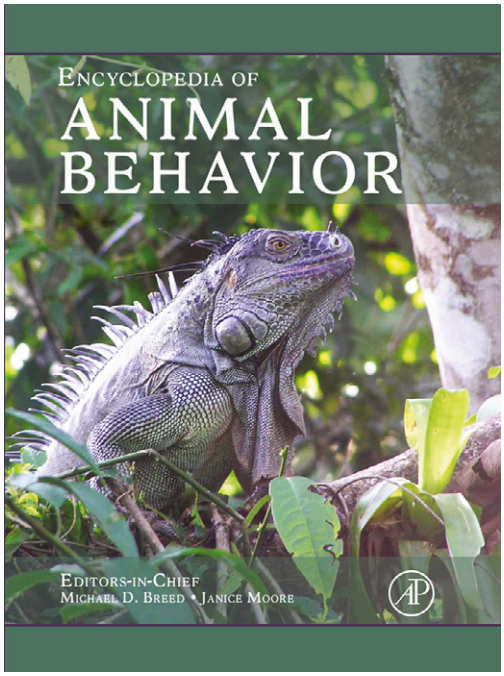

Margulis S.W. (2010) Measurement Error and Reliability. In: Breed M.D. and Moore J., (eds.) Encyclopedia of Animal Behavior, volume 2, pp. 424-428 Oxford: Academic Press.

Measurement Error and Reliability

S. W. Margulis, Canisius College, Buffalo, NY, USA

Introduction

Think about the study of behavior and what do you envision? More likely than not, 'animal behaviorist' conjures up the image of a disheveled, khaki-clad individual with binoculars and a clipboard, sitting in the midst of a jungle, jotting down notes about the fascinating behaviors he or she sees amidst a large and complex group of mammals. The idea that behavioral observation is a subjective, casual endeavor is far from true. With the expansion of ethology in the 1930s, the idea that animals could be observed in natural settings steadily grew in scientific importance. As the field of ethology and behavioral ecology expanded, there came an explosion of research methods, conventions, and practices. While all of these may have been internally valid (i.e., provided quantitative, reliable measures for the particular study for which they were designed), it was difficult, if not impossible, to generalize to a larger population or compare across studies as a result of these methodological and analytical differences. Thus, external validity was compromised because of a lack of standardization and systematic data collection rules. In 1974, the seminal paper published by Jeanne Altmann provided a critical conceptual framework and operational guide for behavioral data collection and quantification. Virtually all observational data ascribe to one of the methods outlined in this paper. These methods were designed not only to provide some degree of standardization to the discipline, but also to reduce bias by structuring observations such that an observer's choice of which subject to watch and what behaviors to record was based on a priori decisions and statistically valid procedures.

As technology has changed, so too have data collection methods. The image of the field researcher with a clipboard has been replaced by the researcher with a PDA (personal data assistant) or other handheld device; live observation may be replaced by digital video recording followed by playback, and analytical methods have grown in complexity as computers have become routine. That being said, as with any type of scientific investigation, there can be sources of error. As in any science, understanding the nature or these errors is essential in order to proactively control for their effects methodologically, or to account for them statistically at the conclusion of the study. Here, the key issues in the measurement and control of inter- and intraobserver reliability in observational research, and the methods and strategies for understanding and controlling these sources of error are discussed.

Behavioral Methodology: A Brief Overview

While every observational research study has its own unique study design and methodology, virtually all studies use as their starting point one of several basic observation techniques. Although other articles in this series will go into the details, a broad brush overview of these methods is warranted here.

Behavioral data collection schemes are based on several key concepts: (1) what does one observe (a single individual or a group); (2) how does one record observations (continuously, or instantaneously), and (3) what behaviors does one observe (establishment of a clearly defined ethogram). By utilizing various combinations of these three conceptual ideas, a study can focus on particular aspects of individuals, groups, and behaviors. Each factor requires careful planning, testing, and training to minimize errors.

Error can be introduced at a number of junctures in a study. For example, if observers are inaccurate in their ability to identify individuals quickly and accurately, they may erroneously ascribe behaviors to the wrong individuals. An additional source of error can be introduced into the data recording scheme if observers fail to time behaviors accurately. Ethograms with incomplete or vague behavioral descriptions can lead to excess variability in how observers interpret behaviors and thus lead to missed or misidentified behaviors.

