Evidence-based Decision Making
Counterfactuals: Experiments

Rui Mata, FS 2026

Version: March 30, 2026

Universität Basel

Bernoulli Network for the Behavioral Sciences

The Bernoulli Network for the Behavioral Sciences
invites you to the research talk

Deliberate Ignorance: Why We Choose Not to Know

Prof. Dr. Ralph Hertwig

Director, Center for Adaptive Rationality
Max Planck Institute for Human Development, Berlin

Tuesday, 14 April 2026 | 18:00

Maurice Müller Hörsaal, Biozentrum
No registration required

Das Interview

Zum Sehen, Lesen oder Anhören

Das ganze Gespräch als Video-Podcast

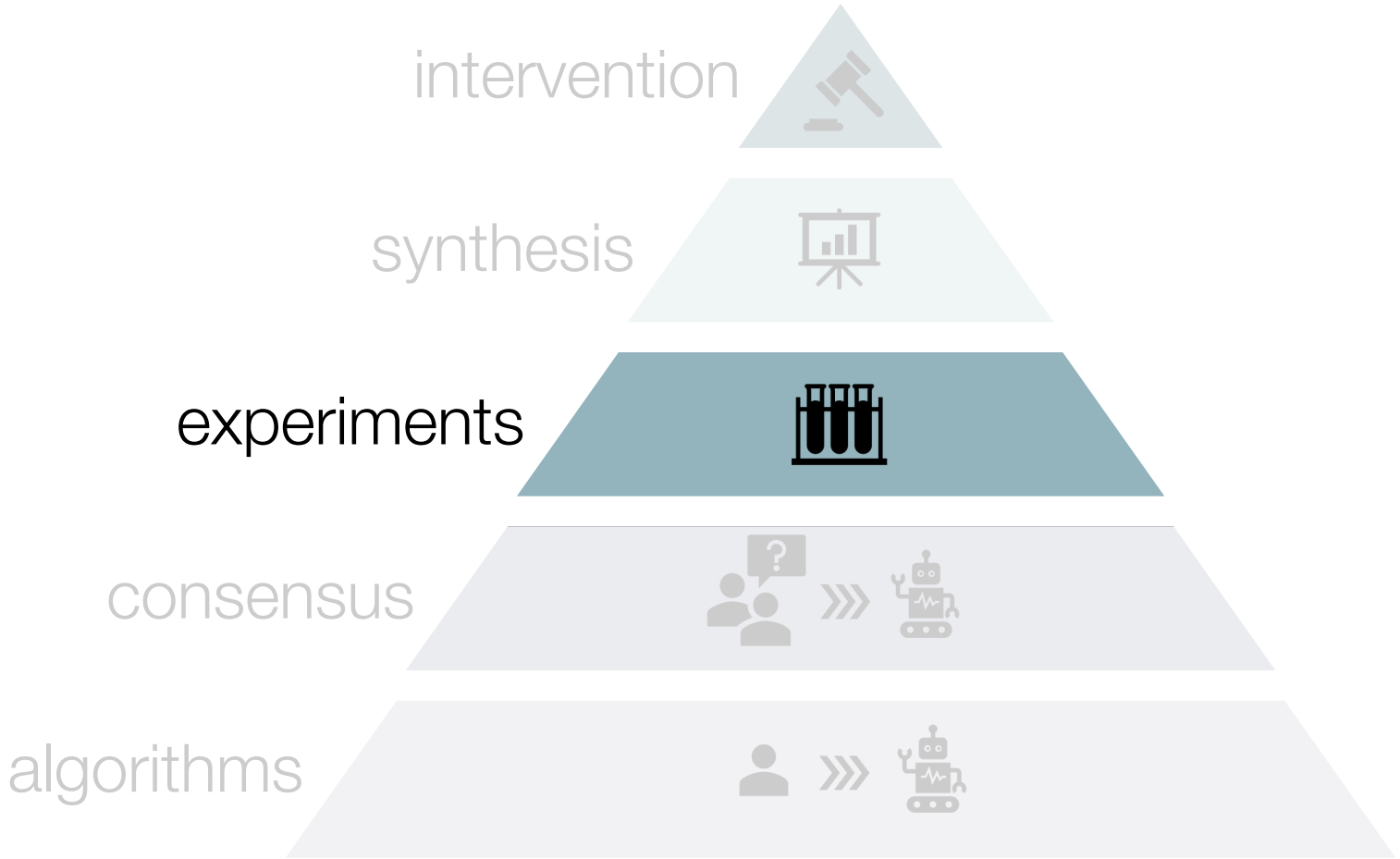
Sehen Sie das Interview auf Spotify

Zerstört das Internet die Demokratie, Ralph Hertwig?

Nur eine Frage

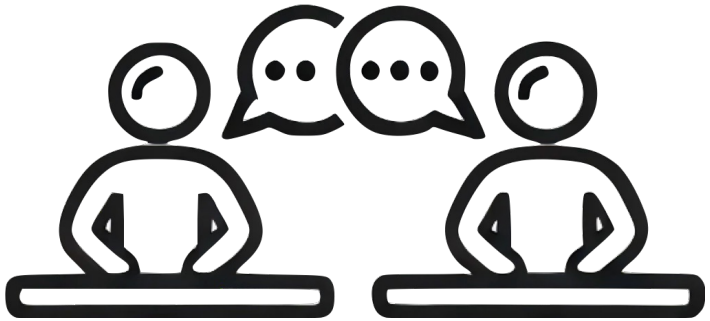
https://www.zeit.de/wissen/2026-01/ralph-hertwig-internet-demokratie-psychologie

Climbing the pyramid of evidence



WHY DO WE NEED EXPERIMENTS?

And when can we (not) use them?



