

Behavioral Observation and Coding

ROGER BAKEMAN

Some time ago students of behavior discovered a land called observational methods (Weick, 1968, 1985), but even today exactly where this land lies and the extent of its domain continue to be debated. Some observational nationalists claim almost unlimited territory. After all, almost every method – from unconstrained narratives (Brandt, 1972) to observing the pointer reading of an instrument measuring the amount of testosterone in saliva – depends on observation in some sense. Others define observational studies as nonexperimental (e.g., Good, 1994, p. 112) because participants are only observed to belong to one group or another and not randomly assigned. Still others (e.g., Pedhazur & Schmelkin, 1991) severely limit the land, describing only ratings and checklists as exemplars of systematic observation.

Nonetheless, I agree with Pedhazur and Schmelkin (1991) that observational methods are located not among the grand lands of independent theory (e.g., behavioral or cognitive) or overarching method (e.g., experimental or correlational) but in the important (one might say foundational) province of measurement. One of the founders was S. S. Stevens (1951),

I am greatly indebted to Vicenç Quera, with whom I co-authored *Analyzing Interaction: Sequential Analysis with SDIS and GSEQ*, and to John M. Gottman, with whom I co-authored *Observing Interaction: An Introduction to Sequential Analysis*. These two colleagues have been a pleasure to work with and have profoundly influenced my thinking regarding many of the matters presented in this chapter. I also thank Connie Russell, Craig Davis, and other students in my graduate observational methods course who provided helpful comments on an earlier draft.

Correspondence should be directed to Roger Bakeman, Department of Psychology, Georgia State University, University Plaza, Atlanta, GA 30303-3083 (e-mail: bakeman@gsu.edu).

who defined measurement as the assignment of numerals to objects or events according to rules. In the context of observational methods, measurement occurs when codes are assigned to events of interest in the passing stream of behavior, which may be either live or recorded, according to definitions and rules clearly described in explicit coding schemes. Thus observational methods may be used whenever the behavior of interest is clearly defined and available to external observers.

Often observational methods are used when studying human infants and animals other than humans, presumably because such nonverbal organisms can not respond to questions or answer questionnaires. Topics studied include such diverse matters as birds courting, monkeys fighting, children playing, and mothers and infants exchanging gleeful vocalizations (Bakeman & Gottman, 1986). But observational methods are also useful for studying verbal behavior, such as couples discussing and therapists and clients conversing, and can even probe internal processes, for example, by having participants view videotapes of their own behavior and provide comments and explanations (Ickes, Bissonnette, Garcia, & Stinson, 1990). Thus the methods can be used in a variety of research contexts (e.g., experimental and correlational) and in the service of diverse theoretical orientations. Still, the methods seem especially useful when the behavior of interest is social, involving interaction between two or more participants, and when hypotheses emphasize not outcomes but process, that is, the ways and means by which interaction unfolds in time.

The remainder of this chapter describes issues, concepts, and techniques that are relevant when systematic behavioral observation and coding are used as a means of measuring behavior. The intent is to

provide readers with enough sense of the lay of the land (the work required, tools currently available, etc.) that they can decide whether their theories and concepts are compatible with, and whether their research would profit from, the application of observational methods.

Still, some general guidance is possible. Think of observational methods as another measurement tool in your toolbox, lying alongside the self-report instruments and other questionnaires that are probably already there. Ask yourself: Do you ever have questions about the validity of the scores derived from pencil and paper measures? How is the behavior of interest to you manifest? Is it private, in the sense that only the target of your investigations can tell you about it? Or is it public, accessible to view by others in some way? Remember, accessible behavior includes, for example, viewing children negotiate how they will play a game (either live or on videotape later), viewing two people discussing topics of dissension between them or reading transcripts of their discussion made from audio or video recordings, or even reading diaries that individuals made about their social interactions. Do your questions concern how often particular behaviors occur, such as conflicts during negotiation, agreements during discussion, or lies during social interaction? Or are you interested in process, and so pose questions concerning how particular behaviors are sequenced? Do you think that using more than one measurement modality would enhance the validity of your study? If you answered yes to any of these questions, you may well find observational methods a useful addition to your measurement toolbox.

BEHAVIORAL OBSERVATION IS SYSTEMATIC

The statement, behavioral observation is systematic, is part definitional, part desire – a goal that is sometimes only imperfectly achieved. In this chapter, I focus on methods of behavioral observation that are systematic; thus *systematic*, even when not stated explicitly, is understood as part of the definition. To qualify as systematic, procedures, should be public (all details of method are clearly described and so accessible to others) and replicable (given similar circumstances, other observers would provide similar results); thus personal characteristics of observers are ordinarily not an issue for systematic observation.

