

Chapter 2.

NATURALISTIC RESEARCH METHODS IN PSYCHOLOGY

Many questions in science are answered through systematic observation of the world around us. Almost all of astronomy, for instance, is observational science. Systematic observation may involve special facilities and sophisticated equipment, as is true of modern astronomy, but observational research does not have to be done in a laboratory nor does it necessarily require specialized apparatus. In many cases, one only needs pencil, paper, a watch, and patience. Perhaps because of their emphasis on laboratory experimentation as a defining feature of their discipline, American psychologists have often seemed reluctant to include observational research methods as scientific techniques for studying psychological questions.¹ Even research that would seem, from a contemporary perspective, to be ideally suited as an observational study, like the characterization done in the 1920's by Arnold Gesell of the normal course of human development in the first 6 years of life, was carried out in a laboratory setting and followed strict protocols.

The introduction of observational methods into psychology has been due largely to the work of researchers in the British Isles and continental Europe. Their recognition of the importance of systematic observation for psychological research has been a significant contribution to modern psychology, a contribution clearly recognized by people outside the academic discipline. For example, in 1973, these methods received a major affirmation when three ethologists, Nikolaas Tinbergen, Konrad Lorenz, and Karl von Frisch, were awarded the Nobel Prize in Medicine for research on animal behavior that relied primarily on observational research methods in naturalistic settings.

Observing people (or animals) in natural settings (settings that have not been structured by the observer) comes closest to what we all do in our daily lives.² Thus, observational studies can bring the highest level of realism possible to scientific research. The cost for achieving this realism is that the observer often has little or no control over what happens and may not know all the factors that are relevant for understanding the data. The limitations of simple observation were illustrated in the story of Clever Hans, when the September commission failed to discover the basis for Hans' cleverness, leaving to Pfungst the task of introducing systematic variation in the visual cues available. Nonetheless, observations that have been made systematically and are preserved in a publically verifiable record can lead to considerable advances in our knowledge. Even observations that are made less systematically, but are integrated with a theoretical understanding of the situation, can provide important information about the meaning or

¹The animus against observational research has abated considerably in the past 30 years.

²“Reality” shows do not provide naturalistic observations, for two reasons: (1) the situations are artificially structured, and (2) what can actually be observed depends on what has been selected for presentation by the show's producers.

significance of what has been observed.

Observational research techniques include an enormous variety of methodologies, and it is not possible to describe them all in this chapter. Instead, the next section identifies two major dimensions along which research methodologies vary and discusses some general issues that apply to many, if not most, methods.

Observational Research Methodologies: 2 Dimensions of Variation

Qualitative versus quantitative methods

Perhaps the most important distinction is that between qualitative and quantitative approaches. *Qualitative research methods* are diverse and come from many different fields and intellectual traditions, but a general characterization has been offered by Norman Denzin and Yvonna Lincoln in *The Landscape of Qualitative Research* (1998), part of their multi-volume exposition of qualitative methods:

Qualitative research is multimethod in focus, involving an interpretive, naturalistic approach to its subject matter. This means that qualitative researchers study things in their natural settings, attempting to make sense of, or interpret, phenomena in terms of the meanings people bring to them. (Denzin & Lincoln, 1998, p. 3)

Qualitative research methods include ethnographic analyses, narratives and personal histories, focus groups, and films and photographic essays. Often, the methods involve collecting a lot of interview data that is then transcribed and analyzed to identify major themes and relationships among concepts. The process of data analysis has been facilitated in some cases by the use of computer software designed to extract themes from nonstructured data.³

Qualitative researchers evaluate the validity of their assessments along at least 5 dimensions. Using an organizational framework presented by Kazdin (2003), these are summarized in Table 2-1: descriptive validity, interpretive validity, theoretical validity, internal validity, and external validity. Validity is achieved in part through a process of finding commonalities among different ways of looking at the data, a process known as **triangulation**.

³One example of such software is NUD*IST, a program designed to analyze Non-numerical Unstructured Data by techniques of Indexing, Searching and Theorizing and produce a loosely organized network of ideas, showing their relationships to each other.