Goals for today

- Understand the nature of causal inference as the comparison of treatment to some counterfactual
- Understand that experiments, and in particular RCTs, have desirable properties for causal inference – but also have limitations…
- Consider alternatives to RCTs to establish the counterfactual

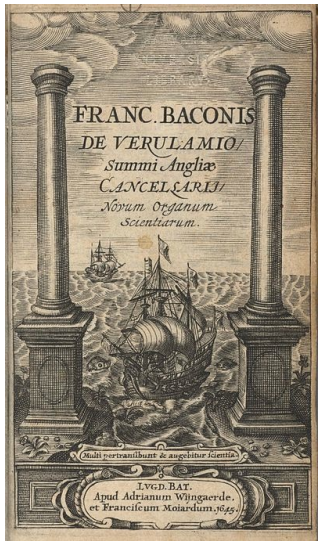
Causality

> : a causal quality or agency
>
> : the relation between a cause and its effect or between regulatory correlated events or phenomena

https://www.merriam-webster.com/dictionary/causality

Science as causal inference

Francis Bacon
(1561-1626)



1620

CrossMark

Causal inference in economics and marketing

Hal R. Varian[a,1]

[a]Economics Team, Google, Inc., Mountain View, CA 94043

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 25, 2016 (received for review May 28, 2015)

This is an elementary introduction to causal inference in economics written for readers familiar with machine learning methods. The critical step in any causal analysis is estimating the counterfactual—a prediction of what would have happened in the absence of the treatment. The powerful techniques used in machine learning may be useful for developing better estimates of the counterfactual, potentially improving causal inference.

causal inference | economics | machine learning | marketing

Suppose you are given some data on ad spend and product sales in various cities and are asked to predict how sales would respond to a contemplated change in ad spend. If y_c denotes per capita sales in city c and x_c denotes per capita ad spend in city c, it is tempting to run a regression of the form $y_c = bx_c + e_c$, where e_c is an error term and b is the coefficient of interest. (We assume all data have been centered; therefore, we can ignore the constant in the regression.) The machine-learning textbook by James et al. describes a problem of this sort (ref. 1, p. 59).

Unfortunately, such a regression is unlikely to provide a satisfactory estimate of the "causal" effect of ad spend on sales. To see why, suppose that the sales, y_c, are per capita box office receipts for a movie about surfing and x_c are per capita television ads for that movie. There are only two cities in the dataset: Honolulu, Hawaii and Fargo, North Dakota.

Suppose that the dataset indicate that the advertiser spent 10 cents per capita on television advertising in Fargo and observed $1 in sales per capita, whereas in Honolulu, the advertiser spent $1 per capita and observed $10 in sales per capita. Hence, the model $y_c = 10x_c$ fits the data perfectly.

However, here is the critical question: Do you really believe that increasing per capita spend in Fargo to $1 would result in box office sales of $10 per capita? For a surfing movie? This outcome seems unlikely, so what is wrong with our regression model?

A Motivating Problem

The problem is that there is an omitted variable in our regression, which we may call "interest in surfing." Interest in surfing is high in Honolulu and low in Fargo. What is more, the marketing executives that determine ad spend presumably know this, and they choose to advertise more where interest is high and less where it is low. Therefore, this omitted variable—interest in surfing—affects both y_c and x_c. Such a variable is called a "confounding variable."

To express this point mathematically, think of (y, x, e) as being the population analogs of the sample (y_c, x_c, e_c). The regression coefficient is given by $b = \text{cov}(x, y)/\text{cov}(x, x)$. Substituting $y = bx + e$, we have

$$b = \text{cov}(x, xb + e)/\text{cov}(x, x) = b + \text{cov}(x, e)/\text{cov}(x, x).$$

The regression coefficient will be unbiased when $\text{cov}(x, e) = 0$.*

If we are primarily interested in predicting sales as a function of spend, and the advertiser's behavior remains constant, the simple regression described in ref. 1 may be just fine. However, usually a prediction of past behavior is not the goal; what we want to know is how box office receipts would respond to a change in the advertiser's behavior.

To put it slightly more formally: we have historical observations that were generated by a process such as "choose spend based on factors you think are important," and we want to predict what would happen if we switch to a data generating process such as "increase your spend everywhere by some amount."

It is important to understand that the problem is not simply that there is a missing variable in the regression. There are always missing variables—that is what the error term represents. The problem is that the missing variable, "interest in surfing," affects both the outcome (sales) and the predictor (ads); therefore, the simple regression of sales on ads will not give us a good estimate of the causal effect: what would happen to sales if we explicitly intervened and changed ad expenditure across the board.

This problem comes up all of the time in statistical analysis of human behavior. In our example, the amount of advertising in a city, x_c, is chosen by some decision makers who likely have some views about how various factors affect outcomes, y_c. However, the analyst is not able to observe these factors—they are part of the error term, e_c. It is therefore unlikely that x_c and e_c are uncorrelated. In our example, cities with high interest in surfing may have high ad expenditure and high box office receipts, meaning a simple regression of y_c on x_c would overestimate the effect of ad expenditure on sales.

In this simple example, we have described a particular confounding variable. However, in realistic cases, there will be many confounding variables—variables that affect both the outcome and the variables we are contemplating changing.

Everyone knows that adding an extra predictor to a regression will typically change the values of the estimated coefficients on the other predictors because the relevant predictors are generally correlated with each other. Despite this well-known phenomenon, many analysts seem comfortable in assuming that the predictors we do not observe—those in the error term—are magically orthogonal to the predictors we do observe.

The "ideal" data, from the viewpoint of the analyst, would be data from an incompetent advertiser who allocated expenditures randomly across cities. If ad expenditure is truly random, then we do not have to worry about confounding variables because the predictors will automatically be orthogonal to the error term. However, statisticians are seldom lucky enough to have a totally incompetent client.

There are many other examples of confounding variables in economics. Here are a few classic examples.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Drawing Causal Inference from Big Data," held March 26–27, 2015, at the National Academies of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Big-data.

Author contributions: H.R.V. wrote the paper.

Conflict of interest statement: H.R.V. is a full-time employee of Google, a private company.

This article is a PNAS Direct Submission.

[1]Email: halhbms.berkeley.edu.

*Note that the problem is not inherently statistical in nature. Suppose that there is no error term, so that the model "revenue = spend + interest in surfing" fits exactly. If we only look at the variation in spend and ignore the variation in surfing interest, we get a misleading estimate of the relationship between spend and revenue.

†It would not have to be that way. Perhaps surfing is so popular in Honolulu that everyone already knows about the movie, and it is pointless to advertise it. Again, this is the sort of thing the advertiser might know but the analyst does not.

www.pnas.org/cgi/doi/10.1073/pnas.1510479113

Downloaded from https://www.pnas.org by IBIOZENTRUM A 4710545 on February 4, 2024 from IP address 131.152.36.128.

