**Types 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:

* If you have a lot of control over features, then you have an **experiment**.
* If you have no control over the features, then you have an **observational study**.
* If you have some control, then you have a **quasi-experiment**.

While the experiment is the main focus of this course, it's also useful to know about the other types of study so that you can use them in effective ways, especially if an experiment cannot be run.

**Experiments**

In the social and medical sciences, 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 in the value of a second feature, like seeing if changing the layout of a button on a website *causes* more visitors to click on it. Having multiple groups is necessary in order to compare the outcome for when we apply the manipulation to when we do not (e.g. old vs. new website layout), or to compare different levels of manipulation (e.g. drug dosages). We also need equivalence between groups so that we can be as sure as possible that the differences in the outcomes were only due to the difference in our manipulated feature.

Equivalence between groups is typically carried out through some kind of randomization procedure. A **unit of analysis** is the entity under study, like a page view or a user in a web experiment. If we randomly assign our units of analysis to each group, then on the whole, we should expect the feature distributions between groups to be about the same. This theoretically isolates the changes in the outcome to the changes in our manipulated feature. Of course, we can always dig deeper afterwards to see if certain other features worked in tandem with, or against, our manipulation.

**Observational Studies**

In an experiment, we exert a lot of control on a system in order to narrow down the changes in our system from one source to one output. Observational studies, on the other hand, 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, perhaps due to ethical concerns or a lack of power to enact the manipulation. This often comes up in medical studies. For example, if we want to look at the effects of smoking on health, the potential risks make it unethical to force people into smoking behaviors. Instead, we need to rely on existing data or groups to make our determinations.

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. (We'll discuss this more later in the lesson.) But simply because an observational study does not imply causation does not mean that it is not useful. An interesting relationship might be the spark needed to perform additional studies or to collect more data. These studies can help strengthen the understanding of the relationship we're interested in by ruling out more and more alternative hypotheses.

**Quasi-Experiments**

In between the observational study and the experiment is the quasi-experiment. This is where some, but not all, of the control requirements of a true experiment are met. For example, rolling out a new website interface to all users to see how much time they spend on it might be considered a 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 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. For the example earlier in this paragraph, it might be that users would have naturally gravitated to higher usage rates, regardless of the website interface.

As another example, we might have two different groups upon which to make a comparison of outcomes, but the original groups themselves might not be equivalent. A classic example of this is if a researcher wants to test some new supplemental materials for a high school course. If they select two different schools, one with the new materials and one without, we have a quasi-experiment since the differing qualities of students or teachers at those schools might have an effect on the outcomes. Ideally, we'd like to match the two schools before the test as closely as possible, but we can't call it a true experiment since the assignment of student to school can't be considered random.

While a quasi-experiment may not have the same strength of causality inference as a true experiment, the results can still provide a strong amount of evidence for the relationship being investigated. This is especially true if some kind of matching is performed to identify similar units or groups. Another benefit of quasi-experimental designs is that the relaxation of requirements makes the quasi-experiment more flexible and easier to set up.

## Types of Experiment

Most of the time, when you think of an experiment, you think of a **between-subjects** 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 like providing a known drug treatment, or an old version of a website. This is known as the **control group**. The other group includes the manipulation we wish to test, such as a new drug or new website layout. This is known as our **experimental group**. We can compare the outcomes between groups (e.g. recovery time or click-through rate) in order to make a judgement about the effect of our manipulation. (Since we have an experiment, we'll randomly assign each unit to either the control or experimental group.) For web-based experiments, 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 aren't limited to just two groups. We could have multiple experimental groups to compare, rather than just one control group and one experimental group. This could form an A/B/C test for a web-based experiment, with control group "A" and experimental groups "B" and "C".

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. For example, if an individual rates three different color palettes for a product, we can know if a high rating for one palette is particularly good compared to the others (e.g. 10 vs. 5, 6) or if it's not a major distinction (e.g. 10 vs. 8, 9).

Randomization still has a part in the within-subjects design in the order in which individuals complete conditions. This is important to reduce potential bias effects, as will be discussed later in the lesson. One other downside of the within-subjects design is that it's not always possible to pull off a within-subjects design. For example, when a user visits a website and completes their session, we usually can't guarantee when they'll come back. The purpose of their following visit also might not be comparable to their first. It can take a lot more effort in control in order to set up an effective within-subjects design.

### Side Note: Factorial Designs

Factorial designs manipulate the value of multiple features of interest. For example, with two independent manipulations "X" and "Y", we have four conditions: "control", "X only", "Y only", "X and Y". Experimental designs where multiple features are manipulated simultaneously are more frequently seen in engineering and physical sciences domains, where the system units tend to be under stricter control. They're less seen in the social and medical realms, where individual differences can impede experiment creation and analysis.

