CS 294-115 Algorithmic Human-Robot Interaction **Fall 2015**

Lecture 5: Experimental Design

Scribes: Jessica Hamrick & Andrew Head

5.1 What is an Experiment?

5.1.1 Origins of Experiments

Some of the first work on experimental design was by Fisher (1929), who studied agriculture. In particular, he was interested in understanding the impact of the use of pesticides on crops. Historically, experiments have also been used in medicine to determine drug effectiveness.

More recently, experiments have been used in psychology to better understand people. In human-robot interaction, experiments can be used to test design decisions for robots. Anca recommends that rigorous experimental design become more commonplace in robotics. It's important to compare new algorithms' performance to competing algorithms, and not all technical works do this to the right extent.

5.1.2 Components of an Experiment

There are four main components that are required to run an experiment. When you are designing an experiment, you should think carefully about each one and make sure you know what they are for your specific case.

5.1.2.1 Conditions or Treatments

First, the conditions in your experiment are what we will vary in order to see some change in behavior. You can think of the conditions as your independent variables (see Section 5.2.4 for more details on independent variables). For example:

- In the case of a medical experiment, you may want to test how well a drug works, in which case your conditions would be ”drug” and ”placebo”.
- In the case of a robot interacting with a human, you may want to test how the human responds to different motion planners, in which case your conditions might be ”random motion” or ”optimal motion”.
- In the case of robotics algorithms, you might want to compare two different algorithms, in which case your conditions might be ”CHOMP” vs. ”GSCHOMP” (where GSCHOMP includes a *goal set* rather than just a single goal position).

5.1.2.2 Responses or Measures

Once you know what you want to test, you need to determine what the metric is that you're actually using to measure differences between your conditions. You can think of the responses or measures as your dependent variables (see Section 5.2.4 for more details on dependent variables). For example:

- In the medical case, you might want to measure symptom progression.
- In the human-robot collaboration case, you might want to measure human comfort or surprise.
- In the algorithm case, you might want to measure success rate or cost.

5.1.2.3 Experimental Units

Next, you need to determine the specific thing that you will be measuring from. This is distinct from the independent variable: while a drug vs. placebo might be your independent variable, you can't actually measure symptom progression from the drug itself, you need something to test the drug on. For example:

- In the medical case, your experimental units would be humans (patients, also called participants or subjects, though ”subjects” is not really used anymore).
- In the human-robot collaboration case, your experimental units would again be humans (users, also called participants).
- In the algorithm case, your experimental units would be different motion planning problems.

5.1.2.4 Assignment Method or Design

Finally, you need to decide how to assign your conditions to experimental units. The next section covers this point in more detail, but to give examples for the three cases we've been going through so far:

- In the medical case, you would probably want to assign participants randomly to conditions.
- In the human-robot collaboration case, you might want every participant to see both types of motions.
- In the algorithm case, you would want each algorithm to be tested on all motion planning problems.

5.1.3 Assigning Conditions to Experimental Units

When we assign conditions to experimental units, we can use one of three methods:

- **between-subjects**: Each experimental unit is assigned one condition (also called *across-subjects*).
- **with-subjects**: All experimental units are assigned all conditions. Measurements are taken for all conditions for each unit.
- **“mixed” design**: There are multiple experimental variables, and units receive one condition for some variables, and all conditions for others. In other words, this experiment is between-subjects with respect to some variables and within-subjects with respect to others.

Table 5.1: Pros and Cons of Assignment Methods

Assignment Method	Pros	Cons
between-subjects	• No confounds are introduced by the ordering of conditions	• More units (users) are required to achieve the same statistical significance
within-subjects	• More data can be obtained with the same number of units (e.g., users)	• Introduces learning-based confounds • Sometimes it's impossible to assign all conditions to all units (e.g., for drug testing)

You should carefully consider which variables should vary between-subjects and within-subjects when designing your experiment. Consider the pros & cons of between-subjects and within-subjects variables shown in Table 5.1. Essentially, you can often achieve greater statistical significance with the same number of users for a within-subjects design, though between-subjects can avoid confounds due to learning.

