15 Discovering the Relationship Between Operational Definitions and Interobserver Reliability

Angela H. Becker
Indiana University Kokomo

This activity is designed for students in research methods and behavior modification classes or the methods section of other content courses. Students observe a brief videotape and collect data on the occurrence/nonoccurrence of a series of six behaviors. The main purposes of this activity are to help students (a) understand the importance of having clear operational definitions, (b) learn to calculate interobserver reliability, and (c) think about ways to improve a study that has low interobserver reliability. In addition, students gain practice in using time sampling and come to realize that observation as a data collection technique is more complex than casual observation.

Concept

This exercise helps students realize that observation as a data collection technique is more complex than casual observation. It introduces students to the use of time sampling, the calculation of interobserver reliability, and the importance of having clear operational definitions.

Materials Needed

You will need a 10-min videotape of human or animal behavior, a VCR and monitor for showing the tape to the class, a watch with a second hand, and enough copies of the handouts described later for each student in the class. Students will need pencils and will probably want calculators.

To be most effective, the videotape should be of a group of humans or animals that are active enough to produce several different types of behaviors. If at least some of those behaviors occur quite frequently and in several individuals at a time, students will come away with a better understanding of why time sampling is useful. My tape is of a group of white geese at a local park. (I would be happy to provide a copy of this taped segment to anyone who sends me a blank videotape.) There are many other possibilities for footage that will meet the previously described requirements. For example, you could videotape small children performing at a school program or playing at a birthday party, or you could get some footage of one of the livelier species at your local zoo. If you or someone you know is planning to visit another country, you may be able to obtain a tape of a festival or other group event from another culture. The videos that some universities make of their graduation ceremonies could also be used. If you have access to a VCR, you could tape a segment of an appropriate televised event. Although many television programs show groups engaging in behaviors that meet the criteria identified at the beginning of this discussion, most do not show this activity, uninterrupted by close-ups of individuals or pans to scenery and other locations

in the story line, for more than a minute or two. There are exceptions, however, that would make good tapes for this activity: Televised New Year's Eve bashes usually show quite long segments of partyers, sporting events such as basketball or volleyball also show fairly long segments of activity on the court, and dance club shows on cable channels run segments of couples dancing for the duration of a complete song.

Prepare three handouts. The first should contain a data collection sheet with a row for each observation interval and a column for each behavior students are to record (see appendix A). The second should contain a list of behaviors (I recommend no more than six to eight), their operational definitions (some of which are purposely clearer than others), and a simple formula for inter-rater reliability (see appendix B). The third handout should contain a set of postobservation questions (see appendix C).

INSTRUCTIONS

This activity should be prefaced with a lecture on the use of observation techniques, including the advantages and disadvantages of time sampling in relation to other observation techniques (e.g., event sampling and narrative recording) and the concept of interobserver reliability. Although the basic observation techniques described in methods textbooks are much the same, the labels given to particular techniques vary. The following definitions of observation techniques are provided to facilitate gathering background lecture material and to prevent misunderstanding.

Time sampling is a technique in which the observer defines several target behaviors, divides the observation period into short intervals, and then alternates from observing to recording every other interval. In contrast, in *event sampling* the observer defines a target behavior and records every instance of that behavior as it occurs throughout the observation period. A *narrative recording* is a running description of behavior in which everything that is said or done during the observation period is recorded. The following are particularly useful sources for lecture material: chapter 6 from Bordens and Abbott (1996); chapters 6, 7, and 8 from Martin and Bateson (1993); and chapter 19 from Martin and Pear (1996).

Give each student all three handouts, and allow them a few minutes to read through the list of behaviors and operational definitions and become familiar with the layout of the data collection sheet. Have students work in 15-s intervals—alternating between 15 s for observing and 15 s for recording observations completed in the previous 15-s interval. For each observation interval, they simply look for whether or not each target behavior occurs. If the behavior occurs *at least once* in the observation interval, they are to place a tally mark in the appropriate column on the data sheet during the recording interval. Rather than having students keep track of their own intervals, use a watch with a second hand to time intervals for them. It works best if you simply call out "observe" or "record" at the beginning of alternate 15-s intervals. Explain to students that data collection will last for a total of 10 min. Each minute represents an observation period and is divided into 30-s sessions. During each 30-s session, students will have a 15-s interval to observe the behavior on the videotape and a 15-s interval to record their observations. It is important for students to understand what they are observing and recording; be sure to explain that they are recording the occurrence of target behaviors. That is, they are looking for whether or not a behavior occurs; they are *not* looking for the number of times a specific behavior occurs.

After 10 min, have students stop observing and work on the postobservation questions (see appendix C). As appendix C illustrates, students will first answer several questions individually, then compare those answers with a partner, and finally calculate interobserver reliability with their partner for each of the target behaviors.

Discussion

Follow up with a class discussion of students' responses to the postobservation questions. Focus on those questions where partners' responses differed most often and on those behaviors that had the highest and lowest interobserver reliability. Then discuss possible reasons for these trends. With my tape of geese and set of behaviors, for example, students generally have very high interobserver reliability for displays and for tail shakes and low reliability for feeding and submission. When exploring possible reasons for these findings, student comments tend to focus on the importance of careful operational definitions and on problems with observation. For example, students notice that my operational definition of display behavior is much more concrete than my definition of submission, that the definition for feeding was too narrow to encompass much of what they wanted to be able to code as feeding behavior, and that several of the behaviors were difficult to identify accurately because of the distance from which the videotape was shot. For example, one student declared that she wanted a better definition of feeding, because "sometimes I thought they might be, but I couldn't see if they really had food in their mouths or not."

