Observational Methods

Description

This assignment is divided into three major sections. The first part discusses the unit of behavior used for measurement. The sec section describes the methods for recording behavior. The last present three measures of reliability when recording behavior.

UNIT OF BEHAVIOR

Behavioral assessment is done in clinics, schools, prisons, homes, and work settings. Before measuring behavior, the observer must identify the specific behaviors of interest called target behaviors.

Behavior is so important to our everyday lives that we have developed elaborate classification schemes to think about behavior more effectively. For example, the behavior of "jogging" could be classified into a more molar (broader) category of "health oriented behavior." Health oriented behavior could be classified into a more molar category of "mature behavior." Mature behavior could be classified into an even more molar category of "well-adjusted behavior," etc.

When doing behavioral assessment, a broader category of behavior is called a more "**molar**" category whereas a more specific category is called more "**molecular**." Tying your shoes can be classified into the behavioral category of getting dressed. Since tying shoes is more specific, it is a more molecular category than getting dressed (i.e., getting dressed is more molar).

When doing behavioral assessment, **molar behavior is defined very molecularly** before actually measuring the individual's behavior. This **minimizes the interpretations** made by observers, thereby incasing the objectivity and the reliability of the recorded observations.

Consider this example. You are a school psychologist told by a teacher that Tom's nervousness while taking tests may be detracting from his performance (he is averaging a D grade in every subject). You decide to work out a plan to objectively measure his nervousness.

Can you directly observe Tom's nervousness? Of course not since someone else's "nervousness" is inferential. Nervousness is too molar to be useful at an observational level. Instead, you ask his teacher what Tom is doing that indicates he is nervous (i.e., you ask her to molecularize his nervous behavior). She replies that he whimpers, shakes, chatters his teeth, and occasionally wets his pants while taking a test. This level is molecular enough to make behavioral recordings.

When assessing behavior, behavior must be defined so molecularly that it is at the descriptive level. Such behaviors do not include a person's ability (e.g., she can wash her face), qualities (e.g., she is wealthy), roles (e.g., she is a teacher), inferred knowledge (e.g., she knows where she is going), emotional states (e.g., she feels bad, or lack of behavior (e.g., opening a checking account).

BEHAVIORAL DIMENSIONS

There are three basic kinds of measurable behavioral dimensions: **Frequency**, **Duration**, and **Amplitude**. Frequency is the number of times a target behavior occurs per unit time. Duration is the length of time a target behavior occurs during a specified interval. Amplitude is the intensity of the target behavior.

For example, say you molecularized a temper tantrum as "lying on the floor kicking and crying." You could measure the number of times a child lies down on the floor kicking and crying each day (frequency). Another approach would be to measure the amount of time a child spends on the floor kicking and crying each day (duration). A third approach would be to use special sound equipment to measure how loud the child cries while lying on the floor kicking and crying each day (amplitude).

The **frequency dimension is most popular in behavioral assessment**. It can be conveniently used to measure both brief (e.g., burping) and prolonged (e.g., talking) forms of behavior. The duration dimension is limited to prolonged forms of behavior since the length of brief behaviors is difficult to measure accurately.

Although amplitude can often be used to measure certain behavior (e.g., loudness of crying), it is generally difficult to use when measuring ongoing behavior.

While this exercise emphasizes the frequency dimension, the methods described in the following sections could also be applied to the other two dimensions.

Products of behavior

A product of behavior is sometimes used as an indirect measure of behavior. For example, if you want to assess the effectiveness of an anti-littering campaign, it would be more practical to measure the amount of litter in a given area before, during, and after the campaign than to try to count the number of times people are observed littering. Not that the behavioral dimension used in this example is amplitude.

Although products of behavior are important measures, most of this exercise focuses on direct behavioral measures. Products of behavior will be briefly discussed in the reliability section.

OBSERVATION METHODS

There are numerous ways to classify the many procedures used for recording behavioral observations. Three ways to classify them are on the basis of:

- the **number of different target behaviors** observed (**single** or **multiple**)
- the **number of persons** being observed (**individual** or **group**)
- the **behavior sampling method** (**Event Sampling** is recording every occurrence of the target behavior whereas **Time Sampling** is recording target behavior only if it occurs at predetermined points in time)

Number of Target Behaviors

A psychologist is occasionally interested in a single target behavior (e.g., smoking cigarettes). In other situations, we want to measure multiple behaviors. For example, a child's aggressive behavior could be molecularized as hitting other people, throwing rocks at other people, and spitting towards another person. In this case, three separate target behaviors would be measured rather than only one.