Unsystematic observation is perhaps the least satisfactory term for the alternatives because it does not convey their richness, historical depth, and continued importance. Long before academic social science

attracted adherents (primarily in Europe and North America, primarily in the 20th century), insightful studies of human behavior appeared in the form of plays, novels, essays, and other treatises, from Lady Murasaki's *The Tale of Genji* to Machiavelli's *The Prince*, among others. In this century, the tradition of humanistic studies has continued to exert influence in the social sciences, although some modern names for it include participant observation, ethnomethodology, phenomenology, qualitative research, or plain old-fashioned journalism. Examples are legion, but include books such as Erving Goffman's *Asylums* (1962), Robert Coles's *Children of Crisis* (1967), and Oscar Lewis's *Children of Sanchez* (1961) and films such as Frederick Wiseman's *Titticut Follies*, Jenny Chamberlin's *Paris is Burning*, and James, Marx, and Gilbert's *Hoop Dreams*.

At their best, qualitative studies such as those just listed represent lively intelligence and informed sensibility. But when not done well, such studies can seem ill-informed, self-contained, and self-indulgent. Yet how do we judge? How do we know when work is good, or at least acceptable? How do we respond to the criticism that authors may have viewed phenomena from a certain preconceived point of view that unduly biased their reports and interpretations? Systematic methods of quantification – including observational methods – have been developed, in part, to solve the sorts of problems qualitative studies present. The intent is to insulate findings from the biases and desires of observers. Systematic observations may have their limits (e.g., predefined codes limit what is seen, or at least recorded) but normally allow us to judge results independently of the observers' personal qualities (other than training). When judging qualitative studies, qualities of the interpreter strongly influence the judgment, but data that result from systematic measurement are rarely acclaimed or discounted because of the particular observers used. As a result, whereas qualitative studies are often viewed as idiosyncratic, we expect quantitative studies to be replicable.

Nonetheless, qualitative studies play an important, precursory role in systematic behavioral observation. Such studies are essential when developing behavioral codes, as discussed in a subsequent section, and often cause investigators to rethink their hypotheses. Thus, for some investigators qualitative studies are an end in themselves, but for users of behavioral observation they help inform development of both research questions and the measuring instruments used (Bakeman, 1991).

SYSTEMATIC OBSERVATION IS CATEGORICAL MEASUREMENT

As I define matters, observation becomes systematic when measurement occurs, that is, when a code (i.e., a category) is assigned an event. In ways I discuss later, this sentence is deceptively simple, but as an example consider how we might measure the act of measurement itself. S. S. Stevens (1951) provided what has become a widely accepted coding scheme for measurement scales. He named four types. To determine which applies (i.e., to measure the act of measurement), all we need do is consider the responses to a series of three yes-no questions:

1. Are codes ordered in some natural way?
 - a. No, then the scale is *categorical*.
 - b. Yes, then:
2. Are intervals between codes equivalent?
 - a. No, then the scale is *ordinal*.
 - b. Yes, then:
3. Does zero indicate truly none of the quantity measured?
 - a. No, then the scale is *interval*.
 - b. Yes, then the scale is *ratio*.

This example presents a coding scheme consisting of four codes. Arguably, these codes are themselves ordinal, in the sense that the quantitative meaning of the scale points increases from categorical, to ordinal, to interval, to ratio scale (although one could argue that the distinction between categorical and other scale types is itself categorical). Usually, however, the codes applied to behavioral observations are clearly categorical (e.g., agrees, approves, complains, excuses, interrupts); that is, the order of the codes is arbitrary and not dictated by some underlying common dimension. Furthermore, determining the appropriate code is an act of measurement, one that defines systematic observation.

The statement, systematic observation involves categorical measurement, is somewhat arbitrary and (deliberately) limiting. Measurement is essential, by definition, but events might be measured on scales other than categorical. For example, we might apply valances to codes like agrees, approves, and so forth, representing some underlying scale of positiveness. Then a sequence of coded statements would become a sequence of numbers representing points on an ordinal scale; or every 5 seconds we might rate a person's sensitivity when interacting with a partner on a 1-7 scale; or, a physiological measure like heart rate could be sampled every second, resulting in a sequence of numbers measured on a ratio scale (Gottman & Bakeman, 1996).

The last example does not quite fit the definition used here for behavioral observation because numbers are recorded automatically by an instrument, unaided by human judgment, and ordinal scaling as in the first two examples is relatively rare. All three examples, however, are amenable to other sorts of analyses; possibilities include time series analysis (Gottman, 1981), hierarchical linear models (Bryk & Raudenbush, 1992), and event history analyses (Allison, 1984; Singer & Willet, 1993).