Baconian empiricism / method:

Sir Francis Bacon (knighted in 1603) was a strong advocate for **observation, experimentation, and inductive reasoning** based on experimental data. He believed that instead of relying on traditional authorities or pure logic (as in Aristotelian thinking), science should be built on careful observation of nature and the **gradual accumulation of knowledge** through methodical experiments. He introduced **"tables of discovery"** to organize experiments—focusing on when phenomena are present, absent, or vary in degree—to uncover the true causes and underlying principles of nature (e.g., heat → friction; social media use → loneliness).

"The critical step in any causal analysis is estimating the counterfactual—a prediction of what would have happened in the absence of the treatment."

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences of the United States of America, 113*(27), 7310–7315. http://doi.org/10.1073/pnas.1510479113

Counterfactuals

A counterfactual is a conditional statement exploring what would be the case if a certain event or condition were true

Example:

"D shoots at V, but only grazes him, leaving V with a slightly bleeding flesh wound. X then comes along and shoots V through the heart, killing him instantly. D's act is clearly not a "cause in fact" of V's death, since V would have died, and in just the manner he did, even if D had not shot him."

- **Singular judgment of causation**: a single cause (e.g., A) is necessary and sufficient for effect (Y) to occur
- In reality, we find **conjunctive plurality of causes** (A&B&C → Y), **disjunctive plurality of causes** (A|B|C → Y)
- Complex regularities (e.g., A&B&C → Y or A|B|C → Y) are rarely (if ever) fully known, thus we **formulate propositions which entail the probability** of a variable being causally connected with an effect

Marini, M. M., & Singer, B. (1988). Causality in the social sciences. *Sociological methodology, 18*, 347-409.

The gold standard: Randomized controlled trials (RCTs)

Consolidated Standards of Reporting Trials

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology and pharmacotherapeutics, 1*(2), 100-107.

The gold standard…

Experiments/Randomised control trials (RCT)

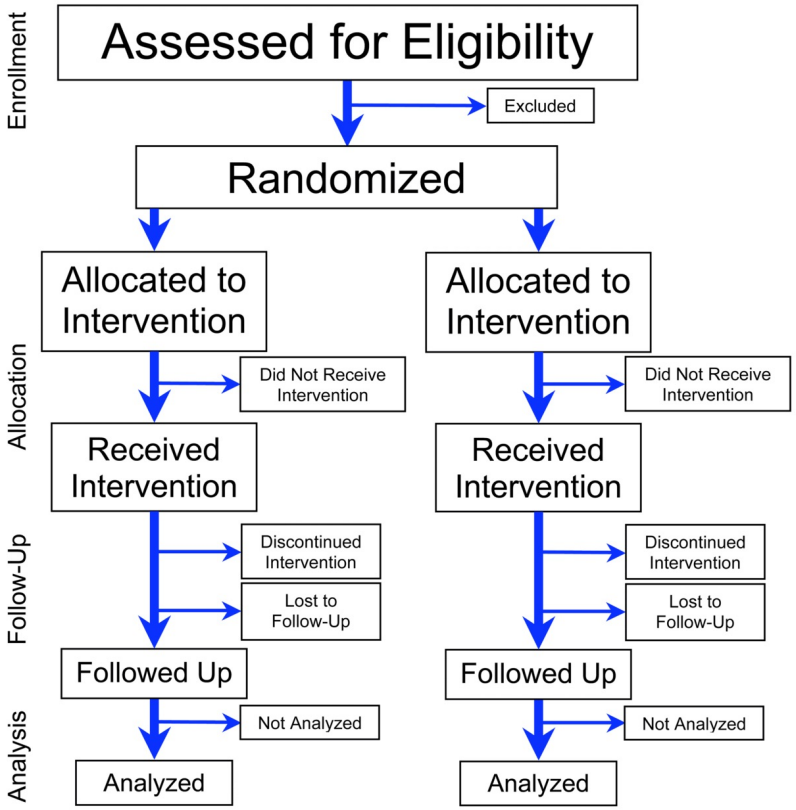
"To find out what happens when you change something, it is necessary to change it."
(Box et al., 2005)

A type of scientific experiment, where the people being studied are randomly allocated to one or other of the different treatments under study. RCTs are considered the gold standard for a clinical trial. RCTs are often used to test the *efficacy* or *effectiveness* of various types of medical intervention and may provide information about adverse effects, such as drug reactions. Random assignment of intervention is done after subjects have been assessed for eligibility and recruited, but before the intervention to be studied begins.

Efficacy:

Effectiveness:

?



Shorter, E. (2011). A brief history of placebos and clinical trials in psychiatry. *Canadian Journal of Psychiatry, 56*(4), 193–197.

The gold standard…

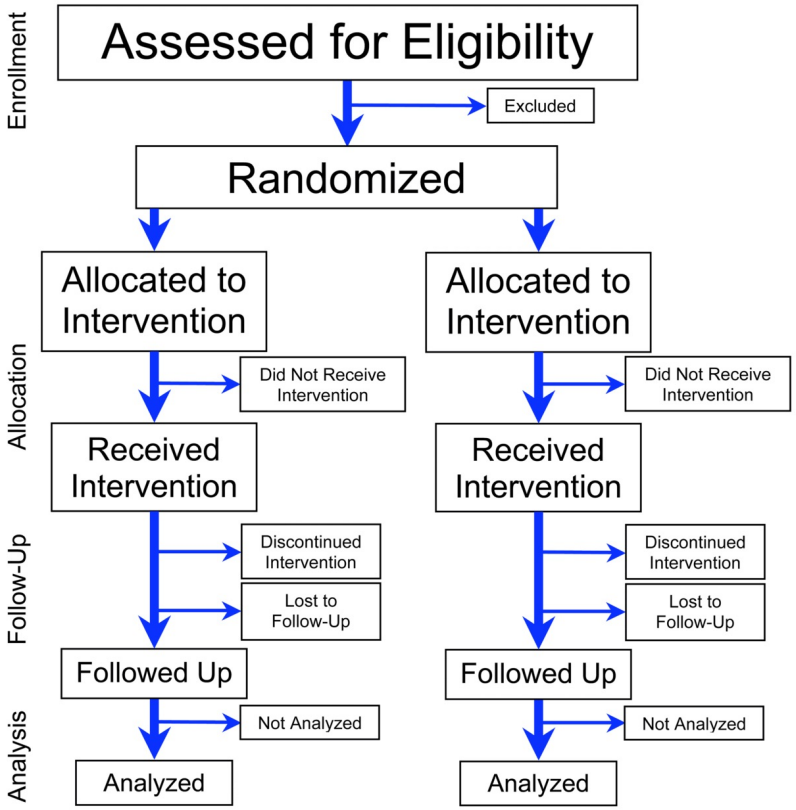
Experiments/Randomised control trials (RCT)

