2

Developing a coding scheme

2.1 Introduction

The first step in observational research is developing a coding scheme. It is a step that deserves a good deal of time and attention. Put simply, the success of observational studies depends on those distinctions that early on become enshrined in the coding scheme. Later on, it will be the job of observers to note when the behaviors defined in the code catalog occur in the stream of behavior. What the investigator is saying, in effect, is: This is what I think important; this is what I want extracted from the passing stream. Yet sometimes the development of coding schemes is approached almost casually, and so we sometimes hear people ask: Do you have a coding scheme I can borrow? This seems to us a little like wearing someone else's underwear. Developing a coding scheme is very much a theoretical act, one that should begin in the privacy of one's own study, and the coding scheme itself represents an hypothesis, even if it is rarely treated as such. After all, it embodies the behaviors and distinctions that the investigator thinks important for exploring the problem at hand. It is, very simply, the lens with which he or she has chosen to view the world.

Now if that lens is thoughtfully constructed and well formed (and aimed in the right direction), a clearer view of the world should emerge. But if not, no amount of corrective action will bring things into focus later. That is, no amount of technical virtuosity, no mathematical geniuses or statistical saviors, can wrest understanding from ill-conceived or wrong-headed coding schemes.

How does one go about constructing a well-formed coding scheme? This may be a little like asking how one formulates a good research question, and although no mechanical prescriptions guaranteeing success are possible, either for coding schemes or for research questions, still some general guidelines may prove helpful. The rest of this chapter discusses various issues that need to be considered when coding schemes are being developed.

2.2 What is the question?

Perhaps the single most important piece of advice for those just beginning to develop a coding scheme is, *begin with a clear question.* For any child, mother–infant pair, couple, animal, or group one might want to study, there are an infinite number of behaviors that could be coded. Without the focus provided by a clear question, it is hard to know just where to look, and it is very easy to be overwhelmed. We have all probably experienced at one time or another how operational goals can take over an activity. For example, a consultant we knew once found the employees of a large college athletic complex diligently recording who used what facility for how long in ways that had become ever more elaborate over the years. It seemed all too easy to think of more complicated ways to encode the passing stream of behavior, but that encoding seemed completely unguided by any purpose. No doubt once there had been some purpose, but apparently it had been forgotten long ago.

Similarly, many investigators, ourselves included, seem tempted to include more and more separate codes, and make ever finer distinctions, simply because it is possible to get observers to record data using such schemes. There is an argument to be made for such a "wide-net" strategy, and it usually goes something like this: Because we are not really sure what will turn out to be important, we need to record everything – or at least lots – and then scan our voluminous data archive for important features. Somehow, it is hoped, coherence will emerge from the multitude. However, we suspect that this happens very rarely, and so when asked, "But what do I do with all this data that I collected so laboriously?" our first suspicion is that a clear question, and not statistical expertise, was lacking. Typically, "categorical overkill" seems to inundate investigators in tedious and not very fruitful detail, whereas studies involving clearly stated questions and tightly focused coding schemes seem far more productive.

For example, consider the following question. Among monkeys, whose early days are spent clinging to their mothers, is it the infants who first leave their mothers as they begin to explore the world, or is it the mothers who first push their reluctant infants out into the world? Given this question, most of us would probably agree on the behavior to be recorded and how it should be categorized. We would want to record separations between mother and infant and categorize each one as being either initiated by the mother or initiated by the infant. Thus our coding scheme would contain two codes – Infant-Initiated Separation and Mother-Initiated Separation – and recording would be "activated" whenever the event of interest – separation – occurred. About the only way to complicate this simple example would be to add a third code: Mutual Disengagement.

With such a simple coding scheme, the progression from data collection to analysis to interpretation would be simple and straightforward. We might find, for example, that over the course of the first several months of life the number of separation episodes gradually increased and then decreased. At first, there would be few separations because the infant is almost always clinging to the mother, whereas later on there might be few because the infant is almost always off the mother, but in between there would be more movement from and to the mother. Further, the data could show that when the number of separations is first increasing, the mother initiated considerably more of them than her infant, whereas later, when the number of separations begins to decrease, the infant initiated more, leading us to conclude that it is the mothers who first push their presumably reluctant infants out into the world.

The point is, developing a coding scheme is theoretical, not mechanical work. In order to work well, a coding scheme has to fit a researcher's ideas and questions. As a result, only rarely can a coding scheme be borrowed from someone else. However, when research questions are clearly stated, it is a much easier matter to determine which distinctions the coding scheme should make. Without clear questions, code development is an unguided affair.

2.3 Physically versus socially based coding schemes

Regardless of whether coding schemes should be borrowed or not – and our previous paragraph objects only to mindless borrowing, not borrowing per se – the fact of the matter is that only rarely are coding schemes used in more than one laboratory. This might seem an undesirable state of affairs to those who believe that scientific work should be replicable. After all, all laboratories interested in temperature use thermometers.

The reason for the commonality of thermometers and the multiplicity of coding schemes has to do, we think, with the nature of the "stuff" being measured – temperature or social behavior. Now it may be that some aspects of social behavior can be measured as precisely and commonly as temperature. Whatever the case, we think it is helpful, when developing a coding scheme, to distinguish between codes that are physically based, on the one hand, and socially based, on the other.

