Scientific Methodology

Designing valid scientific investigations

Valid research in different fields of science

Scientific journals publish studies that involve a wide range of different activities. There are common elements to most scientific studies such as:

- a commitment to deductive testing (i.e. the idea that experimental/observational evidence determines if a theory can be accepted)
- an experimental design that protects against bias
- a consideration of alternative explanations of the results
- the interpretation of data (either qualitative or quantitative) to produce results that are reproducible and generalisable
- a discussion of how the research relates to other work done in that field

[Elements of the above adapted from " Research Design: Qualitative, Quantitative, and Mixed Methods Approaches" 5th Edition, Kindle Edition by John W. Creswell, J. David Creswell, pg. 24]

However, as there are so many different activities that come under the general banner of science, it is not really possible to give a universally applicable prescription for how a scientific study should be done. As we will discover in the topic on the philosophy of science, even defining what is science - and is not - is challenging.

Having said that, within specific fields there are often well established patterns and techniques for conducting research. These patterns have been developed in an attempt to make valid conclusions about the nature of the world around us, while working within the constraints that apply in a particular field. Typical considerations for the design of experiments include:

- The experiment needs to be feasible given constraints on the available time and money.

- As far as possible, biases should be anticipated and avoided or compensated for. Bias can take many forms e.g.
    - Experimenter bias, where the person or persons conducting the experiment consciously or unconsciously prefers a certain result.
    - Selection bias, where the things being measured are not representative of the system being investigated.
    - Or numerous others, see e.g. https://en.wikipedia.org/wiki/List_of_cognitive_biases

- When verifying or investigating causal relationships between variables, alternative explanations should be considered and checked. Is it really the case that A causes B, or might it be that there is some other factor C that causes both A and B ?

- Errors due to the equipment used in the experiment must be compensated for or at least estimated and acknowledged. On a simple level, an instrument such as a thermometer has limited accuracy and may have a systematic error in a particular direction. More complex instruments such as cameras can record quite different images depending on factors such as lighting and exposure time.

We will begin with an overview of the “scientific method” or “fair-testing” which is common in disciplines such as medicine and psychology. These fields often attempt to measure subtle effects that are influenced by numerous factors outside the scientists ability to control or even measure. There is a large body of literature on the design of experiments in these fields.

Fields such as physics and chemistry typically face quite different issues, such as systematic errors produced by complex equipment. Because the issues encountered in experimental design in these fields are frequently quite specific to the particular research being conducted, there is much less literature on generic experimental design in these fields (but very large amounts of literature on the design of specific experiments or use of specific pieces of equipment). We will examine the approach taken in several research papers to see some of the kinds of issues that are encountered and how they might be dealt with.

Over the past few decades there has been a rise in interdisciplinary research that draws on many different areas. For instance climate science involves physics, chemistry, paleontology, ecology and more. Interdisciplinary research draws on many different approaches to experimental design and to science in general.

“The Scientific Method”

The media and some science textbooks often present what can be called the “scientific method”. The general approach is:

1. Ask a question.
2. Form a hypothesis (a potential answer to the question)
3. Design an experiment to test the hypothesis:
    a. Form a control group and a test group
    b. Define two variables, the independent variable which is manipulated by the experimenter and the dependent variable which the experimenter measures. Keep all variables other than the independent variable the same between the control and test group.
4. Record results for the dependent variable. Examine the results to see if they support the hypothesis or not. This may require statistical methods, e.g. p values.

There is a very large body of scientific literature describing experiments conducted in this way, or in a similar way. However we should note from the outset that this approach does not represent all or even most of what is normally considered science. For example very little work done prior to the 20th century fits this mold.

A Brief History of Controlled Trials

Origins

The general idea of performing an experiment to test a hypothesis has been around for a long time. One of the earliest written descriptions of such an approach is found in the Bible. Daniel was a jew living under King Nebuchadnezzar of Babylon, around 500 BC. He was ordered to eat food provided by the King. From Daniel chapter 1:

"But Daniel resolved not to defile himself with the royal food and wine, and he asked the chief official for permission not to defile himself this way. [9] Now God had caused the official to show favor and compassion to Daniel, [10] but the official told Daniel, "I am afraid of my lord the king, who has assigned your[a] food and drink. Why should he see you looking worse than the other young men your age? The king would then have my head because of you."
[11] Daniel then said to the guard whom the chief official had appointed over Daniel, Hananiah, Mishael and Azariah, [12] "Please test your servants for ten days: Give us nothing but vegetables to eat and water to drink. [13] Then compare our appearance with that of the young men who eat the royal food, and treat your servants in accordance with what you see." [14] So he agreed to this and tested them for ten days.
[15] At the end of the ten days they looked healthier and better nourished than any of the young men who ate the royal food. [16] So the guard took away their choice food and the wine they were to drink and gave them vegetables instead."