## Types of Sampling

While web and other online experiments have an easy time collecting data, collecting data from traditional methods involving real populations is a much more difficult proposition. If you need to perform a survey of a population, it could be unreasonable in both time and money costs to try and collect thoughts from every single person in the population. This is where sampling comes in. The goal of sampling is to only take a subset of the population, using the responses from that subset to make an inference about the whole population. Here, we'll cover two basic probabilistic techniques that are commonly used.

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; your sample size depends on the level of uncertainty you are willing to have about the collected data. Since everyone has an equal chance of being drawn, we can expect the feature distribution of selected units to be similar to the distribution of the population as a whole. In addition, a simple random sample is easy to set up.

However, it is possible that certain groups are underrepresented in a simple random sample, especially those that make up a low proportion of the population. If there are certain rarer subgroups of interest, it can be worth adding one additional step and performing **stratified random sampling**. In a stratified random sample, we need to first divide the entire population into disjoint groups, or strata. That is, each individual must be a part of one group, and only one group. For example, you could divide people by gender (male, female, other), or age (e.g. 18-25, 26-35, etc.).

Then, from each group, you take a simple random sample. In a proportional sample, the sample size is proportional to how large the group is in the full population. For example, if you require 1000 data points, and stratified individuals of proportion {0.5, 0.3, 0.2}, then you would take 500 people from the first group, 300 from the second, and 200 from the third. This guarantees a certain level of knowledge from each subset, and theoretically a more representative overall inference on the population.

An alternative approach is to take a nonproportional sample from each group. For example, we could simply sample 500 people from each group. Computing the overall statistics in this case requires weighting each group separately, but this extra effort offers a higher understanding of each subgroup in a deeper investigation.

### Side Note: Non-Probabilistic Sampling

As noted at the start, the goal of sampling is to use a subset of the whole population to make inferences about the full population, so that we didn't need to record data from everyone. To that end, probabilistic sampling techniques were described above to try and obtain a sample that was representative of the whole. However, it's useful to note that there also exist non-probabilistic sampling techniques that simplify the sampling process, at the risk of harming the validity of your results. (We'll discuss the term 'validity' later in the lesson.)

For example, a convenience sample records information from readily available units. Studies performed in the social sciences at colleges often fall into this kind of sampling. The people participating in these tasks are often just college students, rather than representatives of the population at large. When performing inferences from this type of study, it's important to consider how well your results might apply to the population at large.

One notable example of a convenience sample resulting in a grave error comes from the prediction made by magazine "The Literary Digest" on the 1936 U.S. presidential election. While they predicted a healthy victory by candidate Alf Landon, the final result ended with a landslide victory by opposing candidate Franklin D. Roosevelt. This major error is attributed to their methods capturing a non-representative sample of the population, which included looking at the results of a mail-in survey from their magazine readers. Since the mail-ins were voluntary, and the magazine subscribers were already not well-representative of the general population, focusing on the people who returned surveys gave a large bias toward Landon.

## Measuring Outcomes

The goals of your study may not be the same as the way you evaluate the study's success. Perhaps this is because the goal is something that can't be measured directly. Let's say that you have an idea of a website addition that improves user satisfaction. How should we measure this? In order to evaluate whether or not this improvement has happened, you need to have a way to objectively measure the effect of the addition. For example, you might include a survey to random users to have them rate their website experience on a 1-10 scale. If the addition is helpful, then we should expect the average rating to be higher for those users who are given the addition, versus those who are not. The rating scale acts as a concrete way of measuring user satisfaction. These objective features by which you evaluate performance are known as **evaluation metrics**.

As a rule of thumb, 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. This is especially important if a metric isn't directly attached to the goal. For example, measuring students' confidence going into a standardized test might be a proxy for the goal of test preparedness, in the absence of being able to get their test scores directly or in a timely fashion.

Secondly, 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. From the above example, we aren't measuring confidence just to make people feel good about themselves: we're doing it to try and improve their actual performances.

### Side Note: Alternate Terminology

You might hear other terminology for goals and evaluation metrics than those used in this course. In the social sciences, it's common to hear a "construct" as analogous to the goal or objective under investigation, and the "operational definition" as the way outcomes are measured. For example, the construct of "reaction time" could be operationally defined as "time in milliseconds to click on the correctly indicated button."

In general company operations, you might encounter the terms "key results" (KRs) or "key performance indicators" (KPIs) as ways of measuring progress against quarterly or annual "objectives." These objectives and KRs / KPIs serve a similar purpose as study goals and evaluation metrics, and might even be driving factors in the creation of an experiment.

