

Evidence-based Decision Making

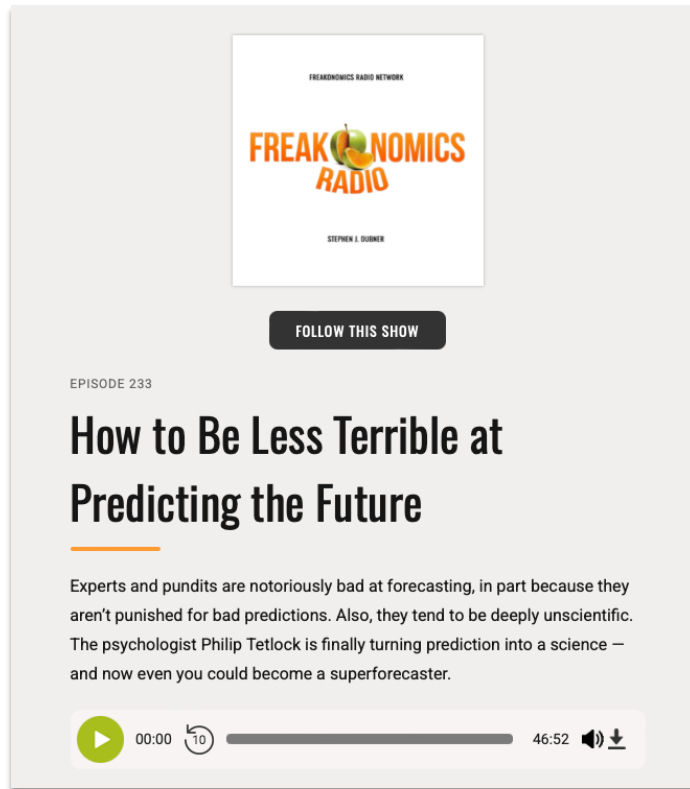
Counterfactuals: Experiments

Loreen Tisdall, FS 2025

Version: April 4, 2025

Last week: Superforecasters

Additional resources (optional)