When observing a child's behavior, a data sheet is used listing each of the three behaviors on separate lines. The observer places a **check mark** in an appropriate box (cell) every time the target behavior occurs during the observation period (Figure 6.2).

When the observation period has ended, the frequency of each target behavior can be kept separate (e.g., 3 hits, zero throws, 5 spits), and/or could be combined into a single frequency classified as "aggressive behavior" (e.g., 3 + 0 + 5 = 8).

Number of People Observed

The observer may be interested in the behavior of only one individual or that of an entire group. The data recording sheet in Figure 6.2 is that for a single individual.

Say an observer wants to measure the aggressive behavior among a group of first graders. The observer would first prepare a data sheet listing the molecularized behaviors (e.g., hitting, throwing rocks, etc.) on the lines going down the left side of the sheet (Figure 6.1). The children's names would then be the column headings across the top of the sheet.

Each time a specific child is seen doing a specific behavior, the observer places a checkmark in the position corresponding to that child (column) and that behavior (row). The checkmarks in Figure 6.1 show that Mary hit someone twice, threw rocks once, and spit once during the one hour observation period.

Observation Period: ___1 hour___
Date: ___Oct 1___

	Mary	Jim	Louise	Rod	Total
Hitting	✓✓	✓	✓✓✓	✓	7
Throw Rocks	✓	✓		✓✓✓✓	6
Spitting	✓	✓	✓	✓✓✓✓	7

Grand Total 20

Figure 6.1 – Response sheet for measuring multiple behaviors among a group of children

After each observation period, the frequencies could be combined across children to get a group frequency for each target behavior. These group totals could also be combined to obtain the group frequency for aggressive behavior. In Figure 6.1, the frequency of aggressive behavior for the group is 20.

It is important to keep each child's frequencies separate when recording group behavior in case the information about specific children is needed at a later time.

Behavior Sampling Method

A person's behavior during an observation period is considered a "sample" of the behavior they display in their everyday life. This sample is used to infer how that person behaves in similar situations on other occasions.

A sample of a person's target behavior may also be compared before, during, and after some form of treatment to assess treatment effectiveness. A variety of methods are used to sample behavior. Two major methods are "event sampling" and "time sampling." Both use the behavioral frequency dimension so don't let the "time" in time sampling confuse you.

Even Sampling

When using event sampling, every occurrence of the target behavior(s) is recorded during the specified intervals. The examples used in the previous section both used event sampling.

When event sampling is used, it is often important to know when in the observation period each occurrence happened. Thus, the observation period (e.g., one hour) is divided into smaller intervals (e.g., 10 minute segments). When preparing a data sheet to record the observations, the target behaviors are listed on the left side of the sheet and the time segments on top (Figure 6.2).

Observation Period: ___60 minutes___
Client: ___John Smith___
Date: ___Oct 1___

TIME SEGMENTS (MINUTES)

	0-10	10-20	20-30	30-40	40-50	50-60	Total
Hitting		✓		✓	✓		3
Throw Rocks							0
Spitting	✓✓	✓		✓	✓		5

Grand Total 8

Figure 6.2 – Data recording sheet for multiple target behaviors of a single individual using an event sampling procedure. Each checkmark represents the occurrence of target behavior in that interval.

When observing the person's behavior during the first 10 minutes, a checkmark is recorded in the first column each time a corresponding target behavior occurs. The next column is used to record every occurrence of the target behaviors during the second 10 minute interval, etc.

At the end of the observation period, the frequencies are combined across each target behavior and listed in the total column. The totals are summed and listed next to the grand total.

Time Sampling

When using time sampling, specific points in time (e.g., 30 second markers) are identified during an observation period (e.g., 3 minutes). The target behaviors occurring at each time point are recorded. Target behaviors that occur before or after each point are ignored.

When preparing a data sheet for time sampling, the target behaviors are listed on the left side and the time points are listed over the columns (Figure 6.3).

Observation Period: ___3 minutes___
Observer: ___Joanne Smith___
Client: ___John Smith___
Date: ___Oct 1___

TIME SEGMENTS (SECONDS)

	30	60	90	120	150	180	Total
Hitting		✓		✓	✓		3
Throw Rocks							0
Spitting	✓	✓			✓		3

Grand Total 6

Figure 6.2 – Data recording sheet for multiple target behaviors of a single individual using a time sampling procedure. Each checkmark represents the occurrence of target behavior at that point in time.