The First Modern Controlled Trial

There are a number of ad-hoc descriptions of controlled trials over the next 2000 years. One of the earliest widely recognized trials that included many of the elements of a modern controlled trial was conducted by a ships doctor, James Lind, in 1747, to investigate treatments for Scurvy. Scurvy is a potentially fatal disease caused by a lack of vitamin C, and was very common on long voyages where access to fresh food was very limited.
The following extract is taken from *Perspect Clin Res. 2010 Jan-Mar; 1(1): 6–10.*
http://www.jameslindlibrary.org/

James Lind is considered the first physician to have conducted a controlled clinical trial of the modern era.[1-4] Dr Lind (1716-94),

whilst working as a surgeon on a ship, was appalled by the high mortality of scurvy amongst the sailors. He planned a comparative trial of the most promising cure for scurvy.[1-4] His vivid description of the trial covers the essential elements of a controlled trial. Lind describes "On the 20th of May 1747, I selected twelve patients in the scurvy, on board the Salisbury at sea. Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of the knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet common to all, viz. water gruel sweetened with sugar in the morning; fresh mutton-broth often times for dinner; at other times light puddings, boiled biscuit with sugar, etc., and for supper, barley and raisins, rice and currants, sago and wine or the like. Two were ordered each a quart of cyder(sic) a day. Two others took twenty-five drops of elixir vitriol three times a day … Two others took two spoonfuls of vinegar three times a day … Two of the worst patients were put on a course of sea-water … Two others had each two oranges and one lemon given them every day … The two remaining patients, took … an electary recommended by a hospital surgeon … The consequence was, that the most sudden and visible good effects were perceived from the use of oranges and lemons; one of those who had taken them, being at the end of six days fit for duty … The other was the best recovered of any in his condition; and … was appointed to attend the rest of the sick. Next to the oranges, I thought the cyder had the best effects …" (Dr James Lind's "Treatise on Scurvy" published in Edinburgh in 1753)

Although the results were clear, Lind hesitated to recommend the use of oranges and lemons because they were too expensive.[3] It was nearly 50 years before the British Navy eventually made lemon juice a compulsory part of the seafarer's diet, and this was soon replaced by lime juice because it was cheaper.

Lind's Treatise of 1753, was written while he was resident in Edinburgh and a Fellow of the Royal College of Physicians, contains not only his well known description of a controlled trial showing that oranges and lemons were dramatically better than the other treatments for the disease, but also a systematic review of previous literature on scurvy.[5]


A mild case of Scurvy.

Statistical Analysis and Experimental Design

Experimental design and the statistical analysis of experiments was put on a firm footing by the statistician R.A. Fisher, with his books "*Statistical Methods for Research Workers*" (published 1925) and "*The Design of Experiments*" (published 1935).

These books covered issues relating to the design and analysis of experiments such as the number of measurements required to be sure of a result to a specified level of confidence.

Modern Controlled Trials

Although the general concept of a controlled trial may seem straightforward, there are numerous ways in which experiments can lead to incorrect conclusions. Well done modern controlled trials go to elaborate lengths to avoid these issues.

Medicine in particular provides numerous opportunities to generate incorrect conclusions from a controlled trial. This paper, "A manifesto for reproducible science" (published in Nature: Human Behaviour https://www.nature.com/articles/s41562-016-0021#t1) describes ways that many of these biases can be removed or reduced for studies with many confounding variables.

Exercise :
Suppose we want to know if a certain drug treats a disease. We find people with the disease, treat some of them with the drug and see if they get better faster than the people we didn't treat.

What are some ways in which bias or methodological flaws could impact this validity of this experiment ?