FREAKONOMICS RADIO NETWORK

FREAKONOMICS RADIO

STEPHEN J. DUBNER

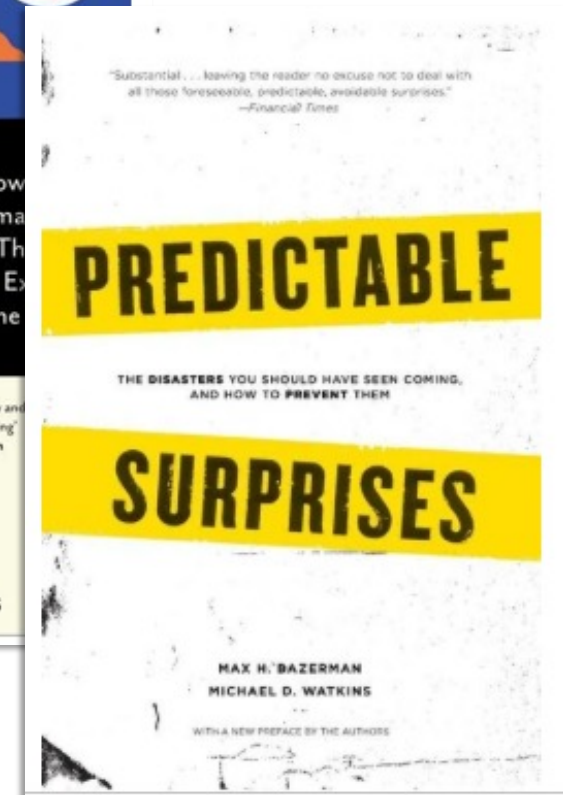
FOLLOW THIS SHOW

EPISODE 233

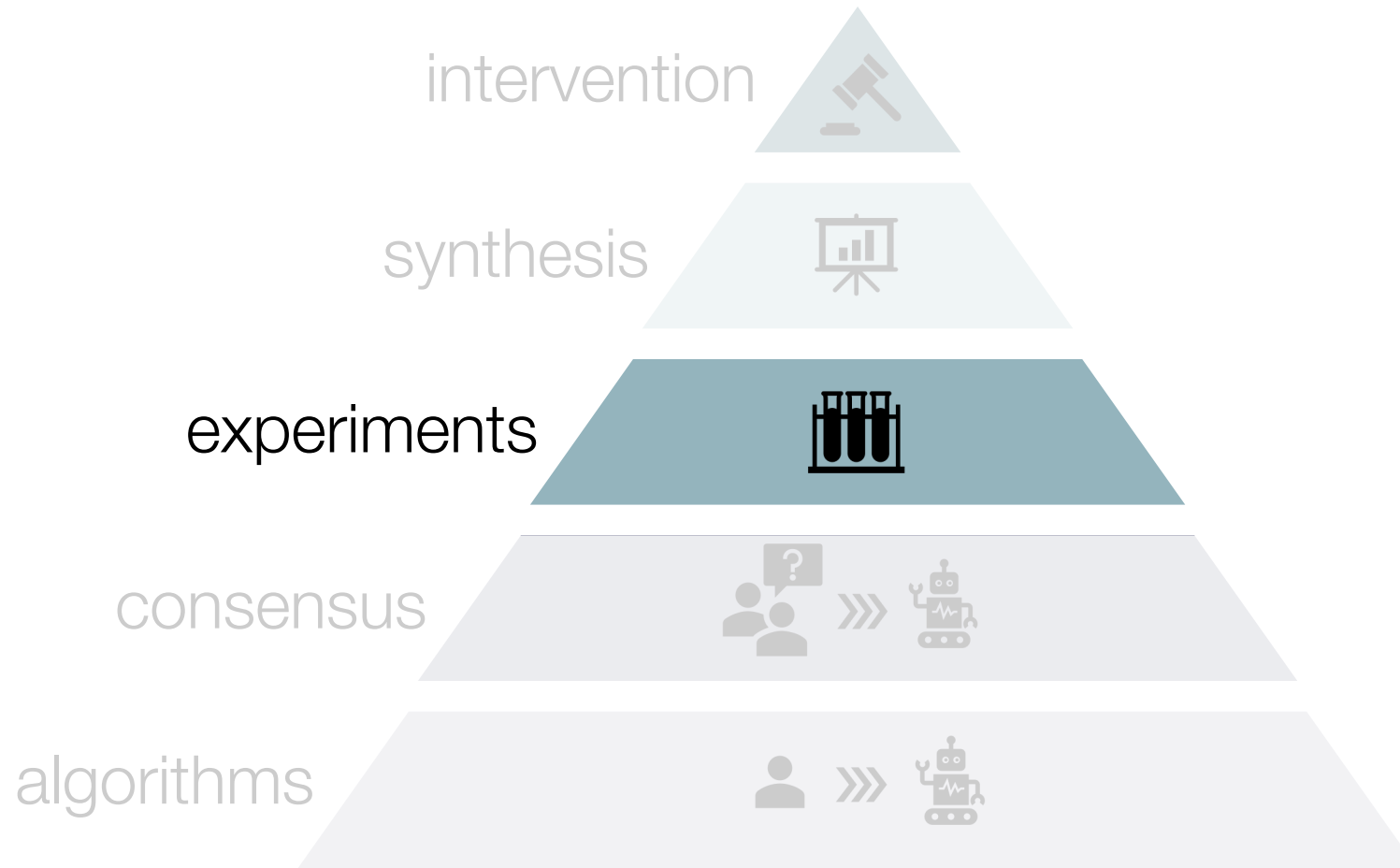
How to Be Less Terrible at Predicting the Future

Experts and pundits are notoriously bad at forecasting, in part because they aren't punished for bad predictions. Also, they tend to be deeply unscientific. The psychologist Philip Tetlock is finally turning prediction into a science – and now even you could become a superforecaster.

00:00 46:52



Climbing the pyramid of evidence



Your turn!

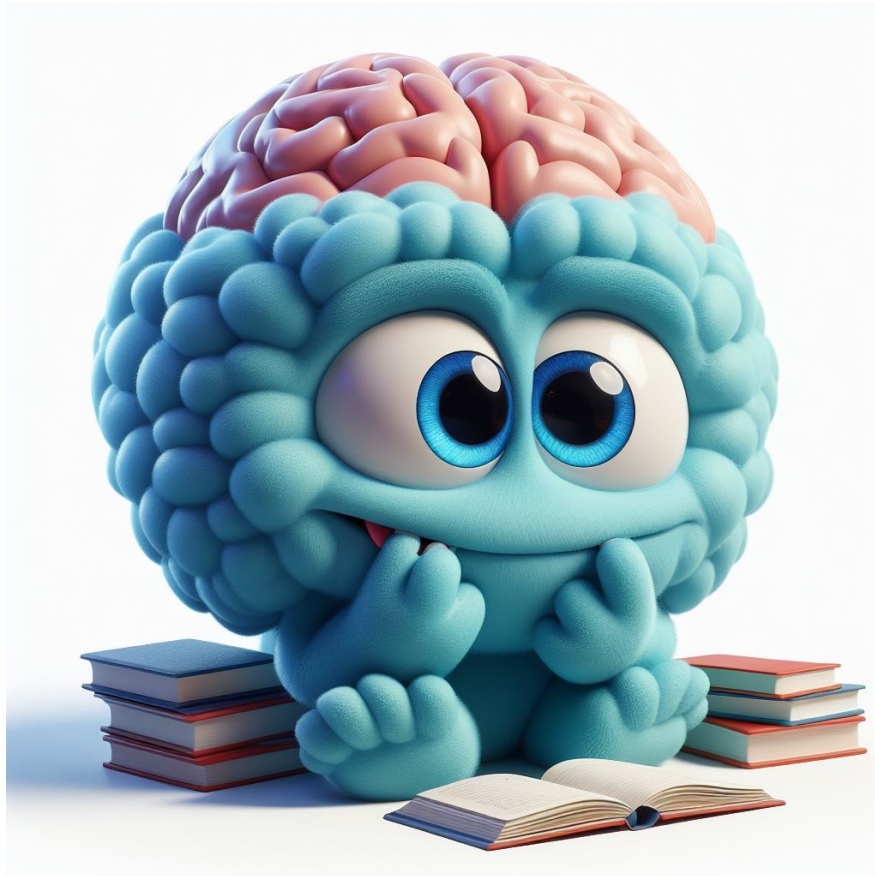


Image created with AI (Bing), January 31, 2024

(Why) Do we need experiments?

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

: someone or something responsible for a result

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

<https://www.merriam-webster.com/thesaurus/causality>

Causal relations as counterfactual relations

Counterfactual = A counterfactual is a conditional statement exploring what would be the case if a certain event or condition, contrary to fact, were true

Example:

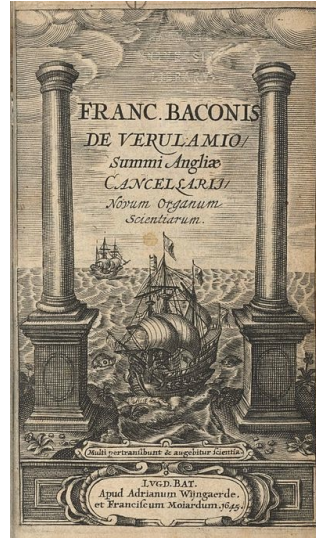
“D shoots at V, but only grazes him, leaving V with a slightly bleeding flesh wound. X then comes along and shoots V through the heart, killing him instantly. D's act is clearly not a "cause in fact" of V's death, since V would have died, and in just the manner he did, even if D had not shot him.”