In this chapter, I limit discussion to behavioral observations made by human observers (not automatic instruments) using categorical codes. Human observers and categorical data entail particular approaches to reliability and data analysis, and both are discussed later. Instruments entail a different set of problems (see Gottman & Bakeman, 1996) and usually result in at least ordinal data, which require different analytic approaches (see Gottman, 1981) than the ones discussed here for categorical data. Still, readers should be familiar with other possibilities and recognize that the approaches detailed here can often be combined with other approaches in extremely fruitful ways.

BEHAVIORAL SEQUENCES CAPTURE PROCESS

The measurement approach described here – human observers assigning codes to selected events – can be labor intensive and time consuming so it is reasonable for investigators to ask what might justify such effort. For example, if one wanted only an assessment of the effectiveness of a discussion group leader or the sensitivity of a mother when interacting with her infant, a rating scale approach (observers rating effectiveness or sensitivity from 1 to 5 after watching for a period of time) might seem preferable to a more microanalytic system (observers recording onset and offset times for a list of selected behaviors). Yet if one wants to capture process – What leadership strategies push groups into conflict as opposed to agreement? What maternal behavioral gambits seem to capture infants' attention? – some characterization of behavior unfolding in time and utilizing more detailed codes seems required.

Still, observational methods need not be sequential in a moment-by-moment continuous sense. For example, in a classic study Parten (1932) coded the play behavior of toddlers and preschoolers. One-minute time samples for selected days occurring over several months were coded unoccupied, onlooking, and solitary, parallel, associative, and cooperative play. From these data Parten was able to derive time-budget profiles (estimated proportions of time that different

children devoted to the various activities) for each child, but she was not able to describe how play unfolded for these children during play sessions. In contrast, Bakeman and Brownlee (1980) coded successive play states for individual children continuously during play sessions, and so were able to discover that parallel play often served as a bridge into more complex play (i.e., more complex play was coded after parallel play more often than chance would suggest; for additional details, see Bakeman & Gottman, 1986).

All of these possibilities – summary assessments (based, e.g., on rating scales), time-budget profiles, and more fine-grained characterizations of process – have their uses. The largest investment of resources is demanded when detailed sequential records are made (typically, using current technology, by recording onset and offset times for selected behaviors, usually from videotape). Summary assessments and time-budget profiles can be derived from the more detailed sequential records (and usually are), but almost always an investigation of process provides the justification for collecting such records in the first place. Thus, although observational methods do not always result in behavioral sequences being recorded, the uses emphasized in the remainder of this chapter assume that behavioral sequences and not isolated codes were recorded.

DEVELOPING BEHAVIORAL CODES

As noted previously, behavioral codes, structured into coding schemes or sets of coding schemes, form the measuring instruments of observational research. Each code can be regarded as a category of behavior (asks question, waves arm, etc.), and each scheme as a dimension of interest (speech function, body movement, etc.). Almost always, codes grouped in a scheme are mutually exclusive (only one code applies) and exhaustive (some code applies), although codes from different schemes may cooccur; indeed, such cooccurrences are often of considerable substantive interest. Thus the essential activity of code development begins with attempts to define the relevant dimensions and the codes each comprises.

Where Do Codes Come From?

It is easy enough to ask colleagues to define the relevant dimensions of the phenomena they are studying currently. Far more difficult is providing clear and coherent answers when similar questions are put to us. Considerable conceptual clarity is required. Further, once dimensions are named, the categories that

appropriately characterize each dimension do not always follow automatically. Ultimately, of course, answers to where-do-codes-come-from questions depend on our ontological commitments. For example, some students of animal behavior, when constructing an ethogram, might claim they are only listing the discrete behaviors of which a particular species is capable (e.g., Tuculescu & Griswold, 1983); in that sense, the codes simply reflect physical reality, and discovering the codes is an important part of the research endeavor. In contrast, some students of human development, when developing codes, might emphasize that the codes – like human language itself – are arbitrary constructions, useful within a specified theoretical or cultural context (e.g., Rogoff, Mishy, Göncü, & Musier, 1993).

On a more practical level, and no matter one's philosophical underpinnings, coding schemes – to be useful – should derive in a fairly direct manner from the research questions asked. As a first step, we should identify other investigators who have asked similar questions using observational methodology and attempt to adopt their coding schemes outright or else adapt them appropriately. Common coding schemes used by several investigators introduce comparability, which permits the accumulation of results and so strengthens the research literature. Marital interaction, for example, is an area of research that has benefited from common coding schemes developed and used by a coherent cadre of investigators (e.g., see Weiss & Summers, 1983; cf. Bakeman & Casey, 1995).