Next ask students to offer possible solutions for the reliability problems they have encountered. We talk about clarifying operational definitions. For example, several students decided that "touching beak to the ground several times in a row" would have defined feeding in a way that would have allowed them to record what they thought was feeding behavior. Students also brought up the possibility of reviewing the tape and discussing discrepancies between observers in order to resolve disagreements or practicing with sample tapes to improve reliability before viewing actual data tapes. This second idea was an elaboration on one student's comment that he "wished we could have watched the whole tape first while reading the definitions and *then* done the recording part." The students also decided that the value of a high-power zoom lens should not be underestimated if one wants to observe detailed behaviors and remain unobtrusive. Overall, students' responses to the activity indicated that not only did they learn a great deal, but that they enjoyed the activity as well.

A minor variation on this activity could allow students to discover for themselves that one of the pitfalls to time sampling is that there will always be lost data (i.e., behaviors that occur during recording intervals rather than observation intervals). Instead of having the entire class observe and record during the same intervals, divide the class in half, and have each half observe and record during opposite intervals. Have each half of the class pool their data and calculate the mean number of intervals in which tail shakes, feeding, grooming, display, aggression, and submission were observed by their group. Students should find that for behaviors that occur frequently, there will be little difference between the means reported by each half of the class. For example, the two halves of the class should be quite similar on mean number of tail shakes, simply because this occurs almost continuously among geese. However, for relatively infrequent behaviors, such as displays of aggression, students are likely to notice differences between

reports by the two halves of the class. This can lead to a discussion of the relative usefulness of time sampling versus event sampling for observing infrequent behaviors. (Obviously, if you use this alternative procedure and you still want students to calculate interobserver reliability, they must do so by pairing up with someone from the same half of the class.)

Writing Component

Instructors who want to provide their students with an opportunity to do more writing than the small amount required to complete appendix C may add one of the following writing components to the activity.

1. Have students reflect on their expectations of observation in general and time sampling in particular. After giving students a brief description of the exercise they are about to engage in, ask them to respond to the question, "What do you think will happen when we do this time sampling observation?" After the exercise is complete, have students reread their earlier expectations and write a response to the following two questions: (a) "Which of your earlier expectations were met and why do you think this happened?" and (b) "Which of your earlier expectations were *not* met and why do you think this happened?" As a follow-up, students could construct a list on the chalkboard of the group's most common expectations, identify those that were not met, and then discuss whether those unmet expectations would make them more or less likely to want to use this method in their own future research.
2. Have students write a report to the researchers who set up the study. In that report, students should point out the strengths and weaknesses of the study and suggest improvements. This writing component could be followed by a small-group discussion in which students compare the strengths and weaknesses they noticed and try to identify the most methodologically sound and practical suggestions for improvement. You could also ask the students to use this small-group time to rewrite the operational definitions that they found lacking.
3. Instructors who have their students keep journals might consider having them include an entry about this observation activity. Students could be asked to respond to the question "What do you feel you learned from this observation exercise?" If content analysis is covered in your course, you could have students use these journal responses as data and attempt to code them into categories.

References

Bordens, K. S., & Abbott, B. B. (1996). *Research design and methods: A process approach* (3rd ed.). Mountain View, CA: Mayfield.

Martin, G., & Pear, J. (1996). *Behavior modification: What it is and how to do it* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Martin, P., & Bateson, P. (1993). *Measuring behavior: An introductory guide* (2nd ed.). Cambridge: Cambridge University Press.

Appendix A

Data Collection Sheet for Time Sampling

Minute	Tail Shake	Feeding	Grooming	Display	Aggression	Submission
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

Each cell represents a 15-s observation interval. Recording intervals are not shown on this sheet.

Appendix B

Target Behaviors and Operational Definitions

Tail shake	Flicking tail back and forth rapidly several times in succession
Feeding	Actually taking food in beak
Grooming	Preening—using beak to fluff or pick at feathers
Display	Full (or almost full) extension of wings accompanied by several flaps, a slight lift in body posture, and a slight extension of neck—usually done when standing or when walking very slowly
Aggression	Nipping or threatening (by chasing or quickly swinging head toward another individual)
Submission	Running from or obviously avoiding close contact with another individual

****

Formula for Calculating Interobserver Reliability When Doing Time Sampling

$$\text{reliability} = \frac{\text{agreements}}{\text{agreements} + \text{disagreements}} \times 100$$

(Agreements = number of intervals in which you both marked that the behavior occurred, and disagreements = number of intervals in which only one of you marked that the behavior occurred.)

****

Appendix C

Postobservation Instructions

1. Individually, tally the number of intervals in which each behavior occurred. Then, answer the following questions:
   - Which behavior occurred *most* often?
   - Which behavior occurred *least* often?
   - Are there any behaviors that at least *appear* to be highly correlated? (That is, are there any behaviors that seem to always, or almost always, occur during the same intervals?)
2. Pair up with another student and compare your answers to the preceding questions. Did you disagree on any of them? If so, which one(s)? *Why* did you disagree?
3. Calculate the interobserver reliability between you and your partner for each behavior category. Identify the category that has the highest interobserver reliability and the category that has the lowest interobserver reliability.

Tail shake =	Display =
Feeding =	Aggression =
Grooming =	Submission =