- **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** ($A \& B \& C \rightarrow Y$), **disjunctive plurality of causes** ($A|B|C \rightarrow Y$)
- Complex regularities (e.g., $A \& B \& C \rightarrow Y$ or $A|B|C \rightarrow Y$) are rarely (if ever) fully known, thus we **formulate propositions which entail the probability** of a variable being causally connected with an effect

Counterfactuals in evidence-based decision making



Francis Bacon
(1561-1626)



1620

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).

Causal inference in economics and marketing

Hal R. Varian¹

¹Economic 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_i denotes per capita sales in city i and x_i denotes per capita ad spend in city i , it is tempting to run a regression of the form $y_i = bx_i + \epsilon_i$, where ϵ_i 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_i are per capita box office receipts for a movie about surfing and x_i 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 $y_i = bx_i$ 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_i and x_i . Such a variable is called a “confounding variable.”

To express this point mathematically, think of (y_i, x_i) as being the population analog of the sample (y, x, ϵ) . The regression coefficient is given by $b = \text{cov}(y, x) / \text{cov}(x, x)$. Substituting $y = bx + \epsilon$, we have

$$b = \text{cov}(bx + \epsilon, x) / \text{cov}(x, x) = b + \text{cov}(\epsilon, x) / \text{cov}(x, x).$$

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



CrossMark
click for updates

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 switched 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_i , is chosen by some decision maker who likely has some views about how various factors affect outcomes, y_i . However, the analyst is not able to observe these factors—they are part of the error term, ϵ_i . If ϵ_i is therefore unlikely that x_i and ϵ_i 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_i on x_i 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, “Estimating Causal Inference from Big Data,” held March 20–21, 2016, at the National Academies of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAE website at www.nationalacademies.org/BigData.

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.

© 2016 Varian; licensee aapm.

These data are not inherently statistical in nature. Suppose that there is no error term, so that the model response is equal to interest in surfing. Its utility is not only that the variation in spend and spend is correlated in surfing interest, we get a misleading picture of the relationship between spend and interest.

It would not have to be that way. Perhaps surfing is no popular in Honolulu that every one already knows about the movie, and it continues to advertise in Fargo. This is the sort of thing the advertiser might know but the analyst does not.

7310-7315 | PNAS | July 5, 2016 | vol. 113 | no. 27

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

“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>

The gold standard...

Experiments/Randomised control trials (RCT)

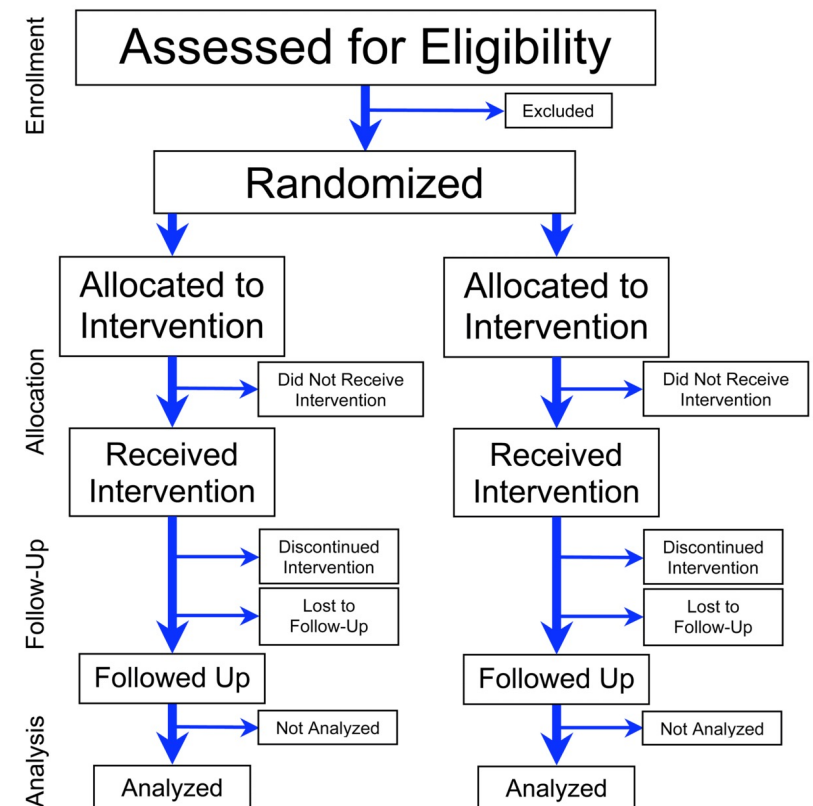
“To find out what happens when you change something, it is necessary to change it.”
(Box et al., 2005)

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:



The gold standard...