Yet, arising perhaps from some misplaced territorial imperative, too often investigators create measuring instruments anew (coding schemes as well as personality and other scales) without first determining that no others exist. This makes as much sense as each student of heat developing his or her own thermometer. Thus, coding schemes should be adopted, or at least adapted, from ones existing in the literature whenever possible, but not if doing so violates a superior principle: Codes must fit both the behavior observed and the questions asked.

Several steps can help ensure that this is so. First, of course, questions need to be refined and clarified. Then the behavior of interest needs to be observed, carefully and openly, as though one planned qualitative narrative reports. Tentative coding schemes need to be piloted, discussed, and refined in the presence of the behavior (or, more conveniently, a videotaped record). All of this takes time and emotional and intellectual energy and requires continual clarification of both underlying questions and individual codes. Often the list of codes grows to include behaviors that

can be discriminated and seem momentarily important, or at least salient, but that ultimately should be pruned in the interests of economy. A simple test helps: If investigators can explain how each proposed code will help answer a research question, it should be retained, but if not, perhaps it should be discarded. In any event, the importance of well-designed and appropriate coding schemes should not be underestimated. They serve as the lens of observational research; if initially faulty, subsequent data manipulation rarely improves the view.

Codes and Coding Units

The definitions of codes cannot easily be separated from the entities (events, time intervals, etc.) to which they are assigned. It makes sense to call the thing coded a *coding unit* because this reminds us of the presence in any study of different kinds of units, each of which may serve a different purpose. For example, participants assigned to treatment conditions are often called *experimental units*, whereas participants (individuals, dyads, families, etc.) in nonexperimental studies are called *study units*. Assumptions are often made concerning such units (e.g., their independence), and such units often figure prominently in statistical analyses. Thus, experimental or study units form components of the design of a study, whereas coding units reflect data recording procedures. Further, as we discuss shortly, representing the recorded data for analysis may depend on yet different units (e.g., units of time).

Often observational studies use only one kind of coding unit. For example, only turns of talk during marital interaction or only play states during preschooler free play might be coded. Other studies might use two or more coding units; for example, an investigator might code both infants' attention states and mothers' interactive strategies. However, there is a third possibility, which often turns out to be quite interesting: Coding units may be related hierarchically, much as real-life behavior seems to be. A good example is the child negotiation study conceived by David Bearison with the help of Bruce Dorval and other colleagues at the Graduate School and University Center of the City University of New York and presented here in a simplified version (Bearison, Dorval, LeBlanc, Sadow, & Plesa, in press). Like many of their colleagues, they used video recorders initially to preserve behavior of interest. Subsequently, coders viewed the videotapes, detecting and recording occurrences of various behaviors. Often onset times for some behaviors, and onset and offset times for others, are noted, using the time

display that is easily made a part of the picture with the sort of video equipment readily available today.

Bearison is not interested in every moment on the tape but only in certain times when the children (who have been instructed to coinvent a board game) enter into what he terms a *negotiation episode*. Asking coders first to identify episodes of a certain sort before moving on to another level of coding is a relatively common and useful strategy. For example, other investigators might be interested in episodes of children's conflict or cooperation, spouses' argument or tenderness, and client-therapist stalemate or breakthrough. The primary coding unit for Bearison consists of a negotiation episode but, reflecting his interest in the process of negotiation itself, he asks coders to consider a second coding unit: the conversational turns individuals take, which are related hierarchically to (i.e., are nested or embedded within) negotiation episodes. Thus coding proceeds on two levels. First negotiation episodes are detected and coded, then the conversational turns within them are coded.

Examples of Coding Schemes That Have Proved Useful

As argued earlier, coding schemes must reflect whatever ideas and questions drive a particular research endeavor; thus principled adaptation but not indiscriminate borrowing is encouraged. Nonetheless, examples are helpful. Literally hundreds are to be found in the literature, several are presented in Bakeman and Gottman (1986, chapter 2) and Bakeman and Quera (1995a, chapter 2), and a few have already been presented here (e.g., Parten's, 1932, scheme for coding children's play). Three additional examples, both of which reflect a moderate amount of complexity and a particular theoretical base, are the child negotiation study introduced earlier, a second study concerned with mother-infant-grandmother triadic interaction, and a third study concerning lying in everyday life.

As described earlier, negotiation episodes and the conversational turns nested within them were coded for the child negotiation study. Several dimensions of negotiation episodes were of concern, including their *topic* and their *outcome*. Topic codes included

1. rules or methods of play and
2. goals or purpose of playing;

and negotiation outcome codes included

1. unresolved,
2. passive acquiescence,

