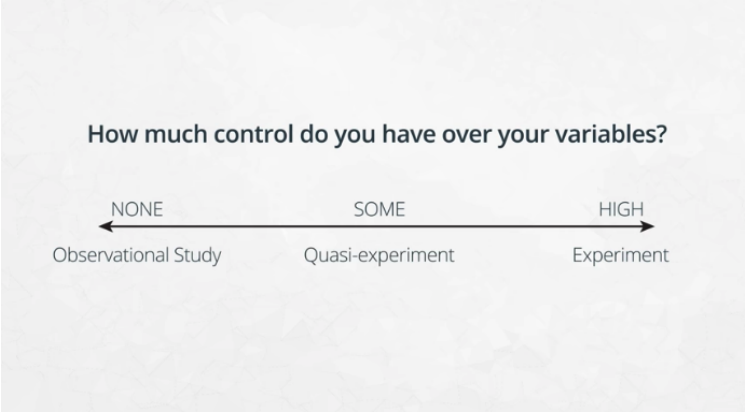
Concepts in Experiment Design

Type of Study

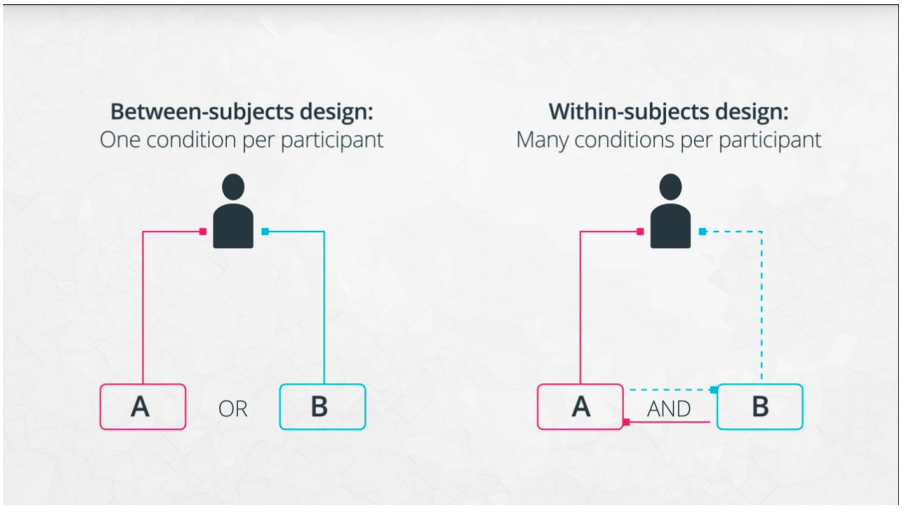
There are many ways in which data can be collected in order to test or understand the relationship between two variables of interest. These methods can be put into three main bins, based on the amount of control that you hold over the variables in play:



- **Experiments**
  An experiment is defined by comparing outcomes between two or more groups and ensuring equivalence between the compared groups except for the manipulation that we want to test. Our interest in an experiment is to see if a change in one feature has an effect on the value of a second feature. Equivalence between groups is typically carried out through some kind of randomization procedure.
- **Observational Studies**
  Observational studies are defined by a lack of control. Observational studies are also known as naturalistic or correlational studies. In an observational study, no control is exerted on the variables of interest. We typically cannot infer causality in an observational study due to our lack of control over the variables. Any relationship observed between variables may be due to unobserved features, or the direction of causality might be uncertain.
- **Quasi-Experiments**
  In between the observational study and the experiment is the quasi-experiment. While the manipulation is controlled by the experimenter, there aren't multiple groups to compare. The experimenter can still use the behavior of the population pre-change and

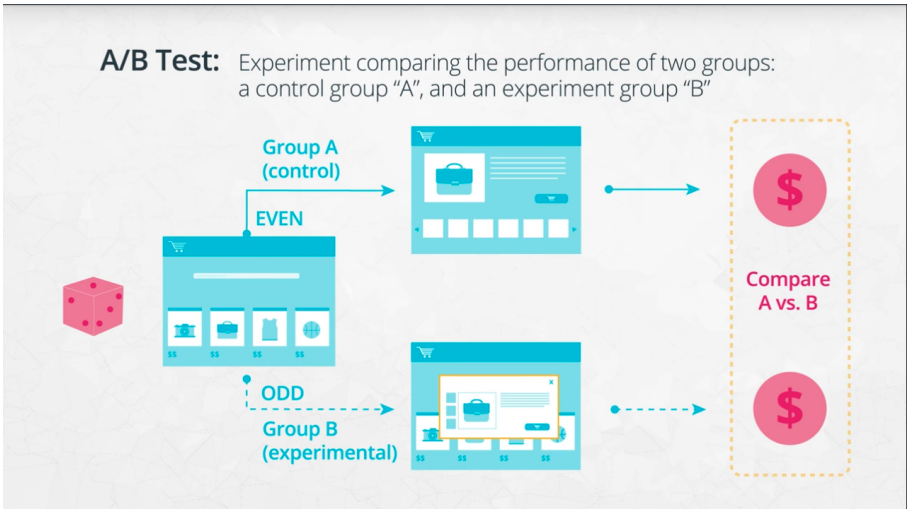
compare that to behaviors post-change, to make a judgment on the effects of the change. However, there is the possibility that there are other effects outside of the manipulation that caused the observed changes in behavior.