**Creating Metrics**

**Funnels**

There are additional concepts and terms that are commonly used for designing experiments, especially for web-based studies. 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. For example, we might think of the following steps for someone to purchase a product in an online store:

* Visit the site homepage
* Search for a desired product or click on a product category
* Click on a product image
* Add the product to the cart
* Check out and finalize purchase

One property to note about user funnels is that typically there will be some dropoff in the users that move from step to step. This is much like how an actual funnel narrows from a large opening to a small exit. Outside of an experiment, funnels can be used to analyze user flows. Observations from these flows can then be used to motivate experiments to try and improve the dropoff rates.

It's also worth noting that the flow through a funnel might be idealized compared to actual user practice. In the above example, users might perform multiple searches in a single session, or want to purchase multiple things. A user might access the site through a specific link, subverting the top part of the funnel. Refining the funnel and being specific about the kinds of events that are expected can help you create a consistent, reliable design and analysis.

**Unit of Diversion**

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 experimental group. The place in which you make this assignment is known as the **unit of diversion**. Depending on the type of experiment you have, you might have different options for diversion, each with its own pros and cons:

* Event-based diversion (e.g. pageview): Each time a user loads up the page of interest, the experimental condition is randomly rolled. Since this ignores previous visits, this can create an inconsistent experience, if the condition causes a user-visible change.
* Cookie-based diversion: A cookie is stored on the user's device, which determines their experimental condition as long as the cookie remains on the device. Cookies don't require a user to have an account or be logged in, but can be subverted through anonymous browsing or a user just clearing out cookies.
* Account-based diversion (e.g. User ID): User IDs are randomly divided into conditions. Account-based diversions are reliable, but requires users to *have* accounts and be logged in. This means that our pool of data might be limited in scope, and you'll need to consider the risks of using personally-identifiable information.

When it comes to selecting a unit of diversion, the consistency of the experience required can be a major factor to consider. For the example provided, we need something more consistent than pageview events. So we then consider the cookie-based diversion. If the differences in interface between control and experiment are fairly minor, then we're probably okay with cookie-based diversion. But if we think that users will notice the change and we believe that it will have a major effect on experience, then we might be inclined to choose an account-based diversion.

**Invariant and Evaluation Metrics**

A funnel will also be of benefit when it comes to deciding on metrics to track and analyze as part of the experiment. The immediate features that come out of a funnel come in the form of counts and ratios. For example, we could count the number of times a search results in a product being selected (a count), or the ratio of selections to searches as adjacent slices in the funnel (a ratio).

There are two major categories that we can consider features: as *evaluation metrics* or as *invariant metrics*. **Evaluation metrics** were mentioned in 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. We might want to see an increased click-through-rate from search results to products, or an increase in overall revenue. 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. For example, in an experiment with cookie-based diversion, the number of cookies generated for each condition would be a good invariant metric. Another metric could compare the distribution of times in which cookies were generated, to check the bias in the randomization procedure.

We're not limited to tracking just one metric of each type. It's not unusual to track multiple invariant metrics as checks on the system's integrity, or multiple evaluation metrics to check different potential facets of a manipulation's effects. Don't think that you need to track every possible metric, however. It's better to focus on a few key metrics, ignoring features that might be less reliable or highly correlated to other, more informative features. We'll discuss statistical considerations surrounding metrics in the next lesson.

## Controlling Variables

If we want to determine causality between two features, there are two main things to control. First of all, 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**. The correlation observed between two variables might be due to changes in a third variable, rather than one causing the other. Another possibility is that there is a causal relationship between the two features, but it is an indirect relationship mediated by a third, intermediate variable. This intermediate variable might be a larger driver of the changes in the output, with the manipulated variable only having a direct effect on the intermediate feature.

For the case where we see a relationship but don't perform a manipulation, we also need to be careful about the direction of effect. A relationship between variables "A" and "B" might be due to "A" having an effect on "B" or the reverse, "B" having an effect on "A". It might even be the case that "A" and "B" are related through some other function like a third variable.

### Additional Reference

Wikipedia: [Correlation does not imply causation](https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation) - Reference page with examples of how an observed correlation between two features might come about.

**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
* Internal Validity
* External Validity

**Construct Validity**

Construct validity is tied to the earlier discussion of how well one's goals are aligned to the evaluation metrics used to evaluate it. Poor construct validity can come about when an evaluation metric does not actually measure something related to the desired outcome concept. Alternatively, it might be that a metric is ill-constructed, such that it does not make clear distinctions on the outcome concept.