At the risk of precipitating a lengthy discussion with the philosophically inclined, let us state that it seems possible to locate coding schemes along a continuum. One end would be anchored by physically based schemes – schemes that classify behavior with clear and well-understood roots in the organism's physiology – whereas the other end would be anchored

by socially based schemes – schemes that deal with behavior whose very classification depends far more on ideas in the mind of the investigator (and others) than on mechanisms in the body. We have called these "socially based," not because they necessarily deal with social behaviors – even though they typically do – but because both their specification and their use depend on social processes. Instead of following quite clearly from physical features or physical mechanisms in a way that causes almost no disagreement, socially based schemes follow from cultural tradition or simply negotiation among people as to a meaningful way to view and categorize the behavior under discussion. Moreover, their use typically requires the observer to make some inference about the individual observed.

For example, some people are paid to determine the sex of baby chicks. The "coding scheme" in this case is simple and obvious: male or female. This is not an easy discrimination to make, and chicken sexers require a fair amount of training, but few people would suggest that the categories exist mainly as ideas in the observers' heads. Their connection with something "seeable," even if difficult to see, is obvious.

Other people (therapists and students influenced by Eric Berne) go about detecting, counting, and giving "strokes" – statements of encouragement or support offered in the course of interaction. In effect, their "coding scheme" categorizes responses made to others as strokes or nonstrokes. For some purposes, therapeutic and otherwise, this may turn out to be a useful construct, but few would argue that "strokes" are a feature of the natural world. Instead, they are a product of the social world and "exist" among those who find the construct useful. Moreover, coding a given behavior as a "stroke" requires making an inference about another's intentions.

Other examples of physically and socially based coding schemes could be drawn from the study of emotion. For example, Ekman and Friesen's (1978) Facial Action Coding System scores facial movement in terms of visible changes in the face brought about by the motion of specific muscle groups (called action units or "AUs"). The muscles that raise the inner corner of the eyebrows receive the code "AU1." The muscles that draw the brows down and together receive the code "AU4." When these muscles act together, they result in a particular configuration of the brow called "AU1+4." The brows wrinkle in specific ways for each of these three action units (see Figure 2.1).

The brow configuration AU1+4 has been called "Darwin's grief muscle" as a result of Darwin's 1873 book on the expression of emotion. This leads us to the point that AU1 and AU1+4 are both brow configurations typically associated with distress and sadness. In fact, there is not always a one-to-one correspondence between sadness or distress and these brow configurations. For example, Woody Allen uses AU1+4 as an underliner

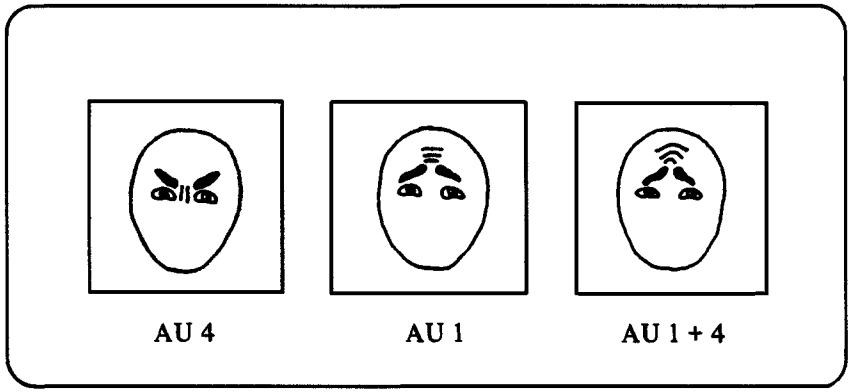

Figure 2.1. Examples of action units from Ekman and Friesen's (1978) Facial Action Coding System.

after he tells a joke. But in most social interaction there are additional cues of sadness or distress.

In a physically based coding scheme we would be coding such things as specific brow configuration, but in a socially based coding scheme we would be coding such things as sadness. The socially based coding scheme requires considerably more inference, and probably requires sensitive observers. However, it is not necessarily less "real" than the system that records brow action. It is simply a different level of description.

Not all researchers would agree with this last statement. Some, especially those trained in animal behavior and ethology, might argue that if the problem is analyzed properly, then any socially based scheme can and should be replaced with a physically based one. We disagree. In fact, we think there are often very good reasons for using socially based coding schemes.

First, it is often the case that physically based schemes, like Ekman and Friesen's mentioned above, are time consuming to learn and to apply, and therefore, as a practical matter, it may be much easier to use a socially based alternative. Even if it is not, a socially based scheme may more faithfully reflect the level of description appropriate for a given research issue. In any given case, of course, investigators' decisions are influenced by the problem at hand and by the audience they want to address, but it seems worth asking, before embarking on an ambitious observational study, whether something simpler, and perhaps more in tune with the research question, would not do as well. Some people will be unhappy with any coding scheme that is not clearly grounded in the physical world, but others, ourselves included, will be tolerant of almost any kind of coding scheme, even one that is quite

inferential, as long as it fits the research concerns and as long as trained observers acting independently can agree (see chapter 4).

Second, often various aspects of the socially created world are exactly what we want to study, and in those cases socially based coding schemes are the appropriate "filter" to capture the behavior of interest. Sometimes investigators are advised to avoid coding schemes that require observers to make any judgments about intentionality. In one sense, this is sound advice. Just from the standpoint of observer agreement, there is no reason to ask observers to make more complex judgments (was the baby trying to get her mother's attention?) when simpler ones will do (did the baby approach within 1 meter of her mother and look at her?). But in another sense, if this advice were followed rigorously, we would end up not studying some very interesting behavior, or else defending some possibly quite silly coding schemes.