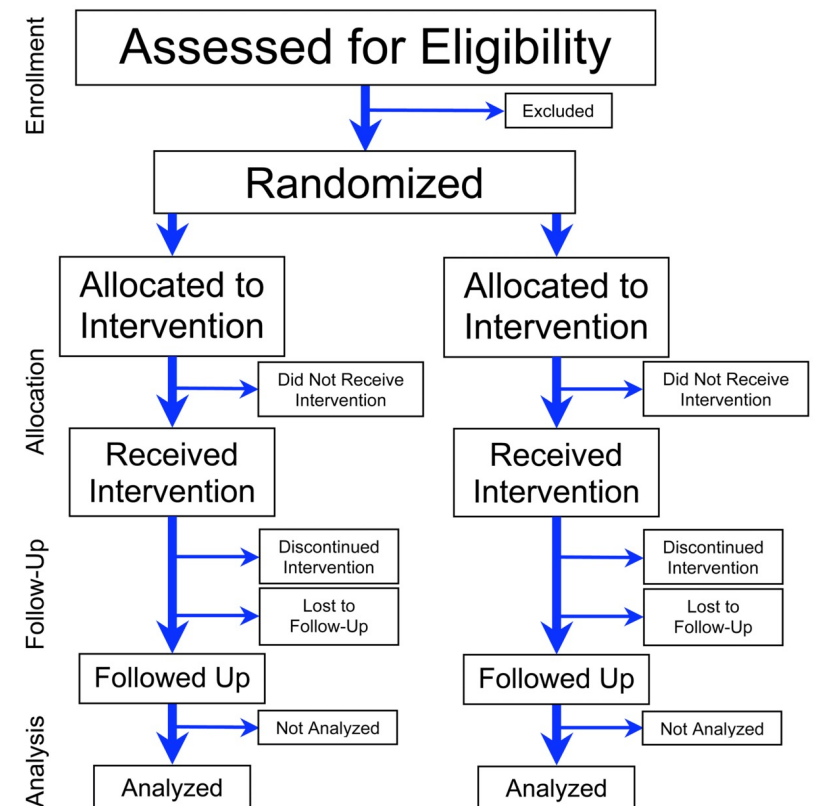
Experiments/Randomised control trials (RCT)

“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 (laboratory) settings

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



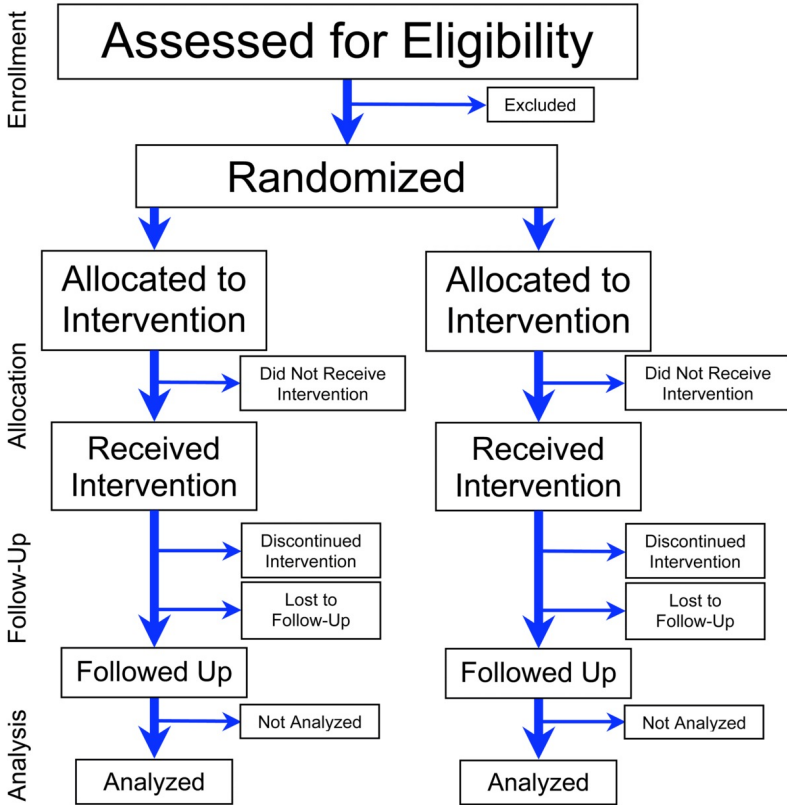
The gold standard...

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.

$$Y = B_0 + B_1 \text{group}$$



The gold standard...

Consolidated Standards of Reporting Trials



The image shows a screenshot of the CONSORT website homepage. At the top, the logo reads "CONSORT TRANSPARENT REPORTING of TRIALS" with a search bar and a "Sign In" button. Below the logo is a navigation menu with links for "Home", "Extensions", "Downloads", "Examples", "Resources", and "About CONSORT". The main content area features a quote from Professor Doug Altman: "To maximise the benefit to society, you need to not just do research but do it well." Below the quote is a "Welcome to the CONSORT Website" section with a brief description of the organization's mission. To the right, there is a "CONSORT 2010 Key Documents" section listing four items: "CONSORT 2010 Checklist", "CONSORT 2010 Flow Diagram", "CONSORT 2010 Statement", and "CONSORT 2010 Explanation and Elaboration Document". At the bottom left, there is a link for "The CONSORT Statement". A dark blue play button icon is visible in the bottom right corner of the screenshot.

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 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 gold standard is not always gold...

Experiments/Randomised control trials (RCT)


- **Efficacy vs. effectiveness:** Trials may not be widely applicable in real-world conditions....
- **Generalizability:** Results may not always generalize to other samples (e.g. inclusion /exclusion criteria)
- **Ethical limitations:** Randomisation requires experimental equipoise (one cannot ethically randomise participants to some treatments that are known to be unsuitable or ineffective)

There are alternatives...

... like synthesis



Donald Campbell
1916-1996



THE CAMPBELL COLLABORATION

Systematic reviews of the effects of interventions
in education, crime and justice, and social welfare,
to promote evidence-based decision-making.

**What
helps?**

**What
harms?**

Based
on what
evidence?



Does education work?

Your turn!



Image created with AI (Bing), January 31, 2024

How could you try to find out if education has an effect on intelligence?