To reduce within-subjects biases, the order of conditions can be **counterbalanced**, or assigned in varying order to each unit. Even this might not be enough to ameliorate the effect of confounds for human-robot interaction (see Section 5.3.2).

Example: *Do we ever want to do a between-subjects design in robotics?* We often want within-subjects design for evaluating robotics algorithms, to make sure that we don't bias the results for one algorithm by choosing a different data set for each one. There may be some exceptional experiments where hysteresis or system memory may bias a robot's performance for tasks that come after the first task. In this case, it may be worth considering a between-subjects design.

Example: *When might we want a between-subjects design for HRI?* If robot behavior is more erratic in one condition than another, a user may be "freaked out" by the time they reach the second condition after they experienced the first one for one condition, and not for the other. In this case, there's an asymmetric impact of experiencing one condition before the other that can't be solved by counterbalancing.

Note: Learning effects may be reduced by having users complete training exercises before any measurements are taken. Training can also bring users up to even performance levels if any were biased by previous experience to perform well on the experimental tasks.

5.1.4 Operationalization of Variables

We define two major types of variables for experiments:

- **Independent variables**: what you manipulate to create the conditions.
- **Dependent variables**: what we measure (see Section 5.2.2.2)

Independent variables can have multiple **levels** – these are values we assign to the independent variable.

While we often think of independent variables that have two levels (e.g., drug vs. placebo, new algorithm vs. baseline), independent variables can have any number of levels.

In a medical experiment that measures the impact of a drug on treating an illness, an independent variable would be treatment (drug vs. placebo) with a dependent variable of some measurement of the symptoms. When testing a motion-planning algorithm, an independent variable may be the algorithm itself (yours vs. some baseline) with a dependent variable of cost, path length, or whether planning was successful.

Your experiment may also have **covariates**, or factors that vary in your population that you don't manipulate. You may want to measure these covariates. You may also want to make sure that they are balanced across conditions (see Section 5.3.2). Examples of covariates are gender for experiments involving human subjects, and the problem difficulty for testing motion-planning algorithms.

Another important type of covariate is known as **stimuli**. In your experiment, you manipulate independent variables, and measure the dependent variables from your experimental units. However, it is not usually enough to take one measurement; stimuli give you a means to measure (ideally) the same thing in slightly different ways. Take for example an experiment in which participants interact with a robot using different motion planners, and are measured for their surprise in how the robot moves. However, the robot will move differently depending on the actual *type* of movement: for example, the movement could be handing the participant an object, taking an object from the participant, shaking hands, waving, etc. Some of these movements might work better under one motion planner, while others might be better under the other. Thus, it is insufficient to test just one of them (unless you are only concerned with a specific type of motion). If you tested participants on multiple different movements for each robot, these movements would be the stimuli in the experiment.

5.1.5 Hypotheses

Hypotheses describe a relationship between an independent variable and a dependent variable. A template for a hypothesis is: *Independent variable x affects dependent variable y.*

Hypotheses can make claims about the directionality of the relationship: for example, circumstance x *positively* affects behavior y, or circumstance x *negatively* affects behavior y.

Practically, we can develop hypotheses to help us to elucidate key insights in our work. By defining the measurements you want to see and the circumstances you want to alter, you can better define what is unique about the algorithms you develop. Essentially, your hypothesis can help you ask: *what was the one change I've made, and its impact on this problem?*

Decide on your hypotheses before running your experiment. Ideally, this hypothesis will be grounded in established theory, observation, or past results. Exploratory studies can generate insights, but are often not appropriate to contribute to generalizable knowledge as they are not rigorously designed to test for a causal relationship without confounds.

A *mechanistic hypothesis* states a causation between a mechanism as an independent variable and a measured phenomenon. *Mechanistic hypotheses* may be described in contrast to *descriptive hypotheses* that, rather than attempting to describe a causation, report observations about a phenomenon.