For example, consider the Parten-like coding scheme for play state, described above, that Peter Smith used. When using this scheme, observers need to discriminate between children playing alone and playing with others, whether in parallel or in a more engaged state. Now one crucial distinction is between being alone and being "with" others; however, whereas observers usually have little trouble agreeing when children are with others, it is not an easy matter to define "withness" in a mechanical way. Is it a matter of proximity? If so, then exactly how near? Is it a matter of visual regard? If so, then how frequently and/or how long? Or does it also include elements of posture, orientation, activity, etc.? We have two choices. Either we define "Alone," "Parallel," and "Group" play in a mechanical way, dependent solely on physical features, and run the risk of occasionally violating our own and our observers' intuitions about the "true" state of affairs; or we view the difference between being alone and being with others as determined by the children's intent and experienced by them subjectively, and we make use of the ability that humans (in this case, our observers) have to judge, more or less accurately (as determined by agreement or consensus), the intentions of others in their culture.

The question is, should we deny or make use of these very human abilities (see Shotter, 1978)? In general, we think we should use them, and that it is especially appropriate to do so when studying social behavior in our fellow humans. (We recognize that this argument begins to break down when animals are being studied.) This is not a new issue in psychological studies, and there is little hope that it will be settled here. Even so, when we argue for more "mentalistic" categories than some traditional psychologists or animal behaviorists would admit, we do not intend this as a license for overly rich and undisciplined speculation and inference. Other things being equal, we much prefer coding categories that are defined in terms of observable

and concrete features, but we are not willing to let this one consideration override all others, especially if the meaningfulness of the data collected or the ability of those data to answer the question at hand might suffer. It is possible, for example, to record accurately how many times an infant approaches within 1 meter of his or her mother and what proportion of time was spent in contact with the mother, within 1 meter of the mother, and looking at the mother, and still not be able to gauge validly the quality of the mother–infant relationship. That task might be accomplished better, as Mary Ainsworth, Alan Sroufe, and others have argued, by asking observers to rate, on 7-point scales, how much the infant seeks to maintain contact with the mother, resists contact with the mother, etc., and then assessing the relationship on the basis of the pattern of the rating scales (see Ainsworth, Blehar, Waters, & Wall, 1978; Sroufe & Waters, 1977).

Still, when one is developing coding schemes (or rating scales, for that matter), it is a very useful exercise to describe each behavior (or points on the rating scale) in as specific a way as possible. For example, Bakeman and Brownlee (1980), in their parallel play study, required observers to distinguish between children who were playing alone and who were playing in parallel with others. First, Bakeman and Brownlee viewed videotapes of 3-year-olds in a free play situation. They continually asked each other, is this child in solitary or parallel play? – and other questions, too – and even when there was consensus, they tried to state in as specific terms as possible what cues prompted them to make the judgments they did. They were thus able to make a list of features that distinguished parallel from solitary play; when engaged in "prototypic" parallel play, children glanced at others at least once every 30 seconds, were engaged in similar activities, were within 1 meter of each other, and were no more than 90 degrees away from facing each other directly.

These features were all described in the code catalog (the written description of the coding scheme), but the observers were instructed to treat these defining features as somewhat flexible guidelines and not as rigid mandates. Once they had thoroughly discussed with their observers what most writers mean by parallel play, and had described parallel play in their code catalog, they were willing to let observers decide individual cases on the basis of "family resemblance" to parallel play. We agree with this procedure and would argue as follows: By not insisting that observers slavishly adhere to the letter of the rules, we then make use of, instead of denying, their human inferential abilities. However, those abilities need to be disciplined by discussion, by training, and – perhaps most importantly – by convincing documentation of observer agreement. The result should be more accurate data and data that can "speak" to the complex questions that often arise when social behavior and social development are being studied.

2.4 Detectors versus informants

In the previous section we suggested that it is useful to distinguish between physically based and socially based coding schemes and mentioned a number of implications of this view. We would now like to suggest one final implication of this view, one that affects how the observers' work is regarded and, ultimately, how observers are trained. It is fairly common to regard observers merely as "detectors," as instruments like any other whose job is passively to record what is "there," much as a thermometer records temperature. For physically based coding schemes, this seems like an appropriate view. In such cases, it is not the human ability to make inferences about fellow humans that is being exploited by the investigator, but simply the quite incredible ability of the visual–perceptual system to "see what is there."

For socially based coding schemes, however, it makes sense to regard the observer more as a cultural informant than as a detector. When socially based distinctions are being made, distinctions that typically require some inference about others' intentions, this strikes us as a more accurate way for both the researchers and the observers to think about their task. At least one code catalog that we know of, which was developed during the course of a study of citizen–police interaction, presents the observers' task to them exactly in this way (Wallen & Sykes, 1974). We think that this is a good idea, and that observers using socially based coding schemes do a better job if they appreciate the "participant–observer" aspects of their observation task.

2.5 The fallacy of equating objectivity with physically based schemes

Sometimes physically based coding schemes are mistakenly labeled as more "objective" than socially based systems. Let us consider why this is a fallacy. In the study of affect, one of our colleagues, Mary Lynn Fletcher, pointed out that if you employed observers to study the conversation of Chinese people, it would be absurd to hire observers who could not understand Chinese, claiming that this makes them more "objective." In an analogous way, one can argue that to code complex social processes such as emotions, it is often necessary for observers to be able to understand the "language" of affect. This requires that coders be expert cultural informants about emotion and also be able to respond subjectively. The key to clarifying the issue of "objectivity" is to stress, not objectivity as such, but replicability instead. Independent observers should agree.