In general, most behavioral data collection schema involve one of two approaches: in some cases, a continuous recording approach, in which a single individual is observed (continuous, focal observation sensu Altmann), and the onset time of every behavior or behavior transition is recorded. Alternatively, an instantaneous approach is used, which may be applied to a single individual or to a group (point or scan sampling). In this case, the behavior in which an animal is engaged at a particular point in time, usually signaled by a stopwatch, is recorded. Each of the methods has its own set of advantages and disadvantages such that no single methodology is appropriate in all cases. Thus, one must be well versed in the particular strengths and pitfalls of each method in order to decide on the best fit for a particular study, and to recognize that each method has inherent sources of error that must be understood and addressed.

Which method to choose is based on the question that the researcher is trying to answer. An instantaneous or scan-sampling approach is most appropriate when behaviors

Encyclopedia of Animal Behavior (2010), vol. 2, pp. 424-428

of interest are defined as state behaviors; this method does not necessarily require that individuals be recognizable (though it is preferable). Detailed interactions are not readily quantified using this method (though it is possible to combine a scan approach with select, continuous observation for highly visible, key behaviors). A continuous, focal approach is often used when interactions are an important component of a study, and when both event and state behaviors are to be recorded. This method often requires more rigorous training before observers attain a sufficient level of comfort with the procedures. Other standard methods are also available to the researcher, but are not discussed here and will be covered fully in other articles.

The Importance of Ethogram Construction

The importance of a clearly defined and consistent ethogram is often overlooked. Recent efforts to develop some level of standardization of ethogram structure and terminology (e.g., EthoSource, described by Martins, and related ontologies described by Midford and colleagues; or SABO, as outlined by Catton) have made progress in this regard. However, there is still considerable variation in the structure, detail, and terminology of ethograms. Use of terms that may be synonymous in ethograms can lead to confusion, and lack of specificity of definitions can result in errors. In most ethograms, behaviors may be defined functionally, in which the presumed use of the behavior is implied, or operationally, in which no specific function is assigned and the description, or definition, provides details on the motor patterns associated with the performance of the behavior. Animal behaviorists often use functional definitions and assumptions; however, in some circumstances, it can be difficult for an observer to reliably identify behaviors functionally. Play and aggression – both functional categories – may involve similar motor patterns, and clear and precise operational definitions may be critical, particularly for novice observers or for a species that has not been well studied such that functionality cannot be satisfactorily ascribed.

The level of ethogram detail is another critical component of study design that influences observer accuracy. A hierarchically structured ethogram can facilitate ease of use, with more detailed, deeper levels of behavioral description used for studies that are narrowly focused or when highly experienced observers are available.

Sources of Error

The nature of behavioral research is fundamentally no different from any other branch of scientific inquiry. Sources of error can be introduced into any study at various levels. While they can be controlled for and minimized, it is impossible to eliminate them. Being aware of these sources and how they might bias data are fundamental to the conduct of good science. The three most common means by which error can be introduced into a behavioral study include observer error, equipment error, and computational error. It must be stressed that these sources of error are common to all scientific investigations and not unique to behavior. The role of the observer is perhaps more critical to behavioral observation than what may be the case for certain types of laboratory sciences, and will be the focus of the remainder of this article. Equipment and computational error are briefly touched upon in the next section.

Equipment and Computational Error

While behavioral data collection has advanced from paper and pencil check-sheets to, in the majority of studies, electronic data collection systems, data collection is nevertheless subject to recording errors. These may involve the failure of electronic devices (particularly in the field), transcription error, and coding error. Careful review and proofreading of all data can alleviate many of these problems. Use of computer-aided data collection tools does not negate the need to review entered data. Tapping an incorrect box on a PDA screen is no less likely than checking the wrong box on a paper check-sheet. There have been numerous times when I have proofread a dataset, confident that it was error-free, only to discover data entry errors.

