Animal Behavior Reliability.

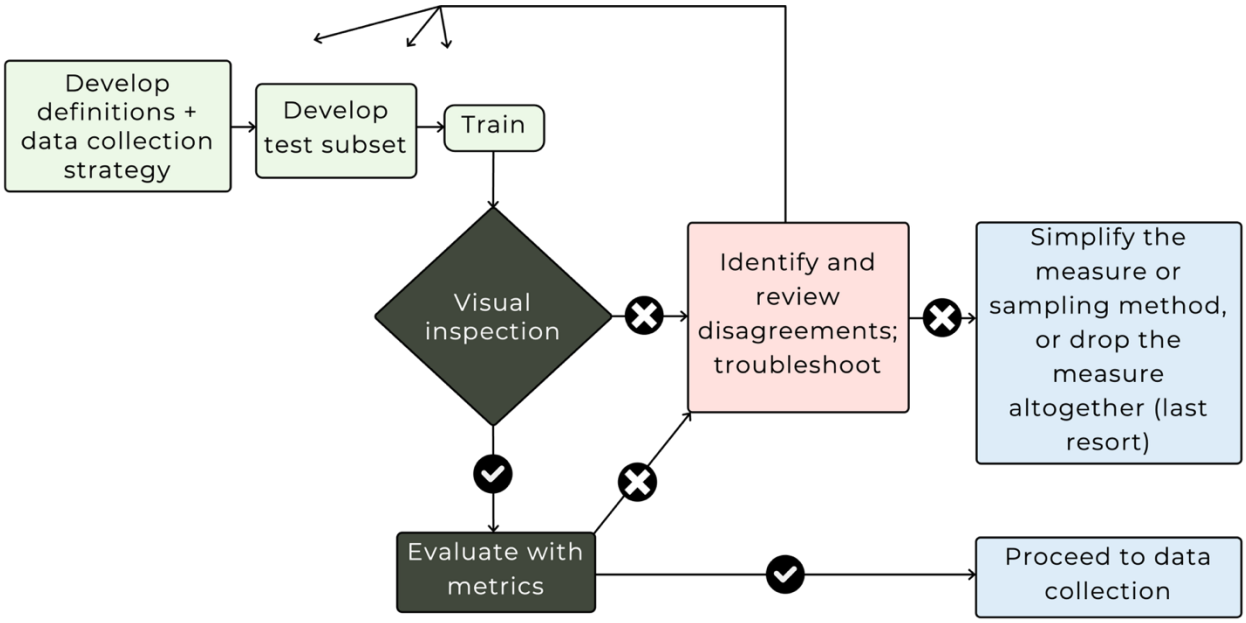
Reliability Guidelines: Essential and Recommended components to address as you begin, complete, and publish experiments. This checklist is designed to be used in combination with the website, or as a reminder for a user who has already worked through the site. For more detail on any listed item, refer to the website.

Before the experiment:

- ☐ **Research questions.** Develop your research proposal, hypotheses, predictions, and dependent variables of interest by reading scientific literature, having discussions with experts, and observing your species. Dependent variables should be relevant/observable in your species and address your research questions.
- ☐ **Definitions or criteria for dependent variables.** Observe your animal and collect photo/video examples of your variables of interest prior to data collection. Consider whether your definitions:
  - ☐ Describe characteristics in a way that is uniformly understood
  - ☐ Are non-circular
  - ☐ Avoid assumptions
  - ☐ Provide relative estimates in view, if estimates of distance or size are required
  - ☐ Provide criteria for starting and ending an event
- ☐ **Pilot testing.** Determine your approach to data collection and pilot test it. This may include testing camera angles, lighting, and image resolution to ensure all relevant dependent variables can be seen before you begin the experiment.
- ☐ **Develop training process: training yourself against an expert.** Develop a reliability training and testing process to compare yourself against a gold standard or another expert (your PI, trusted colleague, etc.).
- ☐ **Build reliability tools: robust test subsets.** Robust test subsets for reliability meet the following criteria:
  - ☐ Captures variation in how individual animals behave
  - ☐ For every outcome, includes multiple examples with presence/absence
  - ☐ Includes challenges observers will face, e.g. camera angles or lighting
  - ☐ Duration of training representative of formal data collection
  - ☐ Balances training goals against investment of time
  - ☐ Matches modality (live vs. video) of formal data collection
  - ☐ Matches sampling method of formal data collection
- ☐ **Build reliability tools: rare outcomes.** If a rare outcome is a dependent variable of interest, try to develop a test subset by reaching out to other researchers; if this is not possible, try to gather examples to develop a task-based training protocol; if neither of

these are possible, obtain 1-3 examples of the parameter to illustrate it for the team conducting data collection.