There is a second issue inherent in the physically based versus socially based systems, namely, the level of description that is of interest to the investigator. For example, expert cultural informants have a complex catalog of physical cues that they can call upon to decide that a husband is angry in a marital discussion. If an investigator is interested in emotional categories such as angry, sad, contemptuous, etc., then it may be fruitless to attempt to construct a complete list of cues for each emotion judgment. Furthermore, it may not even be possible because the kinds of cues that go into a judgment of anger could be infinite, varying with context, the words spoken, pause, stress, rhythm, amplitude, major or minor key, contours of speech, facial expressions, gestures, and the nonadditive interaction of these features. We may not be interested in a description such as the following:

> At 5 : 19 the husband lowered his brows, compressed his lips, lowered the pitch of his voice, disagreed with his wife, while folding his arms in the akimbo position.

Instead, we may simply wish to describe the husband as angry. As long as we know what we mean by "angry," which will be required if we are ever to achieve reliability between observers, incidents such as the one previously described *illustrate* the construct of anger. They need not *exhaust* it.

In other words, competent observers can make extremely complex social judgments reliably and need not be viewed as inherently undesirable measuring instruments. At times they are exactly what we need, and their subjective observational skills can be precise and the best scientific choice.

2.6 Keeping it simple

Some readers, because of their intellectual convictions, the nature of the problems they study, or both, may not have found the foregoing discussion very helpful. Still, they could probably agree with advice to keep coding schemes simple. Like much good advice, this is so familiar that it seems trite, yet like maintaining liberty, maintaining simplicity also requires constant vigilance. The "natural" press of events seems against simplicity, at least with respect to coding schemes. The investigator's question may be only vaguely apprehended, as discussed earlier, a coding scheme may have been borrowed from someone else inappropriately, the conceptual analysis on which a coding scheme is based may not have been worked through sufficiently, or a proposed scheme may not have been adequately refined through pilot observations and critical discussion. All of these circumstances can confuse the issue and result in overly elaborated and unwieldy coding schemes.

Several points may be useful for keeping the coding system simple. First, it is important to have clear conceptual categories that are essentially at the

same level of description. For example, it is not desirable to ask coders to decide whether an utterance was a question and at the same time whether the utterance was responsive to the previous speaker. These judgments are made on two different levels of conceptual complexity. They should not be combined in one coding scheme. Second, codes ought to be reasonably distinct; that is, behavior categories should be homogeneous, which means that even when acts appear somewhat similar, they should not be put in the same category if there is good reason to think that either their causes or their functions are different. Third, it is better to "split" than to "lump" codes. Behavior categories can always be lumped together during data analysis, if that seems useful, but behaviors lumped together by the coding scheme cannot be split out later (unless we have a taped record and recode specific moments).

2.7 Splitting and lumping

At some point in discussions of this sort, the "level of analysis" problem is frequently broached (see, e.g., Hartup, 1979; Suomi, 1979). We can conceive of almost any phenomenon as consisting of levels, hierarchically arranged, with larger and more inclusive or more molar concepts occupying each higher level, and smaller and more detailed or more molecular concepts occupying each lower level. Then the question arises, what level should our coding categories occupy? In the abstract, without the context of a particular question, an absolute answer to this question is hardly possible, but a relative one is. First we need to decide what conceptual level seems appropriate for the question at hand (which is easier said than done); then we should choose coding categories no higher than that level. In fact, there is considerable merit in locating at least some of the categories one level below, on a slightly more detailed level that seems required.

There are at least three reasons why using coding categories that represent a somewhat more molecular level than the level planned for analysis may be a desirable strategy. For example, imagine that our question concerned how often 2-, 3-, and 4-year-old children accompany vocalizations with gestures when directing their attention to others. The question requires only that "bids" for attention be tallied and the presence or absence of vocalizations and of gestures for each bid be recorded. However, we could ask our observers to record whether the bid was directed to an adult or a peer, whether a vocalization involved language or not, whether a nonverbal vocalization was of high or low intensity, whether a verbalization was a question, a command, or a comment, etc. Dropping to a coding level more molecular than required for the question might seem to place additional burdens on

the observers, but paradoxically we think that this strategy increases the chances of collecting reliable data. Just as it is often easier to remember three elements, say, instead of one, if those three are structured in some way, so too observers are often more likely to see and record events accurately when those events are broken up into a number of more specific pieces (as long as that number is not too great, of course). This seems to provide the passing stream of behavior with more "hooks" for the observers to grab.

Further, when data are collected at a somewhat more detailed level than required, we are in a position to justify empirically our later lumping. Given the coding scheme presented in the previous paragraph, for example, a critic might object that the different kinds of vocalization we coded are so different that they should not be dealt with as a single class. Yet if we can show that the frequency of gesture use was not different for the different kinds of vocalizations in the different age groups, then there would be no reason, for these purposes, to use anything other than the lumped category. Moreover, the decision to lump would then be based on something more than our initial hunches.

Finally, and this is the third reason, more detailed data may reveal something of interest to others whose concerns may differ from ours, and at the same time may suggest something unanticipated to us. For example, given the coding scheme described above, we might find out that how gestures and vocalizations are coordinated depends on whether the other person involved is a peer or an adult, even if initially we had not been much interested in, nor had even expected, effects associated with the interactive partner.