"To find out what happens when you change something, it is necessary to change it."
(Box et al., 2005)

A type of scientific experiment, where the people being studied are randomly allocated to one or other of the different treatments under study. RCTs are considered the gold standard for a clinical trial. RCTs are often used to test the *efficacy* or *effectiveness* of various types of medical intervention and may provide information about adverse effects, such as drug reactions. Random assignment of intervention is done after subjects have been assessed for eligibility and recruited, but before the intervention to be studied begins.

Efficacy: how well a treatment/intervention works under ideal, controlled (e.g., laboratory) settings

Effectiveness: how well a treatment/intervention works in real-world (e.g., clinical) settings



Shorter, E. (2011). A brief history of placebos and clinical trials in psychiatry. *Canadian Journal of Psychiatry, 56*(4), 193–197.

The Salk Polio Vaccine Trial & the Cutter Incident

• The 1954, the Salk Polio vaccine trial was the largest RCT (a double-blind, randomized, and placebo-controlled study) ever conducted, involving over 1.8 million children, to test the safety and efficacy of a polio vaccine developed by Jonas Salk.

• The results showed that the vaccine was safe and effective in preventing polio.

• In 1955, shortly after the Salk polio vaccine was licensed, a manufacturing error at one of 5 licensed laboratories, Cutter Laboratories, resulted in the contamination of some batches of the vaccine with live polio virus, which led to an outbreak that affected a few hundred children, including some deaths and cases of permanent paralysis, known as the Cutter incident.

• The Cutter incident led to significant changes in vaccine regulation including the creation of oversight agencies and legislation.

→ The Cutter incident is an example of the problems that may arise from generalizing RCTs – and the continued need for evaluation (also their legal repercussions)…

A manufacturing error at Cutter Laboratories resulted in the contamination of some batches of the vaccine with live polio virus

Offit, P.A. (2005). The Cutter incident, 50 years later. *N Engl J Med*. 352, 1411-1412.

Dawson, L. (2004). The Salk polio vaccine trial of 1954: Risks, randomization and public involvement in research. *Clinical Trials, 1*, 122–130.

The Nuremberg Code

The Nuremberg Code was formulated in 1947 in Nuremberg, Germany, by American judges sitting in judgment of Nazi doctors accused of conducting murderous and torturous human experiments in concentration camps. It remains an important code for ethical experimentation. (https://en.wikipedia.org/wiki/Doctors%27_Trial)

THE NUREMBERG CODE

1. The voluntary consent of the human subject is absolutely essential.

   This means that the person involved should have legal capacity to give consent; should be so situated as to be able to exercise free power of choice, without the intervention of any element of force, fraud, deceit, duress, overreaching, or other ulterior form of constraint or coercion; and should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him to make an understanding and enlightened decision. This latter element requires that before the acceptance of an affirmative decision by the experimental subject there should be made known to him the nature, duration, and purpose of the experiment; the method and means by which it is to be conducted; all inconveniences and hazards reasonably to be expected; and the effects upon his health or person which may possibly come from his participation in the experiment.

   The duty and responsibility for ascertaining the quality of the consent rests upon each individual who initiates, directs or engages in the experiment. It is a personal duty and responsibility which may not be delegated to another with impunity.

2. The experiment should be such as to yield fruitful results for the good of society, unprocurable by other methods or means of study, and not random and unnecessary in nature.
3. The experiment should be so designed and based on the results of animal experimentation and a knowledge of the natural history of the disease or other problem under study that the anticipated results will justify the performance of the experiment.
4. The experiment should be so conducted as to avoid all unnecessary physical and mental suffering and injury.
5. No experiment should be conducted where there is an a priori reason to believe that death or disabling injury will occur; except, perhaps, in those experiments where the experimental physicians also serve as subjects.
6. The degree of risk to be taken should never exceed that determined by the humanitarian importance of the problem to be solved by the experiment.
7. Proper preparations should be made and adequate facilities provided to protect the experimental subject against even remote possibilities of injury, disability, or death.
8. The experiment should be conducted only by scientifically qualified persons. The highest degree of skill and care should be required through all stages of the experiment of those who conduct or engage in the experiment.
9. During the course of the experiment the human subject should be at liberty to bring the experiment to an end if he has reached the physical or mental state where continuation of the experiment seems to him to be impossible.
10. During the course of the experiment the scientist in charge must be prepared to terminate the experiment at any stage, if he has probable cause to believe, in the exercise of the good faith, superior skill, and careful judgment required of him, that a continuation of the experiment is likely to result in injury, disability, or death to the experimental subject.

Shuster E. (1997). Fifty years later: the significance of the Nuremberg Code. N Engl J Med. 337(20):1436-40. doi: 10.1056/NEJM199711133372006

nature medicine

ARTICLES

https://doi.org/10.1038/s41591-022-01780-9

Check for updates

Safety, tolerability and viral kinetics during SARS-CoV-2 human challenge in young adults

Ben Killingley[1,16], Alex J. Mann[2,16], Mariya Kalinova[2], Alison Boyers[2], Niluka Goonawardane[3], Jie Zhou[3], Kate Lindsell[4], Samanjit S. Hare[5], Jonathan Brown[3], Rebecca Frise[3], Emma Smith[6], Claire Hopkins[7], Nicolas Noulin[2], Brandon Löndt[2], Tom Wilkinson[8], Stephen Harden[9], Helen McShane[10], Mark Baillet[11], Anthony Gilbert[4], Michael Jacobs[12], Christine Charman[4], Priya Mande[4], Jonathan S. Nguyen-Van-Tam[13], Malcolm G. Semple[14], Robert C. Read[8], Neil M. Ferguson[15], Peter J. Openshaw[6], Garth Rapeport[6], Wendy S. Barclay[3], Andrew P. Catchpole[2] and Christopher Chiu[3]✉

Would you participate in such a trial?