Potential bias or design issue	What could be done about it ?
In both groups, some people got better and some didn't. How can we know if the difference between the number of people in each group who recovered is really due to	

the drug or just “luck”?	
It is hard to measure if a trial participant actually got better. Maybe the way we measure will give different results ? E.g. DEXA vs QUS for measuring osteoporosis.	
People may say (or feel obliged to say) they are better when they are not. If they feel they are getting treated then may feel better because of their belief in the treatment. Conversely, participants may be upset at being assigned to the control group (and thus not treated for the illness) and might be motivated to lie about how they feel.	
It might not be the drug that is making the participants better. Maybe just the extra attention from the doctor makes them feel better ? Or something else that is special about the treatment group ?	
The participants assigned to the control group might be sicker or healthier than the people assigned to the test group. This might skew the results to suggest that the drug does (or doesn’t) work.	
The participants in the trial may not be representative of the general population. For instance, many experiments are done on 20 year old university psychology students because it is easy to find them and persuade them to participate in exchange for credit in their degree.	

Suppose we use several different ways of measuring how much better people get ? What if some measurements show the control group get better and some show the treatment group get better ?	
What if our study shows the drug doesn't work, so we don't publish it ?	
What if our study shows the drug works, but many other studies come to the opposite conclusion ?	
What if the company that makes the drug is paying us (and wants positive results) ?	

Some specific references :

The memory of water - Dayenas E, Beauvais F, Amara J, Oberbaum M, Robinzon B, Miadonna A, Tedeschit A, Pomeranz B, Fortner P, Belon P, Sainte-Laudy J, Poitevin B, Benveniste J (30 June 1988). "Human basophil degranulation triggered by very dilute antiserum against IgE" (PDF). *Nature*. **333** (6176): 816–818. Bibcode:1988Natur.333..816D.
PLoS One. 2015; 10(3): e0122800.

Placebo Effect in relation to surgery-
Machado, G. C., Ferreira, P. H., Harris, I. A., Pinheiro, M. B., Koes, B. W., van Tulder, M., Rzewuska, M., Maher, C. G., … Ferreira, M. L. (2015). Effectiveness of surgery for lumbar spinal stenosis: a systematic review and meta-analysis. *PloS one*, *10*(3), e0122800. doi:10.1371/journal.pone.0122800

The Salk Polio Vaccine (one of the early very large scale randomised double blind trials)
Dawson, Liza, Clinical Trials 2004: 1, pages 122:130 "The Salk Polio vaccine of 1954 - risks, randomisation, and public involvement in research"

Studies that can be described as "large scale double blind randomised placebo controlled trial" meet all the above requirements for eliminating bias, and are the "Gold standard" in clinical trials.

Questions to ask of our “case study” papers

- What has been investigated and why is this interesting and/or significant?
- Does this research fit easily into the framework of types of scientific research we have explored? I.e. qualitative vs quantitative vs mixed methods research? Is it useful to think of it an example of fair-testing, pattern-seeking, exploring, classifying and identifying, making things or developing systems or investigating models? (the categorisation provided in https://14254.stem.org.uk/Beyond_Fair_Testing.pdf)
- What methodology has been used by the authors to obtain valid and reliable data? (*This is our inquiry question for this section*!). Specifically:
    - how has bias been addressed and minimised?
    - what process has been followed to gather data? (have the techniques of remote sensing or streaming of data been used?)
    - what techniques have been used to minimise and measure uncertainty in the data?
    - how has the data been analysed?
    - have alternative explanations of the results been considered?
    - do the results have some level of general applicability?
    - has the research been reported in a way that it allow other scientists to replicate the results?