Type of Experiment



In a **between-subjects** experiment, each unit only participates in, or sees, one of the conditions being used in the experiment. The simplest of these has just two groups or conditions to compare. In one group, we have either no manipulation or maintenance of the status quo. This is known as **the control group**. The other group includes the manipulation we wish to test. This is known as our **experimental group**.

If an individual completes all conditions, rather than just one, this is known as a **within-subjects** design. Within-subjects designs are also known as repeated measures designs. By measuring an individual's output in all conditions, we know that the distribution of features in the groups will be equivalent. We can account for individuals' aptitudes or inclinations in our analysis.



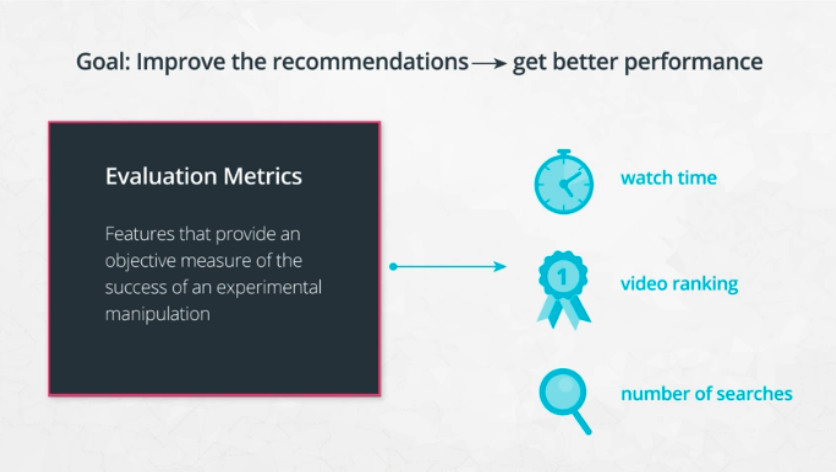
We can compare the outcomes between groups in order to make a judgment about the effect of our manipulation. Since we have an experiment, we'll randomly assign each unit to either the control or experimental group. This kind of basic experiment design is called an **A/B test**: the "A" group representing the old control, and "B" representing the new experimental change. We could have multiple experimental groups to compare to form an A/B/C test, with control group "A" and experimental groups "B" and "C".

Type of Sampling

The simplest of these approaches is **simple random sampling**. In a simple random sample, each individual in the population has an equal chance of being selected. We just randomly make draws from the population until we have the sample size desired.

However, it is possible that certain groups are underrepresented in a simple random sample. If there are certain rarer subgroups of interest, it can be worth adding one additional step and performing the **stratified random sampling**. In a stratified random sample, we need to first divide the entire population into disjoint groups, or strata.

Measuring Outcomes

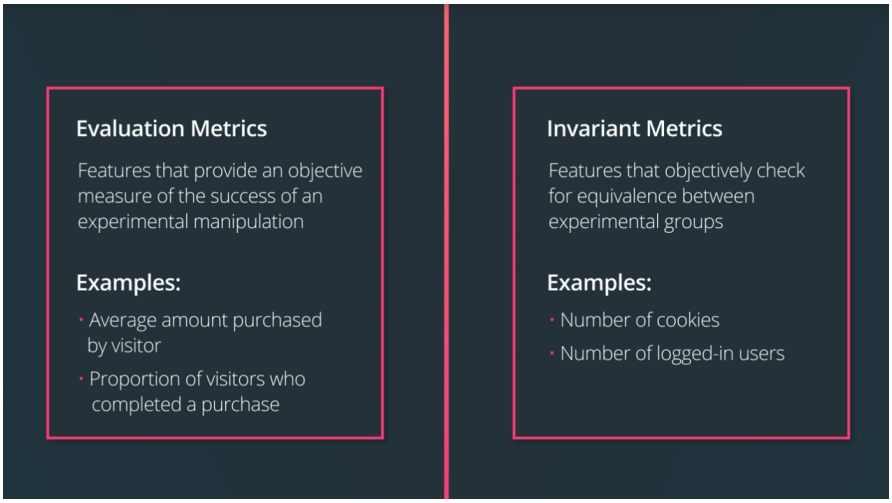


The objective features to evaluate performance are known as **evaluation metrics**. It's a good idea to consider the goals of a study separate from the evaluation metrics. This provides a couple of useful benefits. First, this makes it clear that the metric isn't the main point of a study. It's the implications of the metric relative to the goal that matters. Second, having the metric separate from the goal can clarify the purpose of conducting the study or experiment. It makes sure we can answer the question of why we want to run a study or experiment.