"Since its emergence in 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused hundreds of millions of cases and continues to circulate globally. To establish a novel SARS-CoV-2 human challenge model that enables controlled investigation of pathogenesis, correlates of protection and efficacy testing of forthcoming interventions, 36 volunteers aged 18–29 years without evidence of previous infection or vaccination were inoculated with 10 TCID50 of a wild-type virus (SARS-CoV-2/ human/GBR/484861/2020) intranasally in an open-label, non-randomized study"

"We argue that these human challenge studies can reasonably be considered ethically acceptable insofar as such studies are accepted internationally and by the communities in which they are done, can realistically be expected to accelerate or improve vaccine development, have considerable potential to directly benefit participants, are designed to limit and minimise risks to participants, and are done with strict infection control measures to limit and reduce third-party risks."

Jamrozik E, Selgelid MJ. COVID-19 human challenge studies: ethical issues. Lancet Infect Dis. 2020;20(8):e198–203.

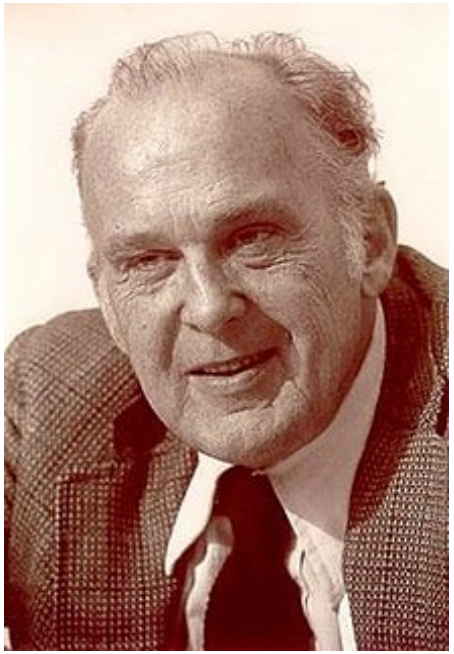
The gold standard is not always gold…

Experiments/Randomised control trials (RCT)

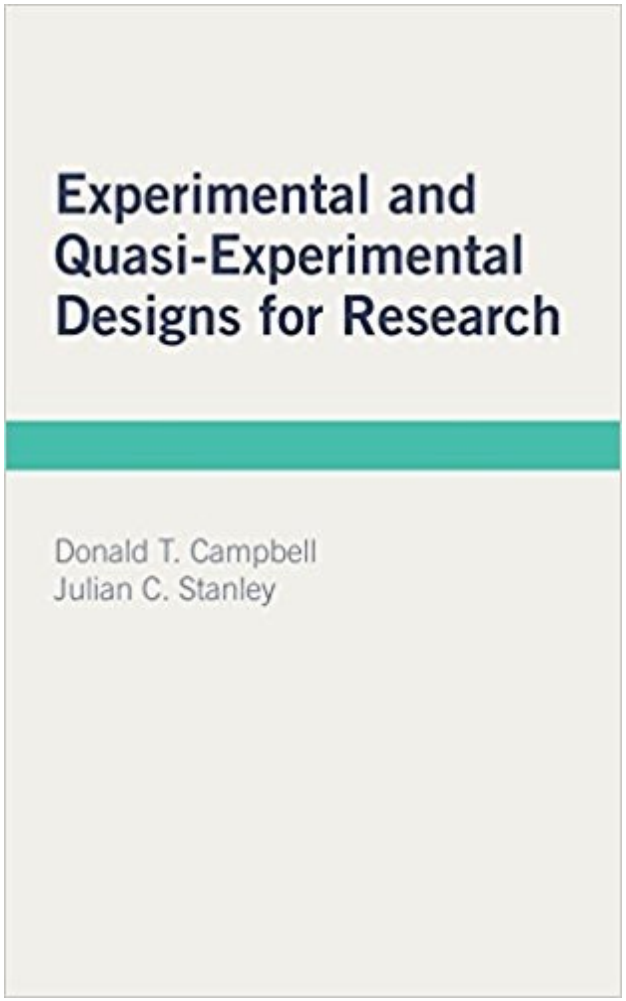
- **Efficacy vs. effectiveness or (lack of) generalizability:** Trials may not be widely applicable in real-world conditions, either because of inability to implement interventions in natural conditions or because interventions do not generalize to other samples/populations

- **Ethical limitations:** Randomization is only ethically justified when there is genuine uncertainty (clinical equipoise)—that is, a state of honest professional disagreement within the expert community regarding the relative merits of interventions; generally, participants should not be randomized to interventions known to be harmful or inferior (in practice, placebo or control conditions are thought ethically permissible under specific conditions, including minimal risk, informed consent, and strong scientific/societal value).

All designs have strengths and weaknesses…

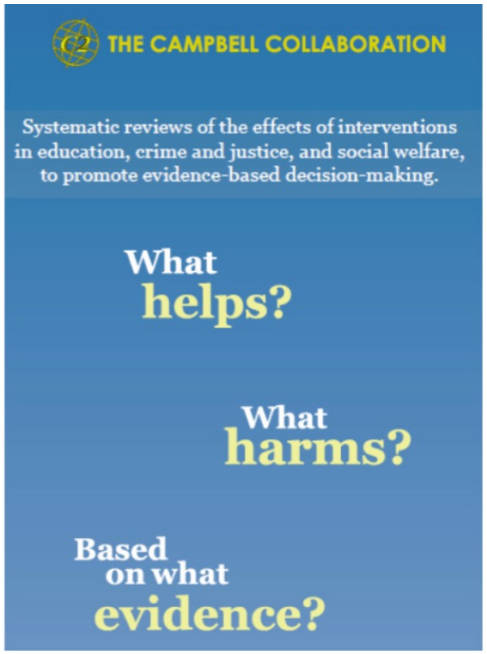


Donald Campbell
1916-1996



1963

named after Campbell…



THE CAMPBELL COLLABORATION

Threats to validity

Internal versus external validity

	Internal validity	External validity (aka generalizability)
Definition		
Key questions		
Threats	???	???
Remedies		

Campbell & Stanley (1963)

Do harsher speeding regulations reduce traffic fatalities?

