitter quinine solution.

In Chapter 1 we showed that one serious problem in attitude research is the practice of unwarranted generalization across different measurement variables. For example, the generalization that "communicator credibility affects attitude change" is comparable to the generalization that "physical effort influences consumption of liquids." Just as a distinction has to be made between milk, quinine solutions, and other liquids, a distinction has to be made between different attitudes, such as attitude toward the church, toward marijuana, toward a political candidate, etc. Further, just as the effects of physical effort on consumption of milk cannot be generalized to consumption of horse meat, the effects of communicator credibility on attitude toward marijuana cannot be generalized to the belief that marijuana is produced in South America or to the intention to sell marijuana.

In contrast, it should be possible to generalize across different measures of the *same* belief, the *same* attitude, or the *same* intention. One major problem in attitude research, then, is not so much the question of generalizing across different measurement variables as the failure to distinguish between different variables and the concomitant view that experimental effects should hold for any measure that is labeled "attitude."

Generalizing across treatment variables. Up to this point we have focused attention on the problems of measuring and distinguishing between different dependent variables relevant to attitude research. Similar problems can be identified with respect to experimental manipulations or treatments, i.e., the independent variables. Different noncorrelated operationalizations can be found for the same concept, and the identical operation is often given different conceptual labels. For example, "distraction" has often been manipulated by delayed voice feedback, performance of an irrelevant task, noise, visual displays, anticipation of a noxious experience, etc. At other times, the same manipulations have been labeled effort, fear, stress, and arousal. Such practices are likely to result in what appear to be inconsistent findings. For example, two studies (Friedman, Buck, and Allen, 1970; Hendrick and Shaffer, 1970) investigated the effects of "arousal" in a communication and persuasion paradigm and reported contradictory results. Closer inspection reveals, however, that one study was concerned with the effects of a drug (epinephrin), and the other considered the volume of a recorded persuasive message. Although it is true that multiple operations of a given concept may increase the generality of a conclusion beyond the specific details of any one experiment, frequent reports of inconsistencies point to the possibility that the different manipulations (even when given the same label) are far from equivalent. It follows that different manipulations may constitute operationalizations of dif-

ferent independent variables. Thus, as with dependent measures, the problem is not so much a question of generalizing across different manipulations of the same variable as the failure to distinguish between different kinds of independent variables.

Recapitulation

The experimental method has typically been viewed as characteristic of laboratory research whereas field research has generally been considered correlational in nature. We have pointed out, however, that there is no necessary relation between research method and research setting. The question of internal validity is relevant only in the context of the experimental method, since it is here that causal inferences are being made; the question of external validity, however, applies equally to experimental and correlational methods. It has frequently been argued that laboratory research has higher internal but lower external validity than field research. We have noted, however, that the experimental method can be applied in field research, thus increasing internal validity.

With respect to external validity, a distinction has to be made between generalization of specific descriptive research findings and generalization of psychological laws or processes. There is little doubt that descriptive research findings may not be generalizable across different populations or across different settings. Since field studies tend to employ more representative populations and settings, descriptive results obtained in such studies will have greater external validity than will descriptive laboratory findings. However, when the investigation is concerned with psychological processes rather than description, generalizations across populations or settings are possible to the same extent in laboratory and field research. We have argued that such generalizations will frequently be justifiable.

Finally, we have tried to show that attitude research is characterized by overgeneralization across and failure to distinguish between different treatment variables and different measurement variables. Thus widely different operations are used to manipulate a given conceptual independent variable. Clearly, this practice is desirable only if different operations produce the same results; unfortunately, investigators have continued to use the same conceptual labels for different operations, even when these operations have led to contradictory results. As we have shown in previous chapters, the same conclusion can be reached with respect to conceptual dependent variables, since a wide variety of different measures have all been given the same label, "attitude." Clearly, then, in order to resolve apparent inconsistencies in the attitude area, greater attention must be paid to the labels that are associated with different manipulations and with different measures.

Data Analysis and Interpretation

Another problem that may be responsible for some of the conflicting findings is related to the analysis and interpretation of data. One cannot fail to be impressed by the widespread mistreatment of data, abuse of statistical procedures, and the

frequency with which invalid conclusions are drawn in attitude research. A complete discussion of these problems is beyond the scope of this book, but some general comments are in order. For an excellent discussion and some examples of problems of data analysis and interpretation in one area of investigation, the reader is directed to the exchange between Barber and Silver (1968a, 1968b) and Rosenthal (1968).

Let us return to the hypothetical attitude-change experiment in which an experimenter pretests two groups of subjects, exposes one to a commercial, and posttests both groups. Assume that the attitude measure is always a single seven-point evaluative *good–bad* scale. First, note that the obtained data can be scored in a variety of ways. It is possible to compute pretest means and posttest means for each group; to compute pre- to posttest change scores for each subject and obtain the average change score for each group; to simply count the number of subjects changing in a positive or negative direction in each group; to compute the proportion of subjects changing in a positive direction in each group; to rank-order the subjects in each group in terms of their pretest, posttest or change scores; etc. Clearly, statistical analyses based on these different scores may yield different results. For example, a significant difference may be found between experimental and control groups in terms of proportion of subjects changing their attitudes in a positive direction, but the average amount of change may not differ significantly. Thus, just as it is important to pay attention to differences in manipulations and measuring instruments, so it is important to realize that different indices (e.g., of attitude change) based on the same data may not be comparable.