5.2 What is a Good Experiment?

Good experiments are controlled, avoid confounds, are reliable, and have good construct validity.

5.2.1 Good experiments are controlled

We call an experiment **controlled** when an experimenter assigns experimental units to treatments or conditions. When the experimenter does not do this, we might call the experiment "observational."

5.2.2 Good experiments avoid confounds

A **confound** is a variable whose effect cannot be distinguished from the effect of another. We are often concerned with confounds when they vary with an independent variable. When they do this, it is difficult to attribute a change in the dependent variable to a change in the independent variable.

Example: Consider this study on estrogen treatment and its effect on health during menopause: An observational study aimed to show that estrogen treatment brought about better health outcomes for women during menopause. 93,676 women participated. This study tracked women over 8 years. The experimenters observed whether women had taken estrogen, and measured health outcomes.

The study found that estrogen treatment was significantly correlated with positive health. One possible takeaway is that estrogen improves health during menopause. However, this is not necessarily the case as *there was a confound*. Those who took estrogen may have just been more health conscious to begin with. In fact, in another controlled study, estrogen led to worse health outcomes.

Example: In another canonical example, one study concluded that those who are left-handed die younger. The study counted the number of people who were left-handed across a large age range. However, there was a confound at play: left-handedness was taboo decades ago. It is likely that few people who were born left-handed in the older age groups reported being left-handed during the study.

One confound that may arise when comparing algorithms is that meta-parameters may be finely-tuned for the experimental algorithm, but not for the control algorithm. However, for generalizability, it's important to emulate how people would use baseline methods in the real world. Take some time to optimize the meta-parameters for the control algorithms on some training set as if it was your own algorithm.

The next few sections go through several tools which you can use to avoid confounds.

5.2.2.1 Randomization

For between-subjects designs, you can avoid some confounds by randomly assigning participants to conditions. Note that this is different from *haphazard* assignment, which would be picking an assignment arbitrarily, such as that every other participant is assigned to the first condition, and everyone else is assigned to the second condition. This type of haphazard assignment is undesirable because it can introduce additional confounds due to the order that participants sign up for your experiment.

5.2.2.2 Counterbalancing

For within-subjects designs, you can avoid confounds by counterbalancing the order of your conditions.

Example: If you have two conditions (drug and placebo), then for half your participants, you should give them the drug first, and for the other half, you should give them the placebo first.

Unfortunately, as the number of conditions in your experiment increases, so does the number of permutations: if you have N conditions, then there will be $N!$ permutations of those conditions that need to be counterbalanced. One method for getting around this issue is called a *latin square* design, which is a particular type of blocking (see Section 5.3.3.2).

5.2.2.3 Stack the cards against yourself

If you design your experiment in a way that is unfavorable to the results that you hope to get, then if you do get those results, you can make a stronger claim.

Example: In the case of comparing motion planning algorithms, what metric do you use to compare them? Typically, the answer would be ”cost”, but ”cost” is different for different algorithms. For RRT+path shortening, the cost is proportional to path length, but for CHOMP, TRAJOPT, and STOMP, the cost is given by a more complex cost function $\mathcal{U}$. In this case, ”stacking the cards against yourself” would mean to evaluate all algorithms on path length, to give RRT a chance.

5.2.2.4 Optimize the baseline algorithm

As mentioned in Section 5.3.2, you should optimize the metaparameters for *all* your algorithms, not just your algorithm. As Anca mentioned in class, if you are going to spend a week tweaking the parameters of your own algorithm, then you should be doing the same for any algorithms you are comparing against. Do not simply use an algorithm out-of-the-box, as it is unlikely to be optimized for your problem domain.

5.2.3 Good experiments are reliable

The term **reliability** refers to experiments that have low experimental error or variance. This is important, because if your experiment has high variance, then you may not get the same results if you run it a second time! See Figure 5.1 for an illustration of reliability versus validity (Section 5.3.4).

The next few sections go through several tools which you can use to help increase reliability.