These two behavior sampling procedures have certain advantages and disadvantages. Event sampling is more sensitive than time sampling for recording brief target behaviors (e.g., sneezing, asking a question) because every occurrence is recorded.

Time sampling is more sensitive than event sampling when recording prolonged behavior. If the target behavior was "talking" and a person talked constantly during a one hour observation period, the event sampling would show a frequency of "1" for talking behavior for each 10 minute interval, yielding a frequency of 6. The time sampling procedure, if one minute time points were used, would sow a frequency of 60.

Time sampling is also an advantage when recording a large number of different target behaviors. When using an event sampling procedure, it is easy to miss the occurrence of some target behaviors while recording target behaviors that have just occurred. Time sampling provides sufficient time to record behavior between successive time points.

RELIABLITY

Reliability means "**consistency**." Reliability of behavioral observations is measured with two observers recording the **target behavior** of the **same** person(s) at the **same** time.

Reliability is expressed as the degree the two sets of observations agree. If there is perfect agreement, the reliability of the observations is perfect. If there is much disagreement, the reliability is poor.

This section presents three methods for expressing the reliability of behavioral observations made by two independent observers. They are called **Observation Reliability**, **Occurrence Reliability**, and **Outcome Reliability**. Each will be discussed separately after introducing the following example.

Say you are using an event sampling procedure to observe multiple behaviors of a single individual. Two observers are first trained to classify the target behaviors, using a variety of different examples. This is important so both willuse the same behavioral definitions when making observations.

After training, the reliability is measured. It is necessary for both observers to record the same behaviors of the same individual at the same time. Care is taken so the two observers can't see each other's recordings. The data from the two observers are shown in Figure 6.4.

Observation Period: ___5 minutes___
Observer: ___Jill Smith___
Client: ___John Smith___
Date: ___Oct 1___

TIME SEGMENTS (SECONDS)

	1-30	30-60	60-90	90-120	120-150	150-180	180-210	210-240	240-270	270-300	Total
Smile	✓✓		✓	✓	✓✓✓		✓✓				9
Hug				✓✓							2
Kiss											0

Grand Total 11

Observation Period: ___5 minutes___
Observer: ___Tom Smith___
Client: ___John Smith___
Date: ___Oct 1___

TIME SEGMENTS (SECONDS)

	1-30	30-60	60-90	90-120	120-150	150-180	180-210	210-240	240-270	270-300	Total
Smile	✓✓		✓	✓		✓✓✓	✓	✓			9
Hug					✓						1
Kiss											0

Grand Total 11

Figure 6.4 – Data from two observers recording multiple target behaviors of a single individual using an event-sampling procedure.

Observation Reliability

The observation reliability for a given target behavior is determined by first counting the pairs of cells in which the two observers agreed. The observation reliability coefficient is then calculated by comparing the number of agreeing cell pairs to the total number of cell pairs.

The formula is:

$$\text{Observation Reliability Coeff.} = \frac{\text{\# of Agreeing Cell Pairs}}{\text{Total \# of Cell Pairs}}$$

For smiling behavior in Figure 6.4, the data for the two observers agreed for intervals 1-30, 30-60, 90-120, 150-180, 240-270, and 270-300. Thus, there were six cell pairs which agreed out of the total of 10 cell-pairs. **The Observation Reliability Coefficient is:**

$$\text{Observation Reliability Coeff.} = \frac{\text{\# of Agreeing Cell Pairs}}{\text{Total \# of Cell Pairs}}$$

Observation Reliability Coeff. = 6/10 = .60 = 60%

This means that for 60% of the intervals, the two observers agreed on the number of times the person smiled. Using this same approach for hugging behavior, the observation reliability coefficient is 9/10 or 90%. The observation reliability coefficient for kissing behavior is 100%.

A "combined observation reliability coefficient" can also be obtained by comparing the cell-pairs for all behaviors. Since there are 30 pairs of cells (10 for each target behavior) and there were 6 agreements for smiling 9 agreements for hugging, and 10 for kissing, the combined observation reliability coefficient would be (6 + 9 + 10)/30 = 25/30 = .83 = 83%.

When actually conducting a reliability test, more than 10 time segments are typically used. The observers also record the behavior of a number of different individuals since the difficulty of judging target behaviors vary between different people.

It is also important that one of the observers who participated in the reliability study does the observing in the actual study. Just because you have shown that two observers can reliably judge a certain target behavior does not mean that a new observer's ratings are also reliable.