Threats to validity: Internal Validity

Example: Before-and-after measures

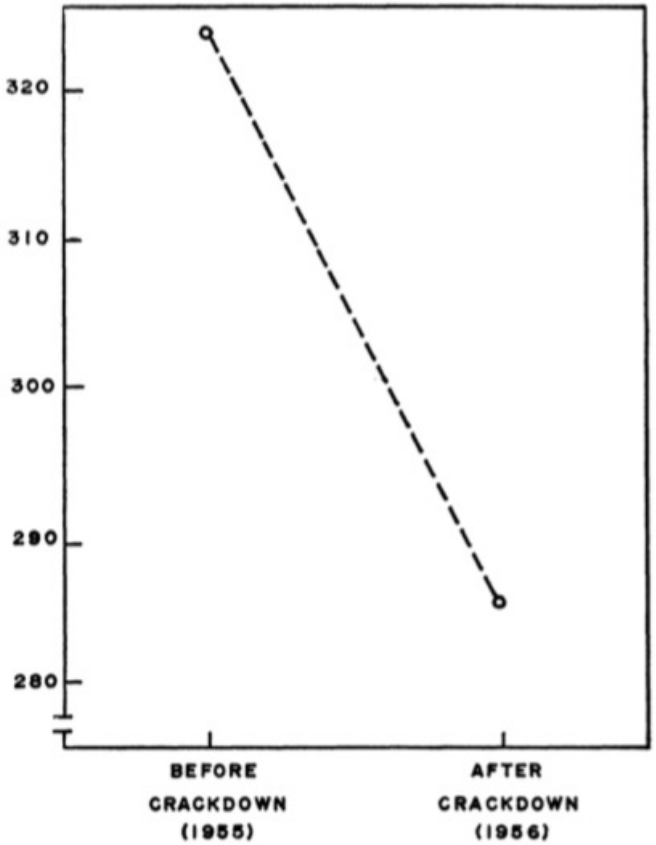


Figure 1. Connecticut Traffic Fatalities, 1955-1956

- was 1956 a special (i.e., dry) year? (history)
- overall declining trend (say due to road safety)? (maturation)
- did publicizing of death rates have an effect? (testing)
- were fatalities counted differently? (instrumentation)
- was 1955 an extreme year in fatalities? (regression)

Why might this NOT be causal?

Campbell, D. T., Ross, H. L. (1968). The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law and Society Review, 3*(1), 33. http://doi.org/10.2307/3052794

Threats to validity: Internal Validity

Example: Interrupted time series

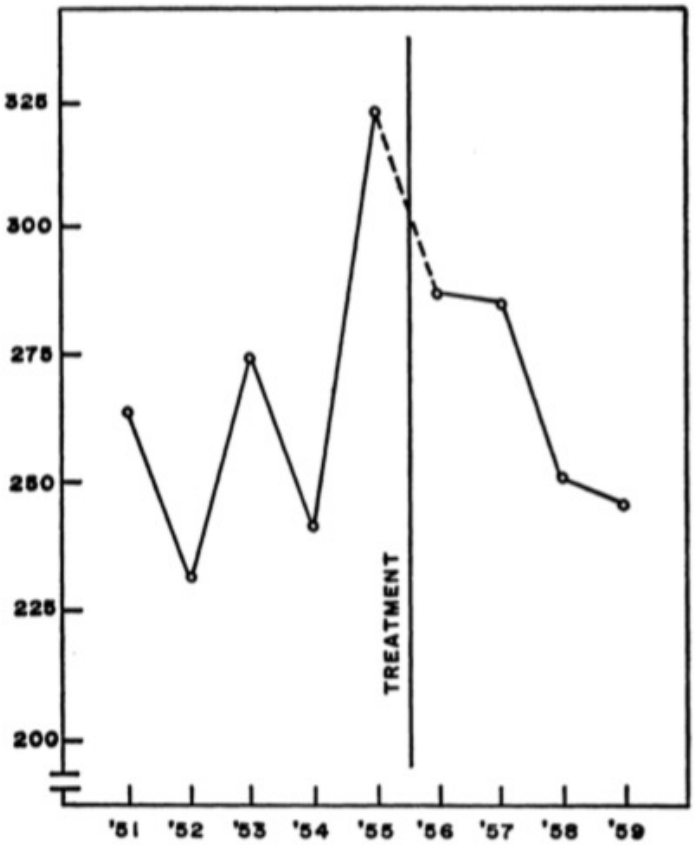


Figure 2. Connecticut Traffic Fatalities, 1951-1959

- was 1956 a special (i.e., dry) year? (history)
- overall declining trend (say due to road safety)? (maturation)
- did publicizing of death rates have an effect? (testing)
- were fatalities counted differently? (instrumentation)
- **was 1955 an extreme year in fatalities?** (regression)

Campbell, D. T., Ross, H. L. (1968). The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law and Society Review, 3*(1), 33. http://doi.org/10.2307/3052794