We should note that often the issue of level of analysis is not the same issue as whether to code data at a detailed or global level. We may have a set of research questions that call for more than one coding system. For example, Gottman and Levenson are currently employing a socially based coding scheme to describe emotional moments as angry, sad, etc. Observers also note if there was facial movement during each emotional moment. These facial movements are then coded with a detailed, physically based coding system, Ekman and Friesen's Facial Action Coding System (FACS). Gottman and Levenson collected psychophysiological data while married couples interacted. One research question concerns whether there are specific physiological profiles for specific categories of facial expressions. The FACS coding is needed to address this question, but in the Gottman and Levenson study a decision had to be made about sampling moments for FACS coding because detailed FACS coding is so costly. The socially based system is thus used as an aid to employing a more detailed, physically based coding system. Coding schemes at different levels of analysis can thus be used in tandem within the same study.

2.8 Mutually exclusive and exhaustive codes

Almost all the examples of coding schemes presented so far consist of mutually exclusive and exhaustive codes. This means that only one code can be associated with a particular event (mutually exclusive) but that there is some code for every event (exhaustive). For example, with respect to Parten's six social participation codes, only one category was appropriate for each 1-minute time sample, but all time samples could be categorized. Observational studies do not require that all coding schemes consist of mutually exclusive and exhaustive codes, but in fact such schemes have several desirable features – their construction requires a certain amount of conceptual analysis, for example, and their use can simplify data analysis – and as a result such schemes are frequently encountered. Some writers even state that coding schemes must consist of mutually exclusive and exhaustive codes, but there are other possibilities, as discussed in the following paragraphs.

In principle, of course, codes for any behavior can be defined in a way that makes them mutually exclusive and exhaustive (ME&E; see S. Altmann, 1965). For example, if we were interested in the coordination of vocal and visual behavior during face-to-face interaction, we might record (a) when person A was looking at, and (b) when person A was talking to his or her partner. Now these two behavioral codes can cooccur and so are not mutually exclusive, but if we regarded their coocurrence as a new or different code, then we could construct an ME&E coding scheme consisting of four codes: (a) A looks, (b) A talks, (c) A both looks and talks, and (d) the "null" code, A neither looks nor talks.

The coding scheme consisting of the two nonmutually exclusive behaviors may offer certain advantages. For one thing, observers have to remember only two behavioral codes and not three (or four, counting the "null" code). When only two nonmutually exclusive "base behaviors" are under consideration, as in the present example, the difference between their number and the number of ME&E "behavior patterns" they generate is not great, but with only a few more base behaviors the number of possible patterns becomes huge. For example, with four base codes there are 16 patterns, with five base codes, 32 patterns, etc., which could rapidly result in observer overload. Moreover, if the times of onsets and offsets for the base behaviors were recorded, data consisting of ME&E categories could always be generated later, if such data were required for subsequent analyses.

It should be noted that coding time may be increased in some ME&E systems because of the nature of the decisions the coder has to make. This applies to schemes that include codes of the sort (a) event A, (b) event B, (c) both event A and B coocurring. If coding time is an issue, an alternative is to have coders use a checklist of items that can cooccur and need not be exhaustive, or a rating system. This is like having coders fill out a brief

"questionnaire" after every interactive unit occurs. Coders decide about each "item" of the "questionnaire" independently of every other item. Each item can still be precisely defined.

2.9 The evolution of a coding system

Rosenblum (1978) described the initial stages involved in the creation of a coding system. First, he discussed the importance of establishing the conditions of observation. In particular, the situation selected will affect the diversity of behavior displayed, which will, in turn, determine the complexity of the coding system. Second, he suggested beginning with informal observation of behavior. He wrote,

> [I]t is best to begin in the most unstructured fashion possible. There is great advantage to beginning such observations with only a pencil and blank pad for recording, putting aside the spectres of anthropomorphism, adultomorphism, or any of the other rigidifying constraints that must be imposed in separating wheat from chaff later on in the development of the research program; it is vital to begin by using the incredible synthesizing and integrative functions of the human mind. . . . At the beginning, the observer must simply watch, allowing individual subjects to arise as separate entities in the group and behavior patterns to emerge as figures against the background of initially amorphous activity. (pp. 16–17)

We suggest that writing narrative summaries is very helpful at this stage. From the narrative a set of codes is generated, ideally an ME&E set.

As research experience is obtained using a particular coding scheme, it can be modified. For example, Patterson and Moore (1979) discussed using interactive patterns as units of behavior. They analyzed the behavior of Tina using the Family Interaction Coding System (FICS) and found an organization of FICS codes in time. These could later become units of observation. They wrote:

> Examination of Tina's aversive behavior showed that they tended to occur in bursts, followed by periods of neutral or prosocial behavior. . . . It seemed, then, that the bursts, themselves, would be a likely area to search next for structure in social interaction. (p. 83)

Next they noted what FICS behavior initiated and maintained these bursts and identified a set of sequences characteristic of the bursts. Using the larger interaction units, they discovered evidence for what they called an "interactional ripple effect," by which they meant the increased likelihood of the initiating event of the chain occurring once a chain has been run off.

There are many consequences of employing a coding system in a series of studies. New codes may appear, new distinctions may be made, or

distinctions may disappear as a new lumping scheme is derived. Gottman (1979a) used sequential analysis of a 28-code system for coding marital interaction to create eight summary codes. Two codes were lumped together only if they were functionally similar and sequentially related. For example, the "yes but" code functioned in a similar way to simple disagreement in the sense that both led to disagreement by the spouse, so they were lumped together. Gottman argued against the logic of using factor analysis to lump codes. He wrote:

> The problem with factor analysis in this application is that it lumps two codes on the basis of high correlations between these two codes across subjects. However, just because subjects who do a lot of code A also do a lot of code B does not imply that these two codes are functionally equivalent. (p. 91)

A coding system can evolve as it is being used by intelligent coders. To make this systematic, require observers to write a narrative summary of each observation session, keeping a log of any new behavior that occurs and seems important. Then ask observers to note the contexts in which the new behavior occurs, its antecedents and consequences. The new behavior may be part of a functional behavior set already described, or it may require a category of its own.