Table 2-1 Forms of Validity in Qualitative Research (after Kazdin, 2003)

Descriptive validity	The extent to which the account reported by the investigator is factually accurate. The account may reflect descriptions of events, objects, behaviors, animals, people, settings, times, and places
Interpretive validity	The extent to which the meaning of what has been described is accurately represented. Is the descriptive material understood adequately? Are the views, intentions, feelings, or other data interpreted in a way that represents or understands the experiences?
Theoretical validity	The extent to which the explanations designed to address how and why a phenomenon or experience has occurred, fit the data. This account is designed to be at a more abstract level of interpretation and to convey the possible reasons and underpinnings of the phenomenon.
Internal validity	The extent to which the influences identified by the investigator provide a necessary and sufficient account of the results. Are there other sources of influence that could have produced the outcomes described?
External validity	The extent to which the findings reported can be generalized across people, cultures, times and situations.

Quantitative research methods have a different focus from that of qualitative research. Rather than seeking to elaborate the rich detail of individual experiences, quantitative methods try to identify characteristics that are relatively general and are found among multiple individuals or across multiple situations. Alan Kazdin (2003), in a thoughtful comparison of qualitative versus quantitative research, describes quantitative research as research in which:

one tries to devise investigations to rule out threats to validity, to test specific hypotheses, to identify the impact of variables on some outcome of interest, and to analyze the data statistically. (Kazdin, 2003, p. 328).

Quantitative observational methods emphasize the same elements as survey and experimental methods, elements that will be elaborated in later chapters of this book.

Kazdin, who comes to research from the perspective of a practicing clinician, offers a useful table for comparing qualitative and quantitative approaches, and his table is reproduced here, with some modifications, as Table 2-2. As Kazdin notes, creating a table tends to give the misleading impression that distinctions are sharply defined, whereas in fact the contrasts between these approaches are much more blurred in actual practice. The best way to think of the two approaches is that they are complements to each other. Good research practices try to make use of both approaches.

Table 2-2. Comparison of Qualitative and Quantitative Research Methods (After Kazdin, 2003)

Characteristic	Qualitative	Quantitative
Goals	Describe and interpret experience; provide new insights, describe and explain with few or no initial hypotheses; generate theory	Test theory and hypotheses; identify causal relations, seek group differences and patterns
How to study	Consider variables as they appear in context with all natural influences; embrace all the complexity of the situation and its context	Isolate variables, control potential artifacts and extraneous influences; rule out rival hypotheses
Subjects	Study one or small number of cases intensively	Study a large number of cases for statistical power; strive for random or representative samples
Use of control conditions	No control group; elaborate on the richness of information about a particular group; describe commonalities and differences within the group	Control or comparison groups are usually included to address threats to validity
Role of subject	Participants are viewed as contributors to interpretation as well as sources of data; subjects are often consulted to ask whether the description and interpretation capture the experience	Participants are sources of data, but not of interpretation; subjects do not reflect on the data or help experimenter make sense of the data
Role of investigator	Investigator is <i>engaged</i> rather than <i>detached</i> and can understand better to the extent the situational meaning is experienced; perspectives are made explicit but not removed	Investigator is as detached as possible; perspectives are removed as much as possible from methods and presentation

The data	Narrative descriptions, full text, interviews, accounts, examples. Meaningful constructs are derived from data	Scores on measures that operationalize constructs; “data” typically refers to information that has been reduced to numbers in some fashion
Data evaluation	Interpret meaning of experience from participants’ point of view; identify themes, patterns, and key concepts that emerge from descriptions	Statistical analyses to find patterns and characterize variation, to identify impact of one variable on another
Criteria for knowledge	Descriptions are coherent and viewed by others as internally consistent, capturing the experience; procedures and findings can be replicated	Procedures and findings can be replicated; findings are sensible in light of quantitative theory
Major contribution	Understanding of experience is elaborated and brought to light in ways that extend our sense of understanding	New theory, relation or hypothesis is brought to light that increases understanding

An instructive example of blending the two approaches is found in a paper by Upshur and Bacigalupe (2001). This paper is from the Mas Salud project, which studies the availability of health care to the Latino community in Massachusetts. Upshur and Bacigalupe (2001) used focus groups to learn how people viewed their access to the health care system. After looking at statistical data on utilization of services, Upshur and Bacigalupe met with Latino health-care professionals and Latino users of health care services and asked for their perceptions of what sorts of access they had and what sorts of barriers limited their access. The comments obtained from the focus groups helped them understand what people were doing that gave rise to the statistical data.