**Internal Validity**

Internal validity refers to the degree to which a causal relationship can be derived from an experiment's results. Controlling for and accounting for other variables is key to maintaining good internal validity. The previous page on controlling variables shows ways in which internal validity might not be met.

**External Validity**

External validity is concerned with the ability of an experimental outcome to be generalized to a broader population. This is most relevant with experiments that involve sampling: how representative is the sample to the whole? For studies at academic institutions, a frequent question is if data collected using only college students can be generalized to other age or socioeconomic groups.

## Checking Bias

Biases in experiments are systematic effects that interfere with the interpretation of experimental results, mostly in terms of internal validity. Just as humans can have a lot of different [biases](https://en.wikipedia.org/wiki/List_of_cognitive_biases), there are numerous ways in which an experiment can become unbalanced.

### Sampling Bias

Many experimental biases fall under the sampling bias umbrella. Sampling biases are those that cause our observations to not be representative of the population. For example, if assignment to experimental groups is done in an arbitrary fashion (as opposed to random assignment or matched groups), we risk our outcomes being based less on the experimental manipulation and more on the composition of the underlying groups.

Studies that use surveys to collect data often have to deal with the **self-selection bias**. The types of people that respond to a survey might be qualitatively very different from those that do not. A straight average of responses would not necessarily reflect the feelings of the full population; weighting responses based on the differences between the observed responses and properties of the target population may be needed to come to reasonable conclusions.

One type of sampling bias related to missing data is the **survivor bias**. Survivor bias is one where losses or dropout of observed units is not accounted for in an analysis. A key example of this was in British World War II operations research, where engineers avoided using survivor bias when they considered where to add armor to their planes. Rather than add armor to the spots where returning planes had bullet holes, armor was added to the spots where the planes didn't have bullet holes. That's because the planes that took shots to those places probably crashed, due to those locations being more vital for maintaining flight, so they didn't "survive" and weren't available for observation.

### Novelty Bias

A novelty effect 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 a manipulation until after the novelty wears off and population metrics return to a level that actually reflects the changes made. This will be important for cases where we want to track changes over time, such as trying to get users to re-visit a webpage or use an app more frequently. Novelty is probably not a concern (or perhaps what we hope for) when it comes to manipulations that are expected to only have a one-shot effect.

### Order Biases

There are a couple of biases to be aware of when running a within-subjects experiment. Recall that in a within-subjects design, each participant performs a task or makes a rating in multiple experimental conditions, rather than just one. 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.

An easy way of getting around order biases is to simply randomize the order of conditions. If we have three conditions, then each of the six ways of completing the task (ABC, ACB, BAC, BCA, CAB, CBA) should be equally likely. While there still might end up being order effects like carry-over effects, where a particular condition continues to have an effect on future conditions, this will be much easier to detect than if every participant completed the task in the exact same order of conditions.

### Experimenter Bias

One bias to watch out for, especially in face-to-face experiments, is the experimenter bias. This 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. In addition, participants may act differently in the presence of an experimenter, to try and act in the 'right' way – regardless of if a subject actually knows what the experimenter is looking for or not.

This is where design steps like blinding are important. In [blinding](https://en.wikipedia.org/wiki/Blinded_experiment), the administrator of a procedure or the participant do not know the condition being used, to avoid that subconscious bias from having an effect. In particular, the **double-blind** design hides condition information from both the administratorand participant in order to have a strong rein on experimenter-based biases.

**Ethics in Experimentation**

Before you run an experiment, it's important to consider the ethical treatments to which you subject your participants. Through the mid-20th century, exposure of questionable and clearly unethical research in the social and medical sciences spurred the creation of guidelines for ethical treatment of human subjects in studies and experiments. 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. Risk of harm isn't just in the physical sense, but also the mental sense. If an experimental condition has potential to negatively affect a participant's emotions or mentality, then it's worth thinking about if the risks are really necessary to perform the desired investigation.
* **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. However, if the benefits are shown to be worth the risks, then it is still possible for the study to be run. This often comes up in medicine, where test treatments should show worthy potential against alternative approaches.
* **Provide informed consent**: Building up somewhat from the previous two points, subjects should be informed of and agree to the risks and benefits of participation before they join the study or experiment. This is also an opportunity for a participant to opt out of participation. However, there are some cases where deception is necessary. This might be to avoid biasing the participant's behavior by seeding their expectations, or if there is a dummy task surrounding the actual test to be performed. 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**: If you're dealing with identifiable information in your study, 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.