Occurrence Reliability

You may have already noticed a limitation of observation reliability in the preceding example. Figure 6.4 shows perfect observation reliability for kissing data since both observers agreed that kissing had not occurred during the 10 segments. Just because two observers agree when a behavior does not occur does not mean they would agree if it were to occur.

Occurrence reliability overcomes this limitation. It is similar to observation reliability except it is based on only those cells in which at least one observer reported an occurrence. It ignores all cells in which both observers agree that the target behavior didn't occur.

The formula for Occurrence Reliability Coefficient is:

$$\text{Occurrence Rel. Coeff.} = \frac{\text{\# Cell Pairs Agreeing on \# of Occurrences}}{\text{\# Cell Pairs Agreeing + \# Cell Pairs Disagreeing}}$$

It is important to remember that the number of Cell Pairs Agreeing refer only to those cell pairs in which the observers reported the occurrence of the target behavior.

To find the Occurrence Reliability Coefficient for Smiling behavior in Figure 6.4, the number of cell-pairs agreeing is three (1-30, 90-120, & 150-180). There were four disagreeing cell pairs. Thus:

Occurrence Rel. Coef. = 3/(3 + 4) = 3/7 = .43

The Occurrence Reliability Coefficient for hugging behavior is 0/1 = 0.00. The Occurrence Reliability Coefficient for kissing behavior is 0/0 which is not 0.00; it is undefined since it is not mathematically possible to divide anything (including 0) b a denominator of 0. This makes sense since no cells occurred in which at least one observer reported the occurrence of kissing.

The Combined Outcome Reliability Coefficient is found by comparing the total number of cell-pairs for which both observers agreed on the number of occurrences (3 for smiling, 0 for hugging, and 0 for kissing) to the total number of cell-pairs agreeing (3) plus the total disagreeing (4 + 1 + 0). Thus:

Comb. Occurrence Rel. Coef. = (3 + 0 + 0)/ (3 + 0 + 0 + 4 + 1 + 0) = 3/(3 + 5) = 3/8 = .38

Outcome Reliability

In some situations, reliability is based on the total frequencies of a given target behavior rather than on agreement between cells. This is called outcome reliability. The formula for the Outcome Reliability Coefficient is:

Outcome Rel. Coef. = Smaller Frequency/Larger Frequency

The total frequency of smiling behavior for the first and second observers in Figure 6.4 is 9 and 9 respectively. Thus their Outcome Reliability Coefficient is:

Outcome Rel. Coef. = 9/9 = 1.00 = 100%

The Outcome Reliability Coefficient for hugs is ½ or .50 since the second observer reported 1 hug and the first recorded 2 hugs. Like the Occurrence Reliability Coefficient, the Outcome Reliability Coefficient for kissing is undefined since the frequency for kissing was zero for both observers.

The Combined Outcome Reliability Coefficient is obtained by dividing the smallest **grand total** by the largest. The Combined Outcome Reliability Coefficient for the data in Figure 6.4 is:

Outcome Rel. Coef. = Smaller Grand Total/Larger Grand Total

Outcome Rel. Coef. = 10/11 = .91

Outcome reliability is important when the **frequency of target behavior is quite high**. Under these circumstances, cell frequencies between two observers are more likely to target behavior is molecular enough and the observers are sufficiently trained, the Outcome Reliability Coefficient will still be quite respectable.

Outcome Reliability is also particularly valuable when the only available data are frequencies rather than a record of occurrences in each cell. For example, say a teacher estimates that Louise got out of her seat 30 times on a particular day and her assistant estimates she got out of her seat 50 times. The Outcome Reliability Coefficient would be 30/50 = 60%; not very impressive.

Outcome reliability can also measure the reliability of estimated behavioral products. Using an earlier example, if you wanted to assess the success of an anti-litter campaign by comparing the amount of litter in a given area before, during, and after the campaign, it might not be practical to physically weigh the amount of litter. Instead, you could take photographs at various times and have observers estimate the amount of litter in each photograph. Two observers could be used to establish an Outcome Reliability Coefficient for the estimates.

Acceptable Reliability Coefficients

The criterion of .90 is generally used when observing well established behavioral categories (e.g., talking) and a criterion of .80 is used for behavioral categories that are not quite as definite (e.g. gesturing).

There are two solutions if your reliability coefficients are not acceptable. One is to define the target behavior more molecularly to reduce the amount of interpretation needed by the observers. The second is to provide your observers with more thorough training.