Threats to validity: Internal Validity

Example: Multiple time series

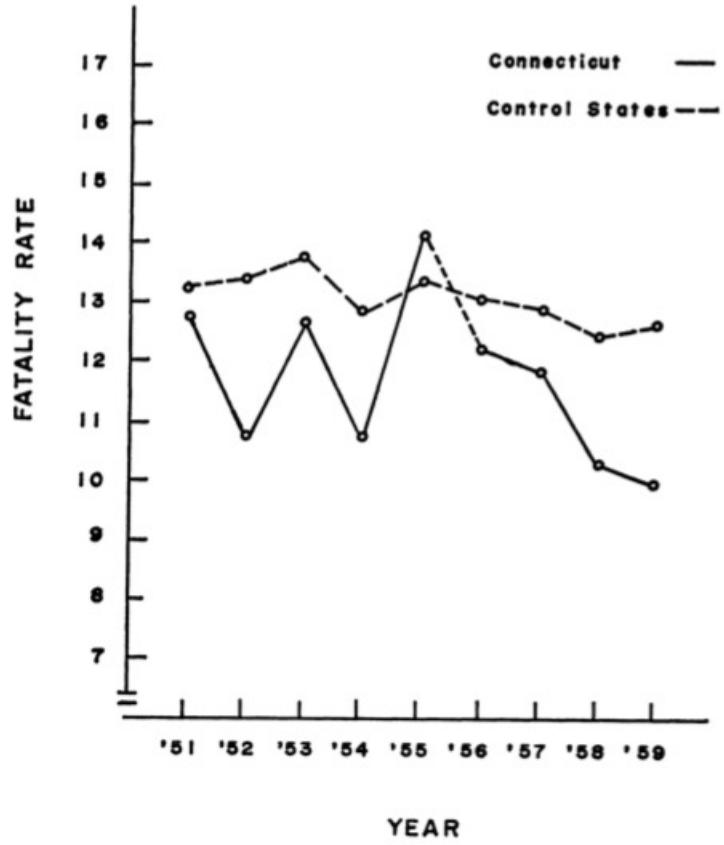


Figure 3. Connecticut and Control States Traffic Fatalities, 1951-1959 (per 100,000 population)

- was 1956 a special (i.e., dry) year? (history)
- overall declining trend (say due to road safety)? (maturation)
- did publicizing of death rates have an effect? (testing)
- were fatalities counted differently? (instrumentation)
- **was 1955 an extreme year in fatalities?** (regression)

Campbell, D. T., Ross, H. L. (1968). The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law and Society Review, 3*(1), 33. http://doi.org/10.2307/3052794

Threats to validity: Internal Validity

Example: Multiple time series

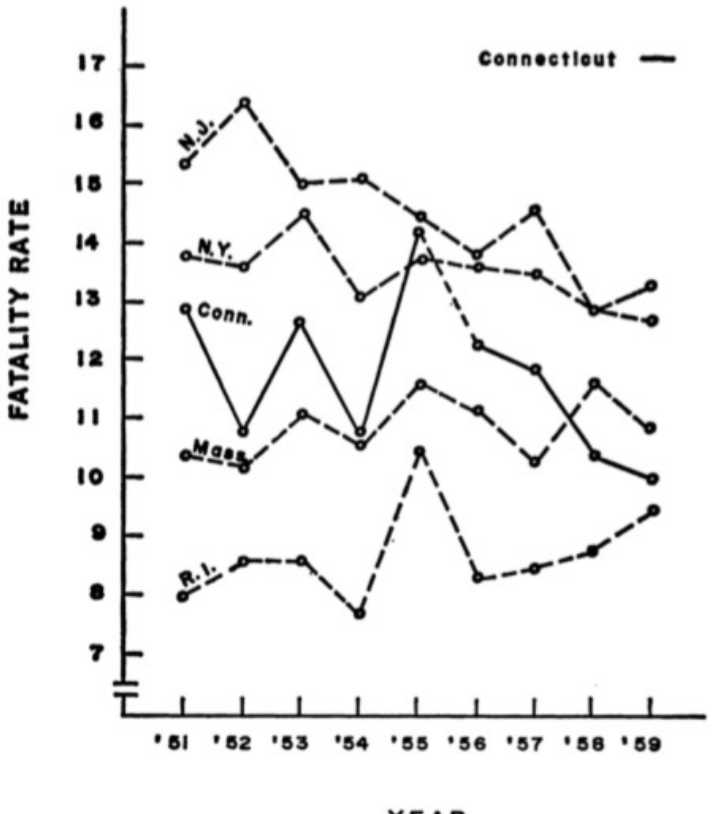


Figure 4. Traffic Fatalities for Connecticut, New York, New Jersey, Rhode Island, and Massachusetts (per 100,000 persons)

- was 1956 a special (i.e., dry) year? (history)
- overall declining trend (say due to road safety)? (maturation)
- did publicizing of death rates have an effect? (testing)
- were fatalities counted differently? (instrumentation)
- was 1955 an extreme year in fatalities? (regression)

Campbell, D. T., Ross, H. L. (1968). The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. *Law and Society Review, 3*(1), 33. http://doi.org/10.2307/3052794

Threats to validity

Internal versus external validity

	Internal validity	External validity (aka representativeness)
Definition	• Assesses the accuracy of causal inferences within the study itself	• Assesses the generalizability of study findings to other populations, settings, and conditions
Key questions	• Did the independent variable manipulation cause changes in the dependent variable? • To what extent can the observed effects be attributed to the experimental treatment?	• Can the findings be applied to other populations beyond the sample studied? • Are the results applicable to real-world situations outside the experimental setting?
Threats	• History, maturation, testing, instrumentation, statistical regression, selection bias, experimental mortality, selection-maturation interaction	???
Remedies	• Random assignment, control groups, counterbalancing, matching, standardized procedures	

Campbell & Stanley (1963)

Threats to validity: External Validity

TABLE 1

SOURCES OF INVALIDITY FOR DESIGNS 1 THROUGH 6

	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation, etc.	Interaction of Testing and X	Interaction of Selection and X	Reactive Arrangements	Multiple-X Interference
Pre-Experimental Designs:												
1. One-Shot Case Study X O	−	−				−	−			−		
2. One-Group Pretest-Posttest Design O X O	−	−	−	−	?	+	+	−	−	−	?	
3. Static-Group Comparison X O --- O	+	?	+	+	+	−	−	−		−		
True Experimental Designs:												
4. Pretest-Posttest Control Group Design R O X O R O O	+	+	+	+	+	+	+	+	−	?	?	
5. Solomon Four-Group Design R O X O R O O R X O R O	+	+	+	+	+	+	+	+	+	?	?	
6. Posttest-Only Control Group Design R X O R O	+	+	+	+	+	+	+	+	+	?	?	

Note: In the tables, a minus indicates a definite weakness, a plus indicates that the factor is controlled, a question mark indicates a possible source of concern, and a blank indicates that the factor is not relevant.

It is with extreme reluctance that these summary tables are presented because they are apt to be "too helpful," and to be depended upon in place of the more complex and qualified presentation in the text. No + or − indicator should be respected unless the reader comprehends why it is placed there. In particular, it is against the spirit of this presentation to create uncomprehended fears of, or confidence in, specific designs.