2.10 Example 1: Interaction of prehatched chickens

In this chapter and the previous one, we have given several examples of coding schemes. For the most part, these have been quite simple, coding just one kind of behavior, like social participation, with just a few mutually exclusive and exhaustive codes. We would like to end this chapter by describing five somewhat more complex coding schemes. The first involves chickens. In order to study interactions between not-yet-hatched chickens and their setting hens, Tuculescu and Griswold (1983) defined 16 codes, organized as follows:

Embryonic behaviors

Distress-type calls

1. Phioo
2. Soft Peep
3. Peep
4. Screech

Pleasure-type calls

5. Twitter

6. Food Call
7. Huddling Call

Maternal behaviors

Body movements

8. Undetermined Move
9. Egg Turn
10. Resettle

Head movements

11. Peck
12. Beak Clap

Vocalizations

13. Cluck
14. Intermediate Call
15. Food Call
16. Mild Alarm Call

To those familiar with the behavior of chickens, these codes appear "natural" and discrete. Trained observers apparently have no trouble discriminating, for example, between a Phioo, a Soft Peep, a Peep, and a Screech, each of which in fact appears somewhat different on a spectrographic recording. Thus "physical reality" may undergird these codes, but human observers are still asked to make the determinations.

These codes are also clearly organized. There are three levels to this particular hierarchy. On the first level, embryonic and maternal behavior are distinguished; on the second, different kinds of embryonic (distress and pleasure calls) and different kinds of maternal (body movements, head movements, vocalizations) behavior are differentiated; and on the third, the codes themselves are defined. Within each "second-level" category, codes are mutually exclusive, but codes across different second-level categories can cooccur. Indeed, cooccurrence of certain kinds of behavior, like embryonic distress calls and maternal body movements, was very much of interest to the investigators.

There are at least three reasons why we think organizing codes in this hierarchical fashion is often desirable. First, it both ensures and reflects a certain amount of conceptual analysis. Second, it makes the codes easier to explain to others and easier for observers to memorize. Third, it facilitates analysis. For example, for some analyses all embryonic distress calls and all maternal vocalizations were lumped together, which is an example of a practice we recommended earlier – analyzing on a more molar level than that used for data collection.

2.11 Example 2: Children's conversations

The second example is derived from a study conducted by John Gottman (1983) on how children become friends. Working from audiotapes, observers categorized each successive "thought unit" in the stream of talk according to a catalog of 42 mutually exclusive and exhaustive content codes. These 42 codes were grouped into seven superordinate categories: (a) demands for the other child, (b) demands for the pair, (c) you and me statements, (d) self-focus statements, (e) emotive statements, (f) social rules, and (g) information exchange and message clarification statements. Here we reproduce definitions and examples just for the first 16 codes (from Gottman, 1983, p. 13):

Demands for the other child

1. Command (Gimme that.)
2. Polite Requests (That one, please.)
3. Polite Request in Question Form (Would you gimme that?)
4. Suggestion (You could make that black.)
5. Suggestion in Question Form (Why don't you make that black?)
6. Asking Permission (Can I play with that now?)
7. Demands in the Form of an Information Statement (I think my crayons are next to you.)
8. Demands in the Form of a Question for Information (Have you got any sixes?)
9. Wanna (I wanna play house.)
10. Question Wanna (Do you wanna play house?)
11. Requirements for the Other Child (You should stay in the lines.)
12. Asks Help (Would you tie this for me?)

We demands (demands for the pair)

13. Hafta Wanna (We have to take a nap.)
14. Let's (Let's play house.)
15. Let's in Question Form (How about drawing now?)
16. Roles to Both (You be the cop and I'll be the robber.)

Unlike with the codes for chicken behavior described above, it is hard to claim that any physical reality undergrids Gottman's content codes. For that very reason, he took considerable pains to demonstrate observer reliability for his codes – which Tuculescu and Griswold did not. Gottman also made finer distinctions when defining his codes than he found useful for later analyses – which is natural tendency when the cleavage between codes is not all that clear. Still, like Tuculescu and Griswold if for slightly different reasons, he found it useful to lump codes for analysis. In fact, the initial 42 content codes were reduced to 20 for his analyses of friendship formation. The three codes derived from the 16 initial codes list above were (a) Weak demands – numbers 2, 3, 5, 6, 7, 8, and 10 above; (b) Strong Demands –

numbers 1, 4, 9, 11, and 12 above; and (c) Demands for the Pair – numbers 13, 14, 15, and 16 above.