Computational errors generally occur during the data analysis phase of a study; however, use of statistical packages minimizes the errors here, provided the user understands the assumptions and rationale of the statistical software being used. Statistical textbooks and software user guides are of course essential; however a number of recent works have emphasized statistical issues that are more common in behavioral studies, particularly those relating to small sample size, repeated measures, and generalizability (see e.g., Kuhar's or Plowman's more extensive treatment of this subject). Behavioral data analysis often involves one or more levels of data tabulation and summary before statistical analyses can be conducted. These may be done in an automated fashion using behavioral or statistical software, or it may be done by hand before data are entered into a computerized system. Again, double-checking and proofreading all such intermediate phases can minimize the probability of such calculation errors.

Observer Error

The role of the observer is critical to the successful collection of behavioral data, but observer error has the

potential to be the most significant error component of behavior studies. However, clear guidelines and methods exist to ensure that such errors are minimized, acknowledged, and controlled. Because of the tremendous variability among observers, we must be cognizant of how to recognize, measure, and control for individual variation to assure a sound study design. Observers have the potential to introduce variation into behavioral investigations in several ways. First, the very presence of an observer may alter the behavior of the subjects. Second, observers may perceive events differently, based on their view of a particular situation or group (errors of apprehension). Third, observers may err because of lack of training or experience or because protocols and ethograms are unclear. Individual observers enter into an investigation with their own personal biases which may have the potential to influence the quality of their data collection as well. Finally, as already discussed, observers may record their observations incorrectly or may have difficulty utilizing equipment. All of these sources of observer error can be addressed and reduced via training and regular assessment of reliability and validity.

Observer Presence

The idea that an observer alters the behavior of the animals he or she observes has been debated for decades and leads to a conundrum: how can we observe natural behavior, if, by definition, we alter the behavior that we are observing simply by our presence? Use of video and remote recording devices is one way to address this concern; however, much observational data collection is – and will always be – done via live observation. Maintaining standard observational protocols holds the observer effect constant and while it may be that behavior is altered, it is in theory altered consistently across all subjects, thus enhancing internal validity. Long-term field studies have demonstrated that most animals can habituate to observer presence, suggesting that the observer's effect on the individuals that are observed may be relatively minor.

Errors of Apprehension

When two observers watch the same animal from different vantage points, differences in perspective may alter the extent to which they perceive a particular event. This is a problem primarily when observers' movements are constrained in how and where they are able to move in the area in which they are observing. This may be the case in a laboratory or captive situation in which animals may be out of view of the observer, or the observer's vantage point may prevent a clear view. In nature, observers' movements may be constrained by the activity of the animal they are watching or by other animals in the group. Ideally, simply changing one's physical position (when the observer is able to do so) to obtain the best possible view of an interaction can mitigate apprehension error. This can be a problem when conducting interobserver reliability tests (to be discussed later).

Observer Error and Bias

As already discussed, there is no single standard protocol for observing behavior. Although there are methodological standards, every study, every individual subject, and every study setting is unique. Thus, training observers is a time-consuming, tedious, but critical component of any investigation and will improve internal validity. It is only through rigorous training and ongoing monitoring and evaluation that one can maintain an acceptable level of interobserver agreement. Even an experienced researcher will require some time to become familiar with their subjects, and to ascertain the validity of their ethogram. Vague or equivocal definitions, for example, can lead to confusion among observers. Lack of experience with data recording systems can be a source of error, until observers have practiced sufficiently and are comfortable with the protocol, the layout of the datasheet, the codes used to record information, and so on. Novice observers often enter into observations with preconceived and oftentimes erroneous notions about behavior, and it may take some time and effort to move them from a subjective view of behavior in which they interpret and read meaning into behavioral patterns and events, to a more objective, consistent ability to record actions without assuming intent or meaning. Dissuading observers of their preconceptions is often the most challenging part of training observers.

Once this challenge of reducing observer bias is met, even a trained observer who has passed standardized reliability tests may diverge from that standard over time. Just as any process may need to be calibrated periodically, so must observer reliability to avoid observer 'drift' in recording of behavioral information. Regular review and repeated reliability testing can address this error.

Reliability and Validity

Reliability is an indicator of how repeatable one's results are, and is critical to maintaining accurate data collection. Unlike measuring weight or length for example, in which the potential exists for getting precisely the same measurement repeated times, it is highly improbable that an animal or a group of animals will perform exactly the same behaviors in the same way if measured multiple times. Careful data collection designs, however, can ensure consistency and standardization, which in turn improves repeatability. This is particularly important in long-term field studies, where data may be gathered by multiple observers over a period of years, or decades.