Some suggested answers to the exercise on potential bias in a medical study

Potential problem	What could be done about it ?
In both groups, some people got better and some didn't. How to know if the difference between the number of people in each group who recovered is real or just luck ?	The larger the sample size, the greater the confidence in the results. The level of confidence can be quantified using statistical tests.
It is hard to measure if a trial participant actually got better. Maybe the way we measure will give different results ? E.g. DEXA vs QUS for measuring osteoporosis.	Use a measurement that has been validated by other researchers. Validate against objective measures. E.g. blood tests, rather than patients' reported feeling of wellness (if this is possible).
People may feel obliged to say they are better when they are not. If they feel they are getting treated then may feel better because of the patient's belief in the treatment. Conversely, participants may be upset at being assigned to the control group (and thus not treated for the illness) and might be motivated to lie about how they feel.	Don't tell the participants whether they are in the control group or the treatment group. That way both groups will have the same average bias. I.e. make the participants "blind" to which group they are in. Use a placebo treatment for the control group which looks identical to the real treatment so the control group cannot detect that they are in the control group
It might not be the drug that is making the participants better. Maybe just the extra attention from the doctor makes them feel better ? Or something else that is special about the treatment group ?	Do not tell the medical staff which people are in which group, so everyone is treated the same way. I.e. make the people running the trial "blind"
The participants assigned to the control group might be sicker or healthier than the people assigned to the test group.	Assign people to the control group or the treatment group in a random way. Try to assign people in a way that has similar people in each group. Use statistical methods that take into account variability within a group (e.g. estimating the variance for each group).
The participants in the trial may not be representative of the general population. For instance, many	Ideally, recruit people to the trial using a random selection (or recruit people who match the type of people that are of interest for the treatment).

experiments are done on 20 year old university student because it is easy to find them and persuade them to participate.	
Suppose we use several different ways of measuring how much better people get ? What if some measurements show the control group get better and some show the treatment group get better ?	Report all your results - the good and the bad. Statistical tests must take into account the number of ways in which you measured the participants. If you measure 20 different variables, it is likely that one of them will have a p value < 0.05.
What if our study shows the drug doesn't work, and so we don't publish it ?	Negative results should be published. When assessing if a drug works, all studies should be considered - say there are 20 studies that show no effect and one that shows an effect, then an unbiased observer would suspect that the study showing an effect was the result of chance or some error. But if only the positive study is published, one would conclude that the drug works. Some funding bodies now require that studies are registered on a public database before they start, to avoid non-publication of negative results.
What if our study shows the drug works, but many other studies come to the opposite conclusion ?	All available evidence should be considered to determine if the result is plausible.
What if the company that makes the drug is paying us (and wants positive results) ?	Bias in favour of the company supplying the money is a real problem. Conflicts of interest should be listed in the study. In some situations, external review committees can assist with the study design to reduce this type of bias. (See: https://www.nature.com/articles/s41562-016-0021#t1)

Brain Drain paper:

- *What has been investigated and why is this interesting and/or significant?*
  Does the presence of your smartphone reduce your cognitive performance? This is interesting and significant as most people use a smartphone - it might impact their ability to perform their job, or learn things at school to the best of their ability

- *Does this research fit easily into the framework of types of scientific research we have explored?*
  Mixed methods (contains numerical measurements of students' performance as well as interviews of how they feel about their smartphone). Independent variable - location of smartphone

- *What methodology has been used by the authors to obtain valid and reliable data? (This is our inquiry question for this section!). Specifically:*
  - *how has bias been addressed and minimised?*
    Participants were randomly assigned to groups in the study (but didn't use a cross-section of the population - results may have limited applicability to people who are not college students
    Duplicate results have been removed, students who did not own phones were excluded, students who didn't follow instructions or had an excessive error rate on tests were excluded
    The authors checked if there was any effect from which lab assistant helped.

  - *what process has been followed to gather data? (have the techniques of remote sensing or streaming of data been used?)*
    - Testing of subjects under controlled conditions to gather quantitative data
    - Questionaire (qualitative data collection)

  - *what techniques have been used to minimise and measure uncertainty in the data?*
    Many participants were used (~500 in first experiment and ~300 in second experiment)
    All smartphones were switched off to control for any noise made by some smartphones and not others.

  - *how has the data been analysed?*
    Using a variety of statistical tests (e.g. MANOVA and ANOVA)

  - *have alternative explanations of the results been considered?*
    The authors checked whether the impact on attention could have been due to participants surreptitiously checking their phones in experiment 1 (as the phones

were silenced (but not turned off), by having participants randomly assigned to two groups in experiments 2 - one group with their phone on silent, and one group where their phone was switched off - and checked they saw the same effect.
The authors also addressed previous research that suggested that being separated from your phone reduces performance due to increased anxiety, by pointing out that in the other study, participants were forced to listen to their phones ringing without being able to answer them.

- *do the results have some level of general applicability?*
  The authors assert general applicability, but the fact their study was limited to college students leaves questions about the true generalisability of the results. The results are most likely generally applicable to young adults at least.

- *has the research been reported in a way that it allow other scientists to replicate the results?*
  The authors are careful to explain what the experimental conditions were, and what tests were used (giving references for these).