Another good example is in the mental models approach to risk perception described by Morgan, Fischhoff, Bostrom and Atman (2002). More will be said about this approach later on.

Direct versus indirect observation

A second important distinction between different approaches to observational research is that between direct observation and indirect observation. Of course, the idea that observation could be indirect may seem contradictory - how can one make observations except by direct experience? With respect to research on other animals, the answer is that one cannot. In research on people, however, we can ask people to describe what they have done, felt, or observed, and their answers can be treated as alternative, but indirect, forms of observation. Note that the distinction between direct versus indirect observation is separate from the distinction between qualitative versus quantitative methodologies: direct and indirect observations can be

carried out in either qualitative or quantitative research. A focus group discussion, for example, is an example of qualitative research using indirect observations of people's behavior, whereas an ethnographic study typically involves direct observation. Similarly, quantitative research may be done using a survey, to provide indirect observations, or it may be done with an observer in a field setting.

For cases in which observations are made directly, a further distinction can be made (e.g., Shaughnessy & Zechmeister, 1997) between research in which the observer tries to avoid active intervention in the situation being studied and research in which the observer is actively involved in the processes that are the focus of the research. Examples of naturalistic observations without intervention include the early work of Jane Goodall on chimpanzees (e.g., Goodall, 1962) and of Dian Fossey on mountain gorillas (e.g., Fossey, 1974) as well as Timothy Perper's studies of flirting behavior in bars (Perper, 1986). Examples of naturalistic observation with intervention include participant observation studies in which a researcher becomes part of a group in order to describe what members of the group do. For example, Douglas Harper rode the rails from Minnesota to Washington in order to learn about hoboes and tramps and wrote about this experience in a qualitative research tradition (Harper, 1982). Other forms of intervention include the use of structured settings, like the conservation tasks developed by Jean Piaget, as well as the "field experiment", in which the researcher deliberately introduces a manipulation into a setting that is otherwise naturalistic. The classical ethologist, Niko Tinbergen, was a master in devising field experiments to study animals' behavior and showed through his success that this can be an invaluable technique to use. The spirit of his effort is seen in a study by Crusco and Wetzel (1984), who asked waitresses to vary how they returned change to a diner. The waitresses either touched the diner on the hand, touched the diner on the shoulder, or did not touch the diner at all. The measure of interest (both to the researchers and to the waitresses) was the size of the tip left by the diner: larger tips were left when the waitress touched the diner on the hand.

Doing observational research: 4 Decisions

Regardless of their general approach, researchers using observational methods must make decisions about specific features of their research. The list of 4 features reviewed here is based on Martin and Bateson (1993), who provide an excellent short guide to observational studies.

Categories, coding and operational definitions

Observations are eventually summarized in some fashion, whether by virtue of identifying themes or by counting examples of different behaviors. To summarize their observations, researchers need to define a set of categories or codes. For example, an observer watching children at play might code different actions as "solitary play", "parallel play", "interactive play" or "no play". *Solitary play* would refer to cases in which a child was playing alone, without involving any other child or adult. *Parallel play* would refer to cases in which two children were playing side by side, but their play was independent of each other, like two children separately coloring in coloring books or swinging on swings. *Interactive play* would refer to cases in which the play of one child involved another, as when children coloring separate images comment on what each other is doing, or when they decide to color a single image

together. *No play* would refer to all the cases in which the child was not playing.

Four aspects of coding are important. First, the codes must have *inter-observer reliability*; they must be ones about which independent observers agree. Ways to measure inter-observer reliability will be described later in the chapter. Second, the codes should be *exhaustive*; they must cover all the cases that might be observed. The categories of play mentioned in the preceding paragraph appear to provide an exhaustive list - children can either be playing alone, playing next to another child, playing with another child, or not playing. Third, the codes should be *mutually exclusive*; they must be assigned so that any given behavior is assigned only one code. The categories of play just mentioned clearly meet this requirement; a child can not both be playing alone and playing with another child, for instance. Fourth, the codes should be *useful*; they must be related to the question or hypothesis with which the research is concerned. Creating codes that exemplify the first 3 aspects of coding is usually straightforward, but creating codes that are useful is likely to require a process of development, in which an initial set of codes is tried out and then revised and tried out again. This process may take considerable time and effort, depending on the situation.