Discuss with your neighbour(s)

~2 minutes

Quasi-experimental designs

control for prior intelligence =

longitudinal studies in which cognitive testing data were collected before and after variation in the duration of education (e.g., before and after university vs. no university)

policy change =

study of the effects of a change in educational duration (e.g., increase of compulsory education by 1 year) on mental testing

school-age cutoff =

studies use regression-discontinuity analysis to leverage the fact that school districts implement a date-of-birth cutoff for school entry (example: compare 3.9-year-olds that are not attending “Kindsgi” vs. 4.0 year-olds that are)

Table 1. Descriptive Statistics for Each Study Design

Design	Control prior intelligence	Policy change	School age cutoff
<i>k</i> studies	7	11	10
<i>k</i> data sets	10	12	20
<i>k</i> effect sizes	26	30	86
<i>N</i> participants	51,645	456,963	107,204
Mean age at early test in years (<i>SD</i>)	12.35 (2.90)	—	—
Mean time lag between tests in years (<i>SD</i>)	53.17 (15.47)	—	—
Mean age at policy change in years (<i>SD</i>)	—	14.80 (2.59)	—
Mean age at outcome test in years (<i>SD</i>)	63.48 (18.80)	47.92 (19.39)	10.36 (1.60)
<i>n</i> outcome test category (composite:fluid:crystallized)	5:20:1	2:23:5	3:67:16
<i>n</i> achievement tests (achievement:other)	1:25	7:23	38:48
Male-only estimates (male only:mixed sex)	2:24	8:22	0:86
Publication status (published:unpublished)	22:4	21:9	64:22

Note: To estimate *N* from studies with multiple effect sizes with different *n*s, we averaged sample sizes across effect sizes within each data set and rounded to the nearest integer. “Unpublished” refers to any study not published in a peer-reviewed journal.

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, 29(8), 1358–1369. <http://doi.org/10.1177/0956797618774253>

Quasi-experimental designs

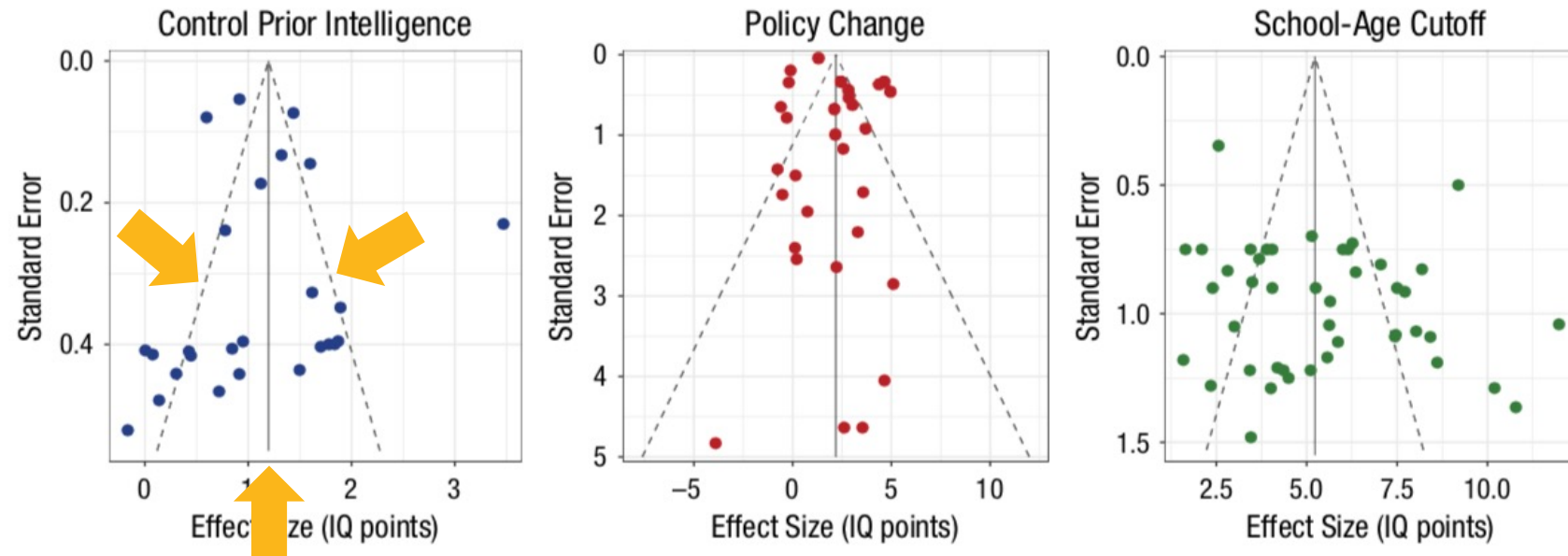


Fig. 2. Funnel plots showing standard error as a function of effect size, separately for each of the three study designs. The dotted lines form a triangular region (with a central vertical line showing the mean effect size) where 95% of estimates should lie in the case of zero within-group heterogeneity in population effect sizes. Note that 42 of the total 86 standard errors reported as approximate or as averages in the original studies were not included for the school-age-cutoff design.

“[...] we found highly consistent evidence that longer educational duration is associated with increased intelligence test scores. [...] Thus, the results support the hypothesis that education has a causal effect on intelligence test scores. The effect of 1 additional year of education—contingent on study design, inclusion of moderators, and publication-bias correction—was estimated at approximately 1 to 5 standardized IQ points.”

Do harsher speeding regulations reduce traffic fatalities?