Using the result of sequential analysis from a detailed coding system, Gottman devised a "macro" coding system. The macro system was designed so that it would be faster to use (2 hours per hour of tape instead of 30) and would code for the sequences previously identified as important. In the process of building the macro system, new codes were added because in moving to a larger interaction unit, he noticed new phenomena that had never been noticed before. For example, the codes escalation and deescalation of a common-ground activity were created. Gottman (1983) wrote:

> Escalation and deescalation of common-ground activity were included as categories because it appeared that the children often initially established a relatively simple common-ground activity (such as coloring side by side) that made low demands of each child for social responsiveness. For example, in coloring side by side, each child would narrate his or her own activity (e.g., "I'm coloring mine green"). This involved extensive use of the ME codes. Piaget (1930) described this as collective monologue, though such conversation is clearly an acceptable form of dialogue. However, in the present investigation the common-ground activity was usually escalated after a while. This anecdotal observation is consistent with Bakeman and Brownlee's (1980) recent report that parallel play among preschool children is usually the temporal precursor of group play. However, the extent of this process of escalation was far greater than Bakeman and Brownlee (1980) imagined. An example of this escalation is the following: Both children begin narrating their own activity; then one child may introduce INX codes (narration of the other child's activity – e.g., "You're coloring in the lines"); next, a child may begin giving suggestions or other commands to the other child (e.g., "Use blue. That'd be nice"). The activity escalates in each case in terms of the responsiveness demand it places on the children. A joint activity is then suggested and the complexity of this activity will be escalated from time to time.
>
> This escalation process was sometimes smooth, but sometimes it introduced conflict. When it did introduce conflict, the children often deescalated that activity, either returning to a previous activity that they had been able to maintain or moving to information exchange. While many investigators have called attention to individual differences in the complexity of children's dialogue during play (e.g., Garvey, 1974; Garvey & Berndt, 1977), the anecdotal observation here is that a dyad will escalate the complexity of the play (with complexity defined in terms of the responsiveness demand) and manage this complexity as the play proceeds. I had not noticed this complex process until I designed this coding system. However, I do not mean to suggest that these processes are subtle or hard to notice, but only that they have until now been overlooked. An example will help clarify this point. D, the host, is 4-0; and J, the guest, is 4-2. They begin playing in parallel, but note that their dialogue is connected.
>
> 18. J: I got a fruit cutter plate.
> 19. D: Mine's white.

20. J: You got white Play-Doh and this color and that color.
21. D: Every color. That's the colors we got.

They continue playing, escalating the responsiveness demand by using strong forms of demands.

29. D: I'm putting pink in the blue.
30. J: Mix pink.
31. D: Pass the blue.
32. J: I think I'll pass the blue.

They next move toward doing the same thing together (common-ground activity).

35. D: And you make those for after we get it together, OK?
36. J: 'Kay.
37. D: Have to make these.
38. J: Pretend like those little roll cookies, too, OK?
39. D: And make, um, make a, um, pancake, too.
40. J: Oh rats. This is a little pancake.
41. D: OK. Make, make me, um, make 2 flat cookies. Cause I'm, I'm cutting any, I'm cutting this. My snake.

The next escalation includes offers.

54. J: You want all my blue?
55. D: Yes. To make cookies. Just to make cookies, but we can't mess the cookies all up.
56. J: Nope.

They then introduce a joint activity and begin using "we" terms in describing what the activity is:

57. D: Put this the right way, OK? *We're* making supper, huh?
58. J: *We're* making supper. Maybe *we* could use, if you get white, *we* could use that too, maybe.
59. D: I don't have any white. Yes, *we*, yes I do.
60. J: If you got some white, *we* could have some, y'know.

As they continue the play, they employ occasional contextual reminders that this is a joint activity:

72. D: Oh, we've got to have our dinner. Trying to make some.

D then tries to escalate the play by introducing some fantasy. This escalation is not successful. J is first allocated a low-status role (baby), then a higher-status role (sister), then a higher-status (but still not an equal-status) role (big sister).

76. D: I'm the mommy.
77. J: Who am I?
78. D: Um, the baby.
79. J: Daddy.
80. D: Sister.
81. J: I wanna be the daddy.
82. D: You're the sister.

83. J: Daddy.
84. D: You're the *big* sister!
85. J: Don't play house. I don't want to play house.

The escalation failure leads to a deescalation.

87. J: Just play eat-eat. We can play eat-eat. We have to play that way.

However, in this case, the successful deescalation was not accomplished without some conflict:

89. J: Look hungry!
90. D: Huh?
91. J: I said look hungry!
92. D: Look hungry? This is dumb.
93. J: Look hungry!
94. D: No!

The children then successfully returned to the previous level of common ground activity, preparing a meal together. Common ground activity is thus viewed in this coding system as a hierarchy in terms of the responsiveness it demands of each child and in terms of the fun it promises. (pp. 55–57)

2.12 Example 3: Baby behavior codes

Our third example describes the behavior of young infants during times when they are being administered a neonatal assessment and was developed by Sharon Landesman-Dwyer (1975). The scheme contains 50 codes grouped into five superordinate categories. The five categories are not mutually exclusive. In fact, they code different possible kinds of behavior, namely, (a) the kind of external stimulation being provided and what the infant is doing with his or her (b) eyes, (c) face (including vocalizations), (d) head, and (e) body. The ten codes within each of these superordinate categories, however, are mutually exclusive and exhaustive. For example, the 10 codes for eyes are:

0. Can't See
1. Closed
2. Slow Roll
3. Squint
4. Open–Close
5. Daze
6. Bright
7. Tracking
8. REM
9. Other

and the 10 codes for face and vocalization are:

0. Repose
1. Small Move
2. Mouth/Suck
3. Grimace
4. Smile
5. Tremor
6. Yawn
7. Whimper
8. Cry
9. Cry/Tremor

2.13 Example 4: Children's object struggles

The fourth example is taken from a study of the social rules governing object conflicts in toddlers and preschoolers, conducted by Bakeman and Brownlee (1982). They defined six codes, organized into three superordinate categories. Each superordinate category can be cast as a question about an object struggle as follows:

Prior possession: Did the child now attempting to take an object from another child play with the contested object at any point in the previous minute?