- X = treatment / event
- O = observation of outcome / effect
- R = randomization
- + = controlled
- - = weakness
- ? = possible concern, depends on context

Interaction of Testing and X

Students get a pretest on math anxiety (O), followed by relaxation (X), then posttest (0). The treatment may appear effective, only because students were primed by the pretest. In the real world (without pretesting), the program might not work as well (see Solomon Four-Group Design)

Interaction of Selection and X

People interested in taking part in a scientific study may be more likely to profit from a new intervention (X), followed by a posttest (O). The treatment may appear effective because it was tested on motivated participants. With more diverse or less motivated individuals, the program might not work as well.

Campbell & Stanley (1963)

TABLE 2

SOURCES OF INVALIDITY FOR QUASI-EXPERIMENTAL DESIGNS 7 THROUGH 12

	Sources of Invalidity											
	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation, etc.	Interaction of Testing and X	Interaction of Selection and X	Reactive Arrangements	Multiple-X Interference
Quasi-Experimental Designs:												
7. Time Series O O O OXO O O O	−	+	+	?	+	+	+	+	−	?	?	
8. Equivalent Time Samples Design X_1O X_2O X_1O X_2O, etc.	+	+	+	+	+	+	+	+	−	?	−	−
9. Equivalent Materials Samples Design M_aX_1O M_bX_2O M_aX_1O M_dX_2O, etc.	+	+	+	+	+	+	+	+	−	?	?	−
10. Nonequivalent Control Group Design O X O --- O O	+	+	+	+	?	+	+	−	−	?	?	
11. Counterbalanced Designs X_1O X_2O X_3O X_4O X_2O X_4O X_1O X_3O X_3O X_1O X_4O X_2O X_4O X_3O X_2O X_1O	+	+	+	+	+	+	+	?	?	?	?	−
12. Separate-Sample Pretest-Posttest Design R O (X) R X O	−	−	+	?	+	+	−	−	+	+	+	
12a. R O (X) R X O --- R O (X) R X O	+	−	+	?	+	+	−	+	+	+	+	
12b. R O_1 (X) R O_2 (X) R X O_3	−	+	+	?	+	+	−	?	+	+	+	
12c. R O_1 X O_2 R X O_3	−	−	+	?	+	+	+	−	+	+	+	

Campbell & Stanley (1963)

TABLE 3

SOURCES OF INVALIDITY FOR QUASI-EXPERIMENTAL DESIGNS 13 THROUGH 16

	Sources of Invalidity											
	Internal								External			
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of Selection and Maturation, etc.	Interaction of Testing and X	Interaction of Selection and X	Reactive Arrangements	Multiple-X Interference
Quasi-Experimental Designs Continued:												
13. Separate-Sample Pretest-Posttest Control Group Design R O (X) R X O R O R O	+	+	+	+	+	+	+	−	+	+	+	
13a. R O (X) / R X O / R O (X) / R X O / R O (X) / R X O / R O / R O / R O / R O / R O / R O	+	+	+	+	+	+	+	+	+	+	+	
14. Multiple Time-Series O O O X O O O O O O O O O O	+	+	+	+	+	+	+	+	−	−	?	
15. Institutional Cycle Design Class A X O_1 Class B_1 RO_2 X O_3 Class B_2 R X O_4 Class C O_5 X [a]Gen. Pop. Con. Cl. B O_6 [a]Gen. Pop. Con. Cl. C O_7												
$O_2 < O_1$; $O_5 < O_4$	+	−	+	+	?	−	?		+	?	+	
$O_2 < O_3$	−	−	−	?	?	+	+		−	?	+	
$O_2 < O_4$	−	−	+	?	?	+	?		+	?	?	
$O_6 = O_7$; $O_{2y} = O_{2o}$		+						−				
16. Regression Discontinuity	+	+	+	?	+	+	?	+	+	−	+	+

[a] General Population Controls for Class B, etc.

Campbell & Stanley (1963)

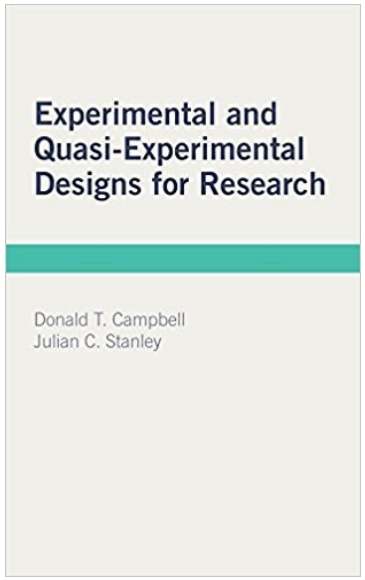
Threats to validity

Internal versus external validity

	Internal validity	External validity (aka representativeness)
Definition	• Assesses the accuracy of causal inferences within the study itself	• Assesses the generalizability of study findings to other populations, settings, and conditions
Key questions	• Did the independent variable manipulation cause changes in the dependent variable? • To what extent can the observed effects be attributed to the experimental treatment?	• Can the findings be applied to other populations beyond the sample studied? • Are the results applicable to real-world situations outside the experimental setting?
Threats	• History, maturation, testing, instrumentation, statistical regression, selection bias, experimental mortiality, selection-maturation interaction	• Reactive/interaction effect of testing, interaction of selection biases and experimental variable, reactive effects of experimental arrangements, multiple-treatment interference
Remedies	• Random assignment, control groups, counterbalancing, matching, standardized procedures	• Representative sampling, cross-validation, field experiments, meta-analysis, external replications

Campbell & Stanley (1963)

All designs have strengths and weaknesses…



"In conclusion, in this chapter we have discussed alternatives in the arrangement or design of experiments, with particular regard to the problems of control of extraneous variables and threats to validity. (…) Throughout, attention has been called to the possibility of creatively utilizing the idiosyncratic features of any specific research situation in designing unique tests of causal hypotheses." (p. 71)

Summary

- **Importance of counterfactuals:** "The critical step in any causal analysis is estimating the counterfactual—a prediction of what would have happened in the absence of the treatment."
- **Limitations for RCTs:** RCTs are great but do not guarantee effectiveness /generalizability, or ethical treatment of participants.
- **Alternatives to RCTs:** Alternatives to experimental designs come in many different forms with different threats to internal and external validity.

Appendix (not mandatory)
The Salk Polio Vaccine Trial & the Cutter Incident