Quasi-experimental designs

Before-and-after measures

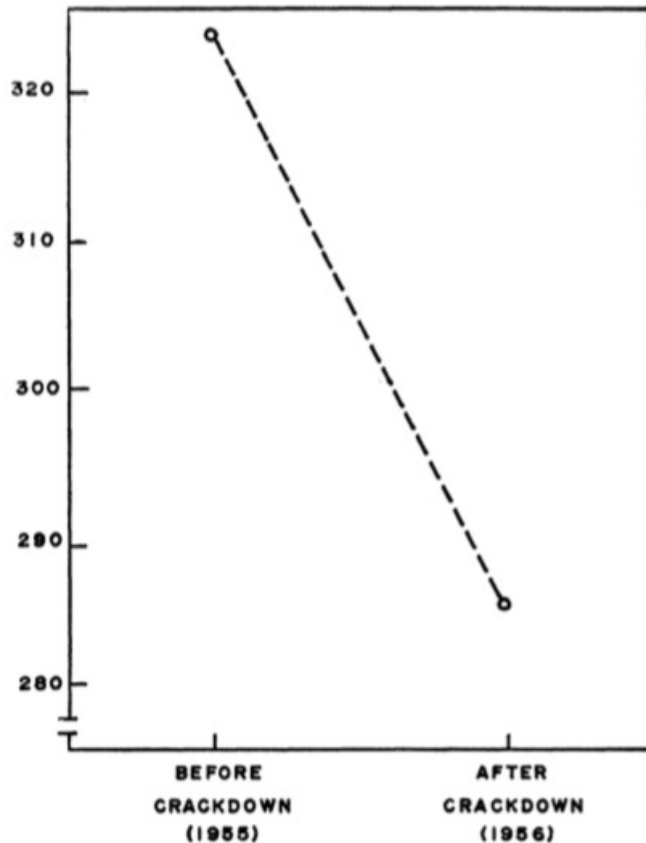


Figure 1. Connecticut Traffic Fatalities, 1955-1956

- was 1956 a dry year? (history)
- overall trends in road safety? (maturation)
- did publicizing of death rates have an effect? (testing)
- were fatalities counted differently? (instrumentation)
- was this a big decrease? (instability)
- was 1955 an extreme year? (regression)

Quasi-experimental designs

Interrupted time series

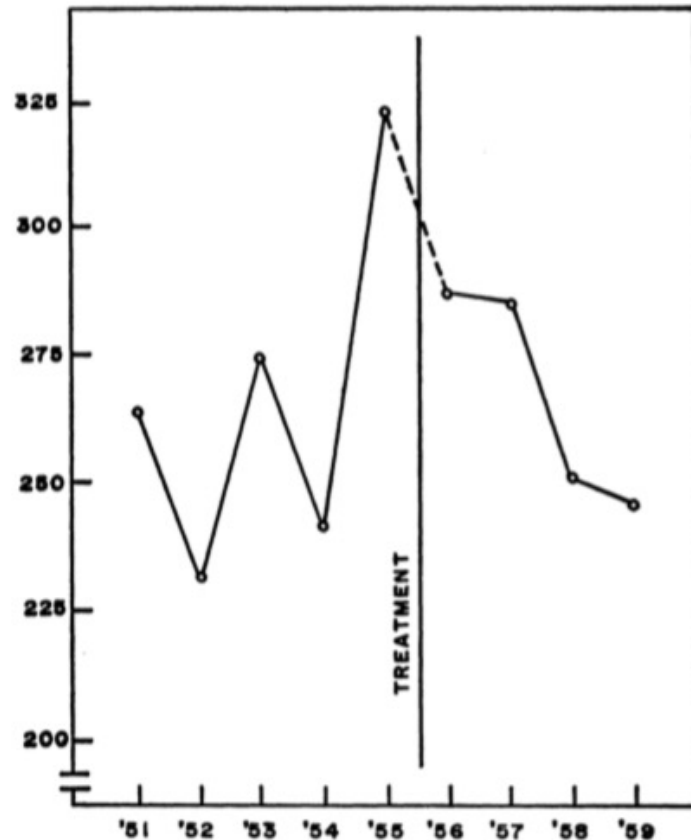


Figure 2. Connecticut Traffic Fatalities, 1951-1959

- was publicizing of death rates similar across years? (testing)
- were fatalities counted differently before and after the intervention? (instrumentation)

Quasi-experimental designs

Multiple time series

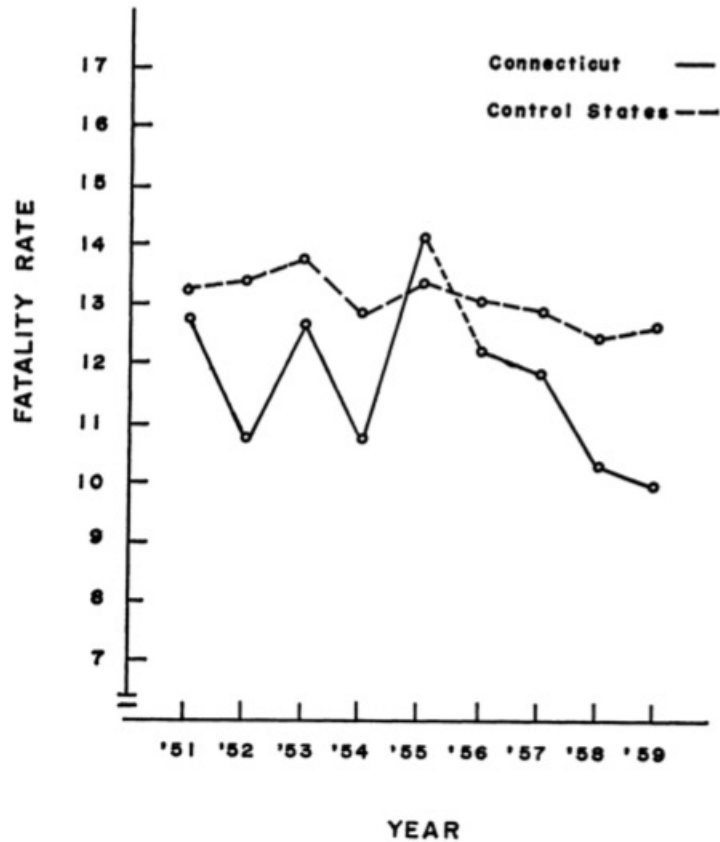


Figure 3. Connecticut and Control States Traffic Fatalities, 1951-1959 (per 100,000 population)