1. Yes
2. No

Resistance: Does the child now holding the object resist the attempted take?

1. Yes
2. No

Success: Is the child who is attempting to take the object successful in getting it?

1. Yes
2. No

The four schemes described above are alike in at least one sense. In each case, the codes defined can be grouped into a relatively small number of superordinate categories. They are also alike in that all codes within a superordinate category are at least mutually exclusive and, with the possible exception of Tuculescu and Griswold, exhaustive as well. Gottman's superordinate categories are themselves mutually exclusive, however, whereas that is not the case for the other three schemes. In fact, the superordinate

categories for the other three schemes represent different modes or kinds of behavior or different questions about a particular behavioral event and are clearly not mutually exclusive (with the exception of embryonic distress and pleasure calls).

The Landesman-Dwyer and the Bakeman and Brownlee schemes are formally identical. Both consist of several sets of mutually exclusive and exhaustive codes. This is a useful structure for codes because it ensures that answers to a number of different questions will be answered: What is the baby doing with his eyes? With his mouth? Did the taker have prior possession? They differ, however, in when the questions are asked. Landesman-Dwyer's scheme is used to characterize each successive moment in time, whereas Bakeman and Brownlee's scheme is used to characterize a particular event and is "activated" only when the event of interest occurs.

2.14 Example 5: Monkeys' activities

The fifth example is from the work of Suomi and his coworkers concerning the social development of rhesus monkeys (Suomi, Mineka, & DeLizio, 1983). They defined 14 codes for monkeys' activities that, although not necessarily mutually exclusive, were designed to be exhaustive. Observers were asked to record frequency and duration information for each. The 14 codes (along with brief definitions) were as follows (paraphrased from Suomi et al., 1983, p. 774):

1. Mother–Infant Ventral (mutual ventral and/or nipple contact)
2. Mother–Infant Reject (any successful or unsuccessful break of Mother–Infant Ventral, or rejection of an infant-initiated Mother–Infant Ventral)
3. Ventral Cling (ventral contact with an animal other than the mother)
4. Self-Clasp (clasping any part of own body)
5. Self-Mouth (oral contact with own body)
6. Environmental Exploration (tactual or oral manipulation of inanimate objects)
7. Passive (absence of all social, exploratory, and locomotor activity; could cooccur with self categories and vocalizations)
8. Stereotypy (patterned movements maintained in a rhythmic and repetitive fashion)
9. Locomotion (any self-induced change in location of self, exclusive of stereotypy)
10. Vocalization (any sound emitted by subject)
11. Huddle (a maintained, self-enclosed, fetal-like position)

12. Play and Sex (any type of play and/or sexual posturing, exclusive of locomotion)
13. Aggression (vigorous and/or prolonged biting, hair pulling, clasping, accompanied by one or more of threat, barking, piloerection, or strutting)
14. Social Contact (contact and/or proximity with another subject, exclusive of Mother–Infant Ventral, Ventral Cling, Aggression, or Play and Sex)

Except for Passive (which can cooccur with Self-Clasp, Self-Mouth, and Vocalization), these codes appear to be mutually exclusive. In some cases, activities that could cooccur have been made mutually exclusive by definition. For example, if an activity involves both Stereotypy and Locomotion, then Stereotypy is coded. Similarly, if what appears to be Social Contact involves a more specific activity for which a code is defined (like Play and Sex), then the specific code takes precedence. Defining such rules of precedence is, in fact, a common way to make a set of codes mutually exclusive.

2.15 Summary

No other single element is as important to the success of an observational study as the coding scheme. Yet developing an appropriate scheme (or schemes) is often an arduous task. It should be assumed that it will involve a fair number of hours of informal observation (either "live" or using videotapes), a fair amount of occasionally heated discussion, and several successively refined versions of the coding scheme. Throughout this process, participants should continually ask themselves, exactly what questions do we want to answer, and how will this way of coding behavior help us answer those questions?

There is no reason to expect this process to be easy. After all, quite apart from our current research traditions, developing "coding schemes" (making distinctions, categorizing, developing taxonomies) is an ancient, perhaps even fundamental, intellectual activity. It seems reasonable to view one product of this activity, the coding scheme, as an informal hypothesis, and the entire study in which the coding scheme is embedded as a "test" of that hypothesis. If the "hypothesis" has merit, if it is clearly focused and makes proper distinctions, then sensible and interpretable results should emerge. When results seem confused and inconclusive, however, this state of affairs should not automatically be blamed on a lack of proper data-analytic tools for observational data. First we should ask, are questions clearly stated, and do the coding schemes fit the questions? We hope that a consideration

of the various issues raised in this chapter will make affirmative answers to these questions more likely.

In this chapter, we have confined our discussion to coding schemes. Five examples of developed schemes were presented, and an additional four are detailed in the companion volume (Bakeman and Quera, 1995a, chapter 2). We have stressed in particular how coding schemes can be organized or structured and have left for the next chapter a discussion of how coding schemes are put into use. This separation is somewhat artificial. How behavioral sequences are to be recorded can and often does affect how codes are defined and organized in the first place. This is especially evident when the Landesman-Dwyer and the Bakeman and Brownlee schemes discussed earlier are compared. Still, a scheme like Landesman-Dwyer's could be recorded in two quite different ways, as we discuss in the next chapter. It is the task of that chapter to describe the different ways behavior can be recorded, once behavioral codes have been defined.