5.2.3.1 Within-subjects design

If you use a within-subjects design, then you are controlling for variance between subjects, this resulting in a more reliable experiment.

5.2.3.2 Blocking

A *block* is a group of homogeneous experimental units, and *blocking* is the practice of arranging experimental units into blocks.

Example: If you have two treatments (drug and placebo) and you have a factor of gender in your participants, then you could block your experiment such that exactly half the males get the drug, and the other half get the placebo, and likewise for females.

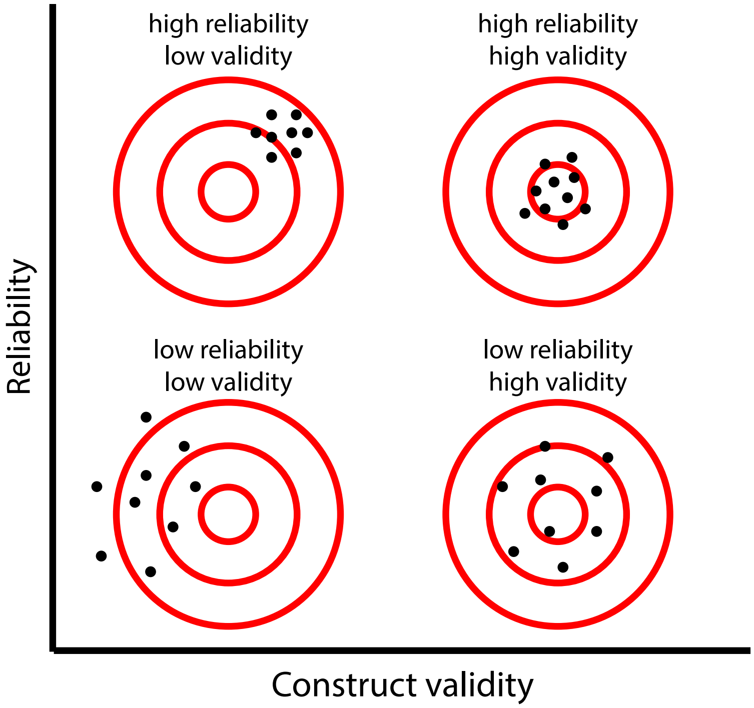

Figure 5.1: An illustration of construct validity vs. experimental reliability. Low validity means you are not measuring the thing you are aiming to measure. Low reliability means that your experiment has high error or variance, and that you might not always get the same results. Ideally, your experiments would have both high reliability and high validity.

5.2.3.3 Multi-item scales

If you want to measure responses from your participants, particularly subjective responses, then it is a good idea to use ”multi-item scales” which involve asking participants related—but different—questions about what you want to measure. Usually, each scale will have a discrete number of items (e.g., 5 or 7). This type of scale is called a *Likert scale* an is widely used in the behavioral sciences.

Example: If you wanted to measure how well participants could anticipate what a robot is going to do, you might ask them about how predictable its motion was, how expected its motion was, and how surprising its motion was.

5.2.4 Good experiments have construct validity

The term **construct validity** refers to measuring what you actually intend to measure. For example, while the IQ test claims to measure ”intelligence”, it is arguably not a very good measure of intelligence. Thus, an experiment measuring intelligence via IQ would not have good construct validity. See Figure 5.1 for an illustration of reliability versus validity.

One tool for improving construct validity is to include both *objective* and *subjective* measures. In the case of human-robot collaboration, an objective measure might be how long it takes the human and robot to jointly complete a task. A subjective measure might be how helpful the participant feels the robot was.

Note: it is particularly important to include multiple measures in experiments involving humans, because people frequently will give you different responses depending on how you ask them the question.[1]

[1] An interesting example on the vast discrepancies you can sometimes find between subjective and objective measures is that regarding how people reason about physical objects. For example, imagine a pendulum swinging from side to side. Now imaging that the string of the pendulum is cut. Where does the pendulum bob go? When prompted, participants usually will fail to draw an accurate trajectory of the bob, but they can easily catch the bob if you ask them to do so.