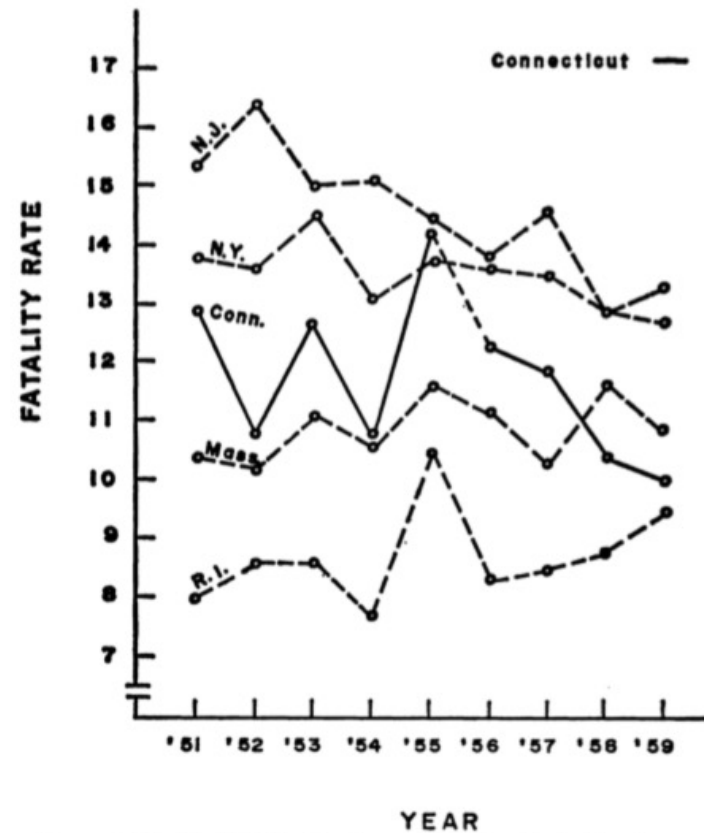


Figure 4. Traffic Fatalities for Connecticut, New York, New Jersey, Rhode Island, and Massachusetts (per 100,000 persons)

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>

Experimental and Quasi-Experimental Designs for Research

Donald T. Campbell
Julian C. Stanley

1963

Factors jeopardizing validity

Internal versus external validity





	Internal validity	External validity (aka representativeness)
Definition	<ul style="list-style-type: none">Assesses the accuracy of causal inferences within the study itself	<ul style="list-style-type: none">Assesses the generalizability of study findings to other populations, settings, and conditions
Key questions	<ul style="list-style-type: none">Did the independent variable manipulation cause changes in the dependent variable?To what extent can the observed effects be attributed to the experimental treatment?	<ul style="list-style-type: none">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		
Remedies		

TABLE 1
SOURCES OF INVALIDITY FOR DESIGNS 1 THROUGH 6

• X = treatment / event
• O = observation of outcome / effect

	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
<i>Pre-Experimental Designs:</i>												
1. One-Shot Case Study X O	-	-				-	-				-	
2. One-Group Pretest-Posttest Design O X O	-	-	-	-	?	+	+	-		-	-	?
3. Static-Group Comparison X O ----- O	+	?	+	+	+	-	-	-		-		
<i>True Experimental Designs:</i>												
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.

TABLE 2

SOURCES OF INVALIDITY FOR QUASI-EXPERIMENTAL DESIGNS 7 THROUGH 12

• X = treatment / event
 • O = observation of outcome / effect

	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
<i>Quasi-Experimental Designs:</i>												
7. Time Series O O O O X O O O O	-	+	+	?	+	+	+	+	-	?	?	
8. Equivalent Time Samples Design X ₁ O X ₀ O X ₁ O X ₀ O, etc.	+	+	+	+	+	+	+	+	-	?	-	-
9. Equivalent Materials Samples Design M _a X ₁ O M _b X ₀ O M _a X ₁ O M _a X ₀ O, etc.	+	+	+	+	+	+	+	+	-	?	?	-
10. Nonequivalent Control Group Design O X O ----- O O	+	+	+	+	?	+	+	-	-	?	?	
11. Counterbalanced Designs X ₁ O X ₂ O X ₁ O X ₂ O ----- X ₂ O X ₁ O X ₂ O X ₁ O ----- X ₂ O X ₁ O X ₁ O X ₂ O ----- X ₁ O X ₂ O X ₂ O X ₁ O	+	+	+	+	+	+	+	?	?	?	?	-
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	-	-	+	?	+	+	+	-	+	+	+	