Inter-observer reliability. How easy or difficult it is to achieve high inter-observer reliability with a coding system depends on several factors. Coding that refers to events which occur as obvious units, for example, will be more reliable than coding that refers to events that are not clearly separated from other activity.

Operational definitions. An operational definition explains the meaning of a term by identifying the processes or procedures by which the term is measured. An operational definition of hunger, for instance, could be expressed in terms of the procedure by which food intake has been restricted - we might say that a person is hungry if that person has not been allowed to consume any food for 24 hours. Operational definitions serve to link psychological terms to observable phenomena, which is a necessary starting point for doing empirical research.

In observational research, the categories or codes used to classify behavior also typically serve as operational definitions. In the example used earlier of observing children at play, the category *parallel play* would also serve as an operational definition of parallel play.

Operational definitions are often expressed in terms that are very specific to a given research setting, and this can give them a particularity that seems excessive. Experience has shown, however, that the particularity is a necessary starting point for developing a definition of a phenomenon, because sometimes small changes in procedures lead to very different constructs. To make their definitions more general, psychologists show that different operational definitions produce the same pattern of results, in which case a more general definition is possible. This process of establishing a more general definition is known as **using converging operations**.

Sampling Rules

Sampling rules are the guidelines that specify when observations are made. These rules can be divided into two general classes, depending on whether the primary basis for sampling is determined by the occurrence of particular events, in which case one does **event sampling**, or the basis for sampling is determined by time, in which case one does **time-based sampling**.

Event sampling is used to record the occurrence of events of interest and to characterize what happens when the event occurs. For example, if one were studying the behavior of young adults in a bar, then one might start making observations when one person approached and spoke to another person. Or, if one were studying the behavior of students in a classroom, one might start making observations when the teacher asked a question of the class. One could also simply record the number of times a particular event or event sequence occurred in a particular situation.

Time sampling is used to record how often events occur and when they occur. In time sampling, observations are made for specific durations, after a given amount of time has elapsed, or at particular moments. Recording behavior at a playground for 30 minutes is an example of time sampling. More commonly, time sampling at specific times, after a given amount of time and the basis for making an observation is that a given amount of time has elapsed, that a particular certain time.

Recording Decisions

Three decisions about recording observations have a lasting influence on the data collected. The first decision concerns the medium by which recordings are externalized. The most readily available medium, and the easiest to use, is that of “paper and pencil”. Whether keeping a narrative account or placing marks on a behavioral checklist, people are very comfortable with using the physical medium of writing down information. Its ease of use is accomplished, however, at the cost of retaining only a very abstracted record of what has taken place. Even a narrative account has only the power of its words to evocatively recreate the events that were observed.

Modern technology has provided many alternative media for retaining a much richer record of observed events. Specialized devices, such as event recorders, for example, can keep a relatively complete record of the time and sequence of multiple events, stored either in an electronic form or as a paper record. Although usually more elaborate than simple paper and pencil records, event recorders keep track of an events at a relatively abstract level, without any of the sensory richness of the actual episode. Modern technology has, however, also provided the means for recording the auditory and visual aspects of experiences. Tape recorders and video cameras, for example, provide reproductions of the sounds or of both the sights and sounds of events in a form that can be examined by many observers and duplicated for distribution.

The richness of detail and extensiveness of the temporal record of events that are made possible by these more sophisticated technologies come with costs, however, and these are worth noting. One cost is the simple but obvious economic cost; both the devices for making the recordings and the media on which the recordings are kept are far more expensive than paper and pencil. A second cost is that the accessibility of the records becomes dependent on technology, and technologies tend to change unpredictably. In video recording, for example, the data storage medium has changed from ½ inch magnetic tape to digital flash memory sticks. Without a ½ inch video tape reader, an observer has no access to video recordings from the 1960's. In contrast, a paper record does not require a technological interface for access. Field notes made in the 1930's are still accessible seventy years later. A third cost is the cost of having more data than one can readily handle. The value of data lies not in having it but in being able to use it.

As a practical matter, one can have more data than can be processed.

The second decision to make about recording observations is whether to use continuous or intermittent recording. Note that this decision relates to the recording *within* an observation period. To use continuous recording means that one attempts to have a continuous record of events from the beginning to the end of the observation period, whereas intermittent recording means that recording is limited to certain periods during the observation interval.