Training and adhering to a standard of accuracy and precision is critical.

The terms 'accuracy' and 'precision' may be considered synonyms in some disciplines (in fact, the thesaurus program of my word processor indicates that they are indeed synonymous), but in the case of behavioral observation, they are subtly but distinctly different. Accuracy refers to how close a recorded observation is to reality ('the truth'), whereas precision refers to how consistently an observer records the same behavior in the same way. Methodological differences sometimes necessitate a trade-off between accuracy and precision. A simple ethogram with clearly defined definitions may facilitate good precision among observers – for example, it is relatively simple to identify an animal as being active or inactive. However, there may be a loss of accuracy in that the behavioral categories may be too broad to adequately answer the study's main questions. In addition, precision may be used as an indicator of intraobserver reliability: that is, to what extent does an individual consistently observe behaviors in the same way? Accuracy is an important element of evaluating how good a study design is at collecting data to answer the question at hand: that is, to what extent do data reflect reality? How suitable is the chosen research design in answering the question that one has posed? Thus, the internal validity of an investigation is closely linked to the applicability of the methods chosen to answer the question posed. External validity is a measure of the generalizability of results to other study populations or species, as the case may be. This may be linked to the ethogram chosen and how broadly applicable it is. Reliability and validity are both essential measures that one must evaluate in terms of both inter- and intraobserver reliability.

Intra- and Interobserver Reliability

Intraobserver reliability can provide a measure of consistency and repeatability. Regular review of methods and ethogram, and reliability testing (to be described below) can provide a quantifiable measure of intraobserver reliability. Because it is common to use multiple observers for behavioral studies, either simultaneously (to maximize efficiency of data collection) or sequentially (to maintain ongoing, longitudinal investigations), it should come as no surprise that maintaining a high standard of interobserver reliability may be the most important aspect of ensuring accurate and precise data for behavioral investigation. Every observer comes into a study with his/her own set of biases and tendencies. Careful and rigorous training are essential to the conduct of behavioral studies. While there is no single training protocol for observers, convention necessitates extensive training on observation methods and animal identification, familiarity with the ethogram and data collection devices, and practice, either supervised or unsupervised, until the observer feels a degree of comfort with the methods. It is at this point that formal interobserver reliability testing should be initiated. Interobserver reliability encompasses a number of statistical approaches that facilitate a comparison between observers: that is, how similar are the data collected by two researchers who observe the same individual at the same time? Theoretically, they should be identical, but in practice, this is rarely the case. Two individuals weighing the same standardized weight on a balance are unlikely to get exactly the same measure, but they should be quite close; similarly, two researchers observing the same individual at the same time may not record exactly the same sequence of behavior, but differences should be minimal and most importantly, they should be random. Often, the conduct of interobserver reliability tests can highlight weaknesses in the study design or protocols. If for example, an observer is consistently misscoring a particular behavior, it may be that the observer needs more training and practice; however, it may also be the case that the behavior is not adequately defined on the ethogram.

Techniques for Measuring Reliability

Most measures of inter- or intraobserver reliability utilize simultaneous observation of the same individual, or independent scoring of videotaped footage. In both cases, the goal is to have observers independently score samples of behavior that should be identical if there were no observer error or bias. When two observers conduct simultaneous, live observations, it is critical that they not communicate with each other as this could influence the outcome by violating assumptions of independence. This can be challenging. If, for example, one observer notices that a second observer is entering a behavior that the first observer may have missed, this could lead the first observer to rethink his/her data entry and add a behavior that he/she might otherwise have erroneously missed. Conversely, two observers are, by definition, viewing a situation from slightly different vantage points and therefore may not be able to see exactly the same sequence of behavior because of errors of apprehension. However, this does not necessarily imply that their data are not reliable, since they may have been unable to adjust their position.

When using live observation, the likelihood that only a small subset of possible behaviors will be seen is high. Should two observers be considered to have high reliability if they both correctly score a subject as sleeping for 20 consecutive scans? The use of videotaped sequences of behavior resolves a number of problems. First, observers are able to watch and score videotape individually and independently, without possible influence from other observers. Second, the researcher can utilize one or more segments of footage that encompass a greater range of

behaviors on the ethogram, thus providing a more rigorous test of observer accuracy. Finally, all observers are able to view the sequence portrayed on the videotape from the same perspective.