TABLE 3
SOURCES OF INVALIDITY FOR QUASI-EXPERIMENTAL DESIGNS 13 THROUGH 16

- X = treatment / event
- O = observation of outcome / effect

	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
<i>Quasi-Experimental Designs Continued:</i>												
13. Separate-Sample Pretest-Posttest Control Group Design	+	+	+	+	+	+	+	-	+	+	+	
$\begin{array}{ccc} R & O & (X) \\ R & & X & O \\ \hline \bar{R} & O & & \\ R & & & O \end{array}$												
13a.	+	+	+	+	+	+	+	+	+	+	+	
$\left. \begin{array}{l} R \\ R' \end{array} \right\} \begin{array}{ccc} R & O & (X) \\ R & & X & O \\ \hline \bar{R} & O & (X) \\ R & & X & O \\ \hline R & O & & O \\ R & & & O \\ \hline \bar{R} & O & & \\ R & & & O \\ R & & & O \end{array}$												
14. Multiple Time-Series	+	+	+	+	+	+	+	+	-	-	?	
$\begin{array}{cccccc} O & O & O & X & O & O & O \\ \hline O & O & O & O & O & O & O \end{array}$												
15. Institutional Cycle Design												
Class A	X	O ₁										
Class B ₁	R	O ₂	X	O ₃								
Class B ₂	R		X	O ₄								
Class C		O ₅		X								
*Gen. Pop. Con. Cl. B	O ₆	O ₆										
*Gen. Pop. Con. Cl. C	O ₇	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	+	+	+	?	+	+	?	+	+	-	+	+

• General Population Controls for Class B, etc.

Factors jeopardizing validity

Internal versus external validity

	Internal validity	External validity (aka representativeness)
Definition	<ul style="list-style-type: none"> Assesses the accuracy of causal inferences within the study itself 	<ul style="list-style-type: none"> Assesses the generalizability of study findings to other populations, settings, and conditions
Key questions	<ul style="list-style-type: none"> Did the independent variable manipulation cause changes in the dependent variable? To what extent can the observed effects be attributed to the experimental treatment? 	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> History, maturation, testing, instrumentation, statistical regression, selection bias, experimental mortality, selection-maturation interaction 	<ul style="list-style-type: none"> Reactive/interaction effect of testing, IA of selection biases and experimental variable, reactive effects of experimental arrangements, multiple-treatment interference
Remedies	<ul style="list-style-type: none"> Random assignment, control groups, counterbalancing, matching, standardized procedures 	<ul style="list-style-type: none"> Representative sampling, cross-validation, field experiments, meta-analysis, external replications

Experimental and Quasi-experimental Designs



“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)

A colorful bouquet of creating counterfactuals

“The stronger the demonstrated consistency of an association under conditions that rule out alternative hypotheses and the stronger the evidence regarding a mechanism that can explain the observed association, the more likely we are to accept the causal hypothesis. Usually the evidence required to confirm a causal hypothesis is cumulated across multiple studies, many of which are, of necessity, observational. **Although a wide variety of research designs and analytic techniques are available to assist in gathering evidence to support a causal inference, they are helpful only to the extent that their use is guided and constrained by appropriate subject-matter considerations. No method or set of methods defines causality.**”

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:** Automation is on the rise, but ethical and safety issues will be crucial! Quasi-experimental designs come in many different forms with different threats to internal and external validity.

Have a good week and see you next Monday!

Appendix (not mandatory)

The Salk Polio Vaccine Trial & the Cutter Incident



Appendix (not mandatory)

On the horizon: Autonomous scientific agents

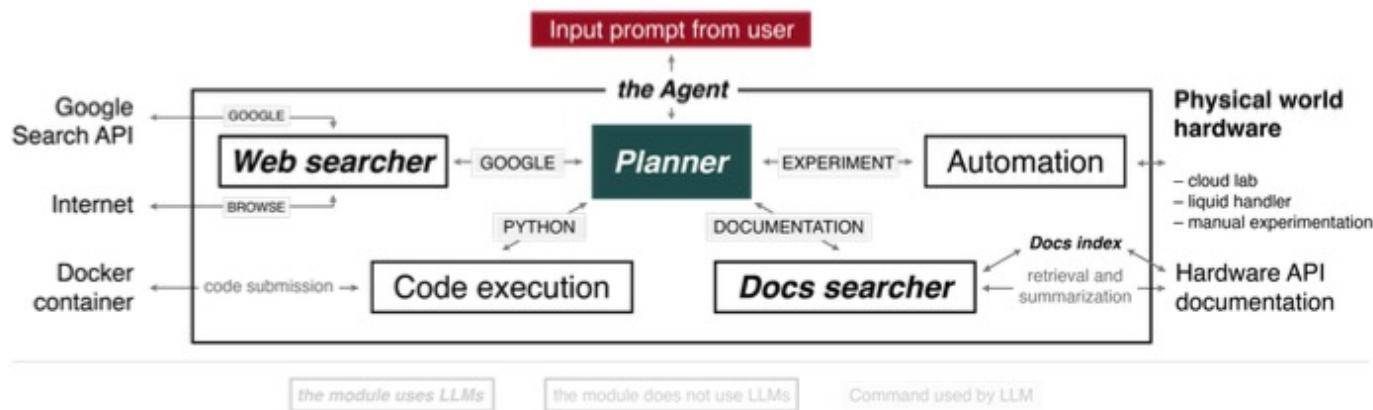


Figure 1. Overview of the system architecture. The Agent is composed of multiple modules that exchange messages. Some of them have access to APIs, the Internet, and Python interpreter.

In this paper, we presented an Intelligent Agent system capable of autonomously designing, planning, and executing complex scientific experiments. Our system demonstrates exceptional reasoning and experimental design capabilities, effectively addressing complex problems and generating high-quality code. However, the development of new machine learning systems and automated methods for conducting scientific experiments raises substantial concerns about the safety and potential dual use consequences, particularly in relation to the proliferation of illicit activities and security threats. By ensuring the ethical and responsible use of these powerful tools, we can continue to explore the vast potential of large language models in advancing scientific research while mitigating the risks associated with their misuse.