Creating Metrics

In a web experiment, you'll often think of the user funnel. A **funnel** is the flow of steps you expect a user of your product to take. Typically, the funnel ends at the place where your main evaluation metric is recorded and includes a step where your experimental manipulation can be performed.

Once you have a funnel, think about how you can implement your experimental manipulation in the funnel. If the goal of the above experiment was to change the way the site looks after a user clicks on a product image, we need to figure out a way to assign users to either a control group or an experimental group. The place in which you make this assignment is known as the **unit of diversion**.



There are two major categories that we can consider features: **as evaluation metrics** or **as invariant metrics**. Evaluation metrics were mentioned on the previous page as the metrics by which we compare groups. Ideally, we hope to see a difference between groups that will tell us if our manipulation was a success. On the flip side, invariant metrics are metrics that we hope will not be different between groups. Metrics in this category serve to check that the experiment is running as expected.

Controlling Variables

There are two main things to control. First, we need to enact the manipulation on one of the features of interest, so that we know that it is causing the change in the other feature. In order to know that it was our manipulated variable and not any other, the second major control point is that we want to make sure that all other features are accounted for. These two requirements make the arguments for causality much stronger with an experiment compared to a quasi-experiment or observational study. If we aren't able to control all features or there is a lack of equivalence between groups, then we may be susceptible to **confounding variables**.

Checking Validity

When designing an experiment, it's important to keep in mind validity, which concerns how well conclusions can be supported. There are three major conceptual dimensions upon which validity can be assessed:

- **Construct validity** is tied to the earlier discussion of how well one's goals are aligned to the evaluation metrics used to evaluate it.
- **Internal validity** refers to the degree to which a causal relationship can be derived from an experiment's results.
- **External validity** is concerned with the ability of an experimental outcome to be generalized to a broader population.

Checking Bias

There are numerous ways in which an experiment can become unbalanced.

- **Sampling biases** are those that cause our observations to not be representative of the population. Studies that use surveys to collect data often have to deal with *self-selection bias*. The types of people that respond to a survey might be qualitatively very different from those that do not. One type of sampling bias related to missing data is *survivor bias*. Survivor bias is one where losses or dropout of observed units are not accounted for in an analysis.
- **Novelty bias** is one that causes observers to change their behavior simply because they're seeing something new. We might not be able to gauge the true effect of manipulation until after the novelty wears off and population metrics return to a level that actually reflects the changes made.
- **Order bias** may occur when running a within-subjects experiment. The order in which conditions are completed could have an effect on participant responses. A *primacy effect* is one that affects early conditions, perhaps biasing them to be recalled better or to serve as anchor values for later conditions. A *recency effect* is one that affects later conditions, perhaps causing bias due to being fresher in memory or task fatigue.
- **Experimenter bias** is where the presence or knowledge of the experimenter can affect participants' behaviors or performance. If an experimenter knows what condition a participant is in, they might subtly nudge the participant towards their expected result with their interactions with the participant.

Ethics in Experimentation

While different fields have developed different standards, they still have a number of major points in common:

- **Minimize participant risk**: Experimenters are obligated to construct experiments that minimize the risks to participants in the study. The risk of harm isn't just in the physical sense, but also the mental sense.

- **Have clear benefits for risks taken**: In some cases, risks may be unavoidable, and so they must be weighed against the benefits that may come from performing the study. When expectations for the study are not clearly defined, this throws into question the purpose of exposing subjects to risk.
- **Provide informed consent**: This is an opportunity for a participant to opt-out of participation. However, there are some cases where deception is necessary. In cases like this, it's important to include a debriefing after the subject's participation so that they don't come away from the task feeling mislead.
- **Handle sensitive data appropriately**: Make sure that you take appropriate steps to protect their anonymity from others. Sensitive information includes things like names, addresses, pictures, timestamps, and other links from personal identifiers to account information and history. Collected data should be anonymized as much as possible; surveys and census results are often also aggregated to avoid tracing outcomes back to any one person.

SMART Mnemonic

The letters of SMART stand for:

- **Specific**: Make sure the goals of your experiment are specific.
- **Measurable**: Outcomes must be measurable using objective metrics
- **Achievable**: The steps taken for the experiment and the goals must be realistic.
- **Relevant**: The experiment needs to have a purpose behind it.
- **Timely**: Results must be obtainable in a reasonable time frame.