Details of reliability measurements can be found in sources listed at the end of this article, and a particularly clear example of how to calculate the various reliability metrics can be found in Lehner's book; however, they are briefly described here.

Assessing Reliability via Concordance

A number of statistical methods exist for quantifying observer reliability, and all are based on a similar premise: to what extent do data collected by two individuals (or by one individual at different points in time) agree? In its simplest form, this may mean evaluating percent agreement. For example, consider an animal that is observed for 10 min, and the state behavior in which it is engaged is noted every minute on the minute (an instantaneous sampling approach). If two observers record data on the same individual for these 10 min, the 'agreement' between their datasets is easily calculated: How many of the 10 point observations are the same? If all are identical, then the agreement is 100%; if nine out the ten are identical, then agreement is 90%. A variation on this is the kappa coefficient, which corrects for chance agreement.

Kendall's coefficient of concordance can evaluate reliability evaluations with more than two observers; however, data must be converted to ranks to accommodate this nonparametric approach. Most behavioral studies look for agreement at or above 90% before an observer is considered to be 'reliable.' There is no hard and fast rule on this, however, so this value should be thought of as a guideline only. Most often, new observers are tested against a standard (the lead investigator, or main field assistant, for example).

Assessing Reliability via Correlation

Several statistical tools are available to measure correlations between nominal, ordinal, interval, and ratio data. The Phi coefficient measures correlation between nominal variables; for example, comparing the number of times two observers score a particular behavior. Similar standard statistical measures of correlation are appropriate for evaluating interobserver reliability. Spearman correlation is used for ordinal or ranked data, and Pearson correlation for interval or ratio data. Correlation coefficients range from 0 to 1, with higher values indicating better agreement. In general, a correlation coefficient > 0.7 is considered a strong correlation.

Maintaining Reliability and Consistency

The goal of behavioral research, as with any scientific endeavor, is to collect accurate, reliable data that allow the scientist to answer the question posed. The methodology chosen should fit the question at hand; it should be tested and modified to maximize its effectiveness, and its efficacy evaluated before finalizing data collection plans. It is imperative that observers be trained and their reliability – their accuracy, precision, repeatability, and validity – tested prior to utilizing their data, and regularly throughout the period of data collection.

See also: Ethograms, Activity Profiles and Energy Budgets; Experiment, Observation, and Modeling in the Lab and Field; Experimental Design: Basic Concepts.

Further Reading

Altmann, J (1974). Observational study of behavior: Sampling methods. *Behaviour* 49: 227–266.

Caro, TM, Roper, R, Young, M, and Dank, GR (1979). Inter-observer reliability. *Behaviour* 69: 303–315.

Catton, C, Dalton, R, Wilson, C, and Shotton, D (2003). SABO: A proposed standard animal behaviour ontology. www.bioimage.org/pub/SABO/SABO.

Kuhar, CW (2006). In the deep end: Pooling data and other statistical challenges of zoo and aquarium research. *Zoo Biology* 25: 339–352.

Lehner, PN (1996). *Handbook of Ethological Methods,* 2nd edn. Cambridge: Cambridge University Press.

Martin, P and Bateson, P (2007). *Measuring Behaviour: An Introductory Guide,* 3rd edn. Cambridge: Cambridge University Press.

Martins, EP (2004). EthoSource: Storing, sharing, and combining behavioral data. *Bioscience* 54: 886–887.

Midford, PE (2004). Ontologies for behavior. *Bioinformatics* 20: 3700–3701.

Paterson, JD (2001). *Primate Behavior: An Exercise Workbook*. Long Grove, IL: Waveland Press.

Ploger, BJ and Yasukawa, K (eds.) (2003). *Exploring Animal Behavior in Laboratory and Field.* New York, NY: Academic Press.

Plowman, AB (2008). BIAZA statistics guidelines: Toward a common application of statistical tests for zoo research. *Zoo Biology* 27: 226–233.

Stamp Dawkins, M (2007). *Observing Animal Behaviour*. New York, NY: Oxford University Press.