Different tests of significance applied to a given set of data may also yield different results. In part, this circumstance may be attributable to the fact that different significance tests often require different indices. (For example, some tests are based on means, others on ranks or proportions.)

Another problem is that many studies published in the attitude literature are presented as supporting the researcher's hypotheses when either the results obtained fail to reach acceptable levels of statistical significance by some small margin or they are not significant in one statistical analysis but are shown to be significant in another. When the findings of a study are close to being significant, the investigator may simply discuss the effects as if they were significant, pointing to differences between groups that are in the expected directions. The American Psychological Association *Publication Manual* specifically cautions against "inferring trends from data which fail by a small margin to meet the level of significance adopted. Such results are most economically interpreted as a function of chance and should be reported as non-significant." Although we are aware that the five percent significance level is an arbitrary convention, its utility is evidenced by the literature. Much contradiction and controversy might have been avoided if findings not reaching this criterion had been rejected.[11]

---

11. In addition to observing the requirement of a minimum level of significance, research reports would be more useful if they also included estimates of the percent of variance accounted for by the experimental effects.

A similar suggestion is that significant results obtained on the same set of data by one statistical analysis but not another should also be interpreted as a function of chance and should lead to the rejection of the hypothesis unless the different analyses were initially predicted to yield different results.[12]

CONCLUSION

This chapter has dealt with a number of methodological issues that are of importance in attitude research. We have discussed reliability and validity of measurement techniques as well as problems of internal and external validity of research designs; we have pointed out that attention must be paid to the exact nature of independent and dependent variables; and we have considered some of the problems in data analysis and their interpretation. We have seen that beliefs, attitudes, and intentions can be empirically distinguished and that reliable and valid techniques for measuring these concepts are available. Results obtained with respect to a given dependent variable may not generalize to some other variable, and different manipulations of a given independent variable may also lead to apparently conflicting findings, since they may actually be operationalizations of different variables. The question of generalization across populations or settings was seen to be of greater relevance to descriptive research than to theory- or process-oriented investigations.

To justify the proposed distinction between beliefs, attitudes, intentions, and behaviors, it is necessary to demonstrate that different laws apply to these concepts. That is, it must be shown that different factors influence these variables or that they are differentially related to other variables. In Chapter 1 we outlined a conceptual framework of the relationships between beliefs, attitudes, intentions, and behaviors. The remainder of this book examines empirical research in the attitude area. We shall see that much of the research deals with the relations between beliefs, attitudes, intentions, and behaviors and with the effects of various factors on these variables.

Earlier in this chapter we suggested that the effects of a given event or manipulation on a given dependent measure can be understood only in terms of a larger frame of reference involving the person's interpretation of the event. In the

12. Unfortunately, for the purpose of salvaging a nonsignificant experiment, data are often manipulated until a significant result is obtained. In addition to shifting to another (and usually weaker) statistical test or transforming the dependent variables, the investigator can often accomplish this goal post hoc by combining conditions, subdividing the sample and performing internal analyses, shifting from a two-tailed to a one-tailed test of significance, etc. Although some of these practices may serve an exploratory purpose and provide ideas for future research, they do not provide conclusive evidence for a hypothesis, and the obtained results are best interpreted as a function of chance. For an excellent discussion of the use of many of these techniques, the reader is directed to the exchange between Barber and Silver (1968a, 1968b) and Rosenthal (1968) on experimenter bias effects.

framework of our conceptual system, this implies that a person exposed to some stimulus situation will form beliefs about it. The beliefs themselves can serve as a focus of research. For example, an investigator may vary the frequency with which he rewards different subjects for producing response x. He can then obtain a measure of the subjective probability (i.e., belief) that response x leads to obtaining the reward. Thus the subject forms beliefs that are descriptive of the stimulus situation to which he is exposed.

Usually, however, an investigator is less interested in these descriptive beliefs than in the effects of the stimulus situation and particularly his manipulation on some other belief, attitude, intention, or behavior. For example, he may be interested in the effects of the reinforcement schedule on the subject's belief that the experimenter expects him to perform response x; he may be interested in the subject's attitude toward the person who administered the rewards; in the subject's intention to help the experimenter in a future study; or in the actual frequency with which behavior x is performed by the subject.

Our conceptual framework suggests that all these effects are mediated by the different descriptive beliefs formed in the situation. Specifically, given the belief that behavior x leads to reward, the subject may infer that the experimenter expects him to perform the behavior. He may also infer that the experimenter is honest; in conjunction with other descriptive and inferential beliefs about the experimenter, the belief concerning honesty may influence the subject's attitude toward the experimenter. In a similar fashion, the different descriptive beliefs can mediate intentions and behaviors.

Thus, according to our conceptual framework, the effects of any given stimulus variable on a response are mediated by processes within the organism. In studying the effects of a commercial about product x, for example, the investigator may be interested in changing beliefs about product x, changing attitudes toward the product, changing intentions to buy the product, or influencing actual purchasing behavior. Our conceptual system deals with the different kinds of intervening processes that must be studied in order to understand the effects of a stimulus (the commercial), on the different dependent measures. Conflicting findings are to be expected when these intervening processes are not taken into account.