In the formal sciences, an experiment proposal must go through a review board before it can be run, to ensure that ethical principles have been followed. It's likely that you won't have a review board to submit your designs to prior to running an experiment. You'll need to evaluate these principles for yourself or with your colleagues to check for potential ethical issues before going forward with a study design.

One particular point worth further discussion is that of informed consent for web-based experiments. It's often the case that when an experiment is run, users who are included in an experiment often don't know that they're participating in an experiment. If a manipulation carries no risk and is so minor as to be hidden away from the user (e.g. a change in recommendation engine), perhaps there is no need for informed consent. And when it comes to bias, it's known that peoples' behaviors can change when they know they are under observation. In practice, informed consent is often not considered when performing a web experiment.

However, informed consent is still an important ethical principle, so there is continuing debate on how to best obtain consent for users of a website. One option could be to allow users to opt out of experiment participation, with the default user agreement implying consent to participation in unobtrusive experiments. An opposing option would only run experiments on users who opt-\_in\_ to participation, asking the user to set their preference on their initial visit or registration. The opt-in approach is more in line with the core idea of informed consent, but also risks fewer users available for testing changes.

**Examples in Experimental Ethics**

Here are three studies in the social and medical sciences that are often brought up as examples of violations of experimental ethics and progenitors of the movement to establish ethics guidelines and boards:

* [Tuskegee Syphilis Study](https://en.wikipedia.org/wiki/Tuskegee_syphilis_experiment): This study was started in 1932, where hundreds of African-American men were tracked over the course of up to forty years to study the natural progression of syphilis. The subjects were denied treatment, even after the development of effective syphilis treatments like penicillin, and there were active steps taken to hide the truth of their conditions and treatments to the subjects.
* [Milgram obedience study](https://en.wikipedia.org/wiki/Milgram_experiment): During the mid 1960s, psychologist Stanley Milgram tested to what degree people follow authority figures. Participants were asked to administer gradually increasing shocks to an acting confederate participant at the behest of a lead experimenter based on mistakes made on a dummy memory test used as a cover story. While no shocks were *actually*administered, the study did bing forth questions on what constitutes adequate debriefing and what level of information and informed consent needs to be provided to a participant in a study that includes necessary deceptive elements.
* [Stanford Prison Experiment](https://en.wikipedia.org/wiki/Stanford_prison_experiment): This 1971 study conducted by psychologist Philip Zimbardo was built to test the dynamics and effects of power differences, using a prison scenario. Study volunteers were divided into prisoner and guard groups and their dynamics observed; the study had to be ended after less than a week due to the increasingly harsh conditions the 'guards' had settled into treating the 'prisoners'. It has since served as a major ground for ethical criticisms, violating now-established guidelines around risks to participants and clarity of purpose. There are also methodological criticisms, as experimenter biases may have shaped the behavior of the 'guards' group in their interactions with 'prisoners'.

And here's a [Techcrunch article](https://techcrunch.com/2014/06/29/ethics-in-a-data-driven-world/" \t "_blank) discussing the ethics of the 2014-published Facebook study on the impact of changing the affect of posts seen on users' feeds to their own posting habits.

**Additional Resources**

For further reading, here are a few links to documents both historical and current on how to conduct experiments involving humans:

* [Nuremberg Code](https://history.nih.gov/research/downloads/nuremberg.pdf) - Ten principles for human subjects research stemming from the Nuremberg Trials published in 1949.
* [The Belmont Report](https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html) - 1979 report laying out general principles for research involving human subjects
* [APA Ethics Code](https://www.apa.org/ethics/code/) - Guidelines published by the American Psychological Association for general psychological practices (not just research)
* [BPS Code for Human Research Ethics](https://www.bps.org.uk/news-and-policy/bps-code-human-research-ethics-2nd-edition-2014) - Code for human-subjects research published by the British Psychological Society

**SMART Mnemonic for Experiment Design**

There's a mnemonic called **SMART** for teams to plan out projects that also happens to apply pretty well for creating experiments. The letters of SMART stand for:

* **S**pecific: Make sure the goals of your experiment are specific.
* **M**easurable: Outcomes must be measurable using objective metrics
* **A**chievable: The steps taken for the experiment and the goals must be realistic.
* **R**elevant: The experiment needs to have purpose behind it.
* **T**imely: Results must be obtainable in a reasonable time frame.

There are also other words possible for the mnemonic, such as **A**ctionable and **R**ealistic. Ultimately, however, the message is pretty much the same, with the dimensions switched around a bit. Note, however, that the mnemonic doesn't cover certain considerations important to experiment design. Considerations of ethical issues or bias will need to be considered separately, so don't just take the mnemonic as the final judge of whether your experiment is ready to proceed!