From the perspective of having as complete a record as possible, continuous recording seems obviously preferable to any protocol that deliberately omits some of the events. What has to be recognized, again, is the practical question of whether one can process and store all the data that are obtained.

The third decision about recording, which is perhaps more subtle than the first two, is whether the recording is sequential or non-sequential in nature. It may seem strange that this is a decision, because we are accustomed to the fact that behavior takes place over time and is inherently sequential in our experience of it. However, observations must be recorded in particular ways if the record is to retain an accurate account of the sequential nature of events. It is worth noting that our ability to remember the actual sequence in which events occur is often poor, even when such information is very important. Consider the role of instant replay in professional football; there is usually good agreement that a player caught a ball and that the player has gone out of bounds, but which of those two occurred first and which occurred second is often the subject of heated disputes.

Calculating Inter-observer reliability: Cohen's kappa

Agreement among observers is a necessary element for distinguishing public events from private experiences, but such agreement can be assessed in ways that range from informal assent to formal evaluations. Measuring agreement in a well-defined way is essential for establishing the objectivity of observational methods, though.

One numerical measure of agreement is the *percentage of agreement*, which is calculated as the number of agreements (N_A) divided by the sum of the number of agreements and the number of disagreements (N_D), expressed as a percentage:

$$N_A = \frac{N_A}{(N_A + N_D)} \times 100\%$$

A more useful measure, however, is Cohen's kappa, κ , because it corrects for the possibility that two observers might agree or disagree just by chance. To illustrate the calculation, consider the data summary in Table 2-3. The table shows data from a hypothetical study, of people at a bar, in which two researchers distinguish among 4 behavior categories:

Unoccupied - Person is not engaged in anything specific or is an onlooker, watching other

people at the bar.

Solitary - Person sits alone and does not appear to be affected by what others are doing.

Together - Person is sitting with another person, but is not engaged in any interaction with that person.

Group - Person is physically with one or more other people and is engaged in eating, drinking, or conversational behaviors that are shared and interactive.

Table 2-3. Observational Data						
			Observer 2			
		Unoccupied	Solitary	Together	Group	Total
	Unoccupied	7	0	1	0	8
Observer 1	Solitary	1	24	0	0	25
	Together	1	1	17	4	23
	Group	0	0	3	41	44
	Total	9	25	21	45	100

The table is organized to show the agreements and disagreements between the two observers with respect to 100 observations of individuals at the bar. The righthand side of the table shows the total number of times Observer 1 made each of the 4 classifications among the 100 observations. The bottom row of the table shows the total number of times Observer 2 made each of the 4 classifications.⁴ Within the table, each cell shows the number of times that Observer 1 made one classification and Observer 2 made a particular classification. For example, in the first row, one cell shows that Observer 1 and Observer 2 both classified a behavior as “Unoccupied” on 7 occasions and Observer 1 classified a behavior as “Unoccupied” on 1 occasion on which Observer 2 classified the same behavior as “Together”.

The total number of agreements, N_A , is the sum of the cells on the main diagonal, namely, $7 + 24 + 17 + 41$, or $N_A = 87$. Because the total number of observations was 100, the percentage of agreements is $87/100 \times 100\% = 87\%$.

The percentage of agreement doesn’t take into account the possibility that the two observers might agree by chance, however. To correct for chance agreement, Cohen developed the measure, κ , that is calculated as follows:

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

⁴ These are called the *marginal totals*, because they are in the margins.

where \underline{P}_O is the observed proportion of agreement and P_C is the proportion of agreement expected by chance.

The observed proportion of agreement can be calculated by dividing the number of agreements, N_A , by the total number of observations, N . The proportion of agreement expected by chance is found from the marginal totals (the numbers in the righthand column and the bottom row) of Table 2-3. Multiply the marginal total for Row 1 by the marginal total for Column 1 (9 x 8), the marginal total for Row 2 by the marginal total for Column 2 (25 x 25), etc. and add all these products together, then divide by the grand total (100) multiplied by itself (100 x 100), to get

$$\begin{aligned} P_C &= \frac{(9 \times 8) + (25 \times 25) + (21 \times 23) + (17 \times 15)}{100 \times 100} \\ &= .225 \end{aligned}$$

Consequently, Cohen's kappa is

$$\kappa = \frac{.87 - .225}{1.0 - .225}$$

$$\kappa = .844$$