- ☐ **Refine training process and tools used: multiple training sets.** If test subset does not meet all criteria (e.g. using video when data will be collected live, or sets that show only some of the targeted outcomes), create multiple rounds of testing and compare reliability consistently across observers (e.g. average all trainings, only use the final training, etc.).
- ☐ **Develop training process: evaluating your own consistency.** Use the same training process that you used with an expert to evaluate your own consistency (intra-rater reliability).
- ☐ **Refine training process and tools used: minimize memorization.** Develop a reliability testing process that can be randomized or conducted over a sufficiently long time period to reduce memorization, particularly for your own training and if re-training may be required for your team members.
- ☐ **Develop training process: training others involved with data collection.** Develop a reliability training and testing process that is relevant for the size and experience level of your team. Teams that involve beginners (to the research process, species, or just the dependent variables of interest) can require a scaffolded approach.
- ☐ **Implement your training process.** Conduct reliability training with all team members involved in data collection. All team members, including you, should be exposed to the training used to generate metrics reported in the manuscript, regardless of individual differences in experience.
- ☐ **Evaluate effectiveness of your training process.** Use a combination of visual observation and a robust reliability test metric that fits your data type to assess if team members are consistent with an expert or gold standard. Only team members that meet criteria should continue to data collection; re-training on the same test subset, or training on new subsets, may be required for those that do not meet criteria.



During the experiment:

- ☐ **Use metrics to check for drift: within or between observers.** Check intra- or inter-rater reliability for each team member involved in collecting data to evaluate whether there was drift from the beginning of the experiment, either from themselves or against the gold standard.
- ☐ **If there is evidence of drift, revise.** If individuals are not consistent at this point, revision (re-training individuals and re-scoring data collected by them, if possible, or revising the approach to data collection more broadly) may be needed.

After the experiment:

- ☐ **Use metrics to check for drift: within or between observers.** Check intra- or inter-rater reliability for each team member involved in collecting data to evaluate whether there was drift from the beginning of the experiment, either from themselves or against the gold standard.
- ☐ **If there is evidence of drift, revise.** If individuals are not consistent at this point, revision (re-training individuals and re-scoring data collected by them, if possible, or revising the approach to data collection more broadly) may be needed.
- ☐ **If metrics are not possible: rare outcomes.** If rare outcomes of interest were included and did not have enough examples prior to data collection to allow for formal metrics, and data collection is a modality that can be revisited (e.g. video), review each instance identified by observers to confirm correct identification.

Writing the manuscript:

- ☐ **Report the dependent variables: definitions or criteria.** Describe all definitions or criteria used, providing the same level of detail given to all team members during training and data collection.
- ☐ **Report the dependent variables: examples.** Provide photo or video examples of the definitions to demonstrate what they look like in practice.
- ☐ **Report your training process: size and composition.** Describe your team size used for data collection; include who was used as the expert or gold standard.
- ☐ **Report your training process: test subsets.** Describe features of your test subset(s) used in training: modality, how many images or sessions, how long were the videos used in the test, how many unique individuals were used in the images, whether all possible dependent variables of interest were included in the training, etc.
- ☐ **Report your training process: repetition or revision needed.** If multiple levels of training were used (e.g. initial orientation and a final, formal test), describe these and note which dataset was used to generate your reliability metric(s). If individuals varied in how many trainings they received (e.g. because some needed re-training), note the variation, such as by describing the minimum and maximum number or hours of training required to reach success.
- ☐ **Report your training process: metrics.** Describe the inter-observer reliability metric(s) used, and the cutoff value(s) that observers needed to meet to complete training; describe the intra-rater metric used for the expert. Describe whether cutoff values or training differed for any outcomes.

- ☐ **Report your training process: timing.** Describe if training occurred before, during, or after data collection. If reliability was evaluated at multiple timepoints, describe the process at each point and what the metric and cutoff values were. If the same training was used at each timepoint, describe whether efforts were made to limit memorization.
- ☐ **Report rare outcomes.** If rare outcomes of interest were included, consider reporting and discussing them in the manuscript even if the outcome was not analyzed with statistical models; consider uploading examples to an open-source platform to help other researchers with future development of a test subset.
- ☐ **Acknowledge drift or poor reliability.** If individuals were not reliable overall or consistent over time and revision was not possible (e.g. data were collected live on growing animals and cannot be re-scored), poor reliability should be addressed in the manuscript.