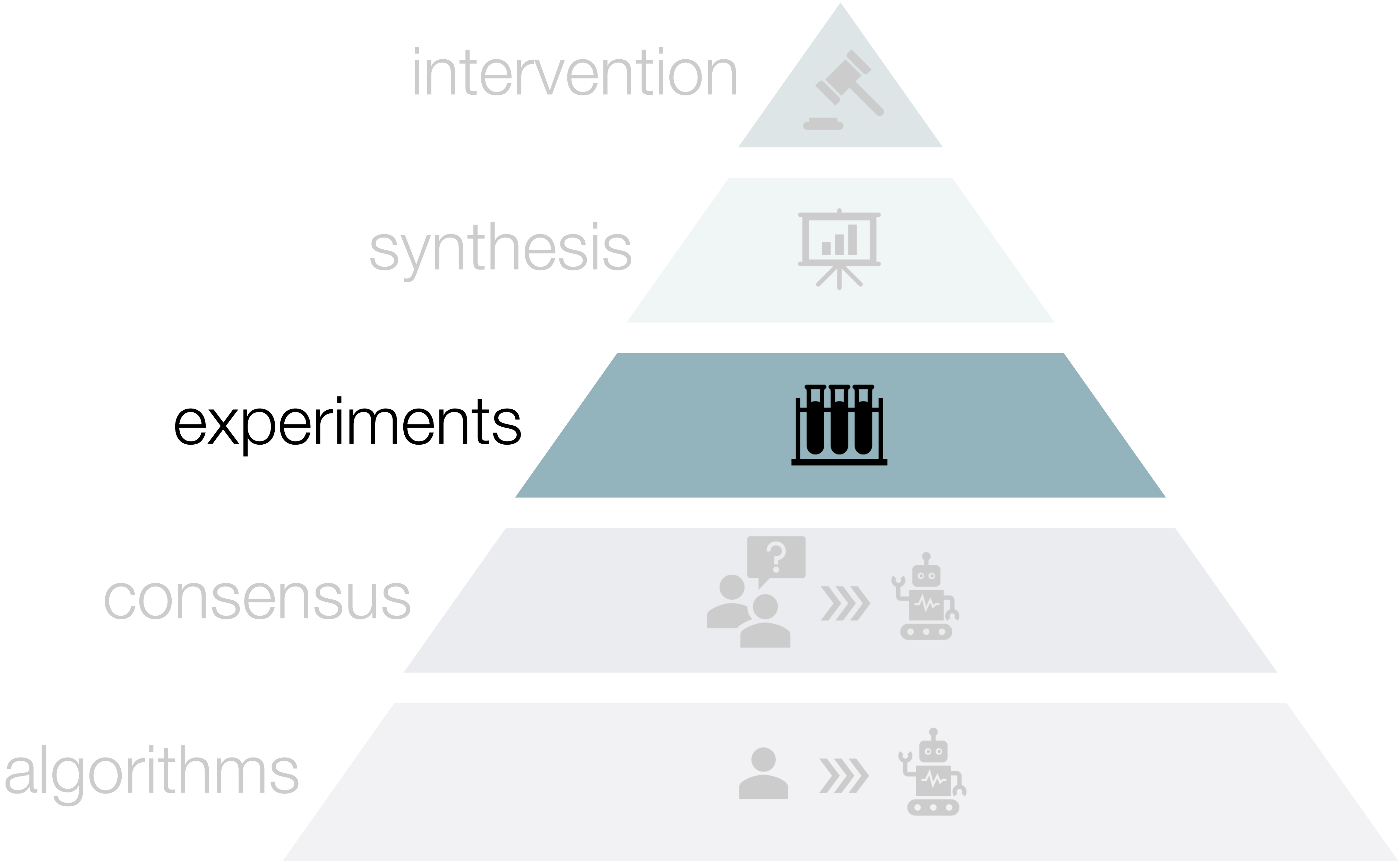
Evidence-based Decision Making

Counterfactuals: Experiments

Loreen Tisdall, FS 2024

Version: February 23, 2024



intervention

synthesis

experiments

consensus

algorithms

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

> *"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."*

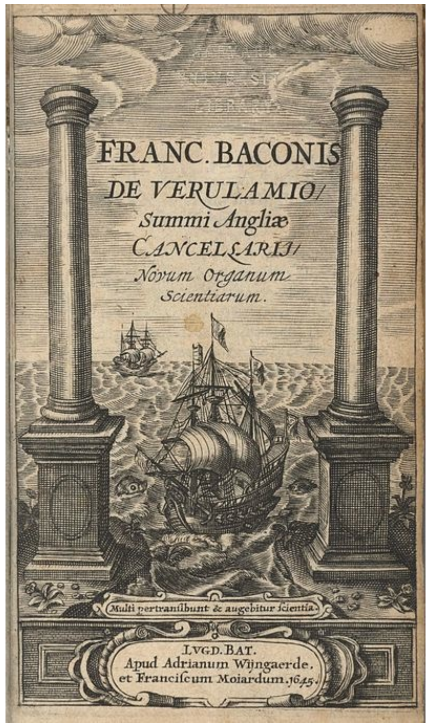
- Singular judgment of causation: a single cause (e.g., A) is necessary and sufficient for effect (Y) to occur
- **In reality:** conjunctive plurality of causes (A&B&C → Y), disjunctive plurality of causes (A|B|C → Y)
- Complex regularities (e.g., A&B&C → Y) are rarely (if ever) fully known, thus we 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.

Evidence-based decision making

MORTI MELIORA
The right Hon.ble Francis Lo: Veru-
lam, Viscount St. Alban. mortuus 9 Aprilis,
Anno Dni. 1626. Annoq Ætat 66.

Francis Bacon
(1561-1626)



FRANC. BACONIS
DE VERULAMIO/
Summi Angliæ
CANCELLARII/
Novum Organum
Scientiarum.

Multi pertranſibunt & augebitur ſcientia.

LVGD. BAT.
Apud Adrianum Wijngaerde.
et Franciſcum Moiardum. 1645.

1620

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 $b = \mathrm{cov}(x, y)/\mathrm{cov}(x, x)$. Substituting $y = bx + e$, we have

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

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

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: hal@sims.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

Bacon suggests that one can draw up a list of all things in which the phenomenon to explain occurs, as well as a list of things in which it does not occur. Then one can rank the lists according to the degree in which the phenomenon occurs in each one. Then one should be able to deduce what factors match the occurrence of the phenomenon in one list and do not occur in the other list, and also what factors change in accordance with the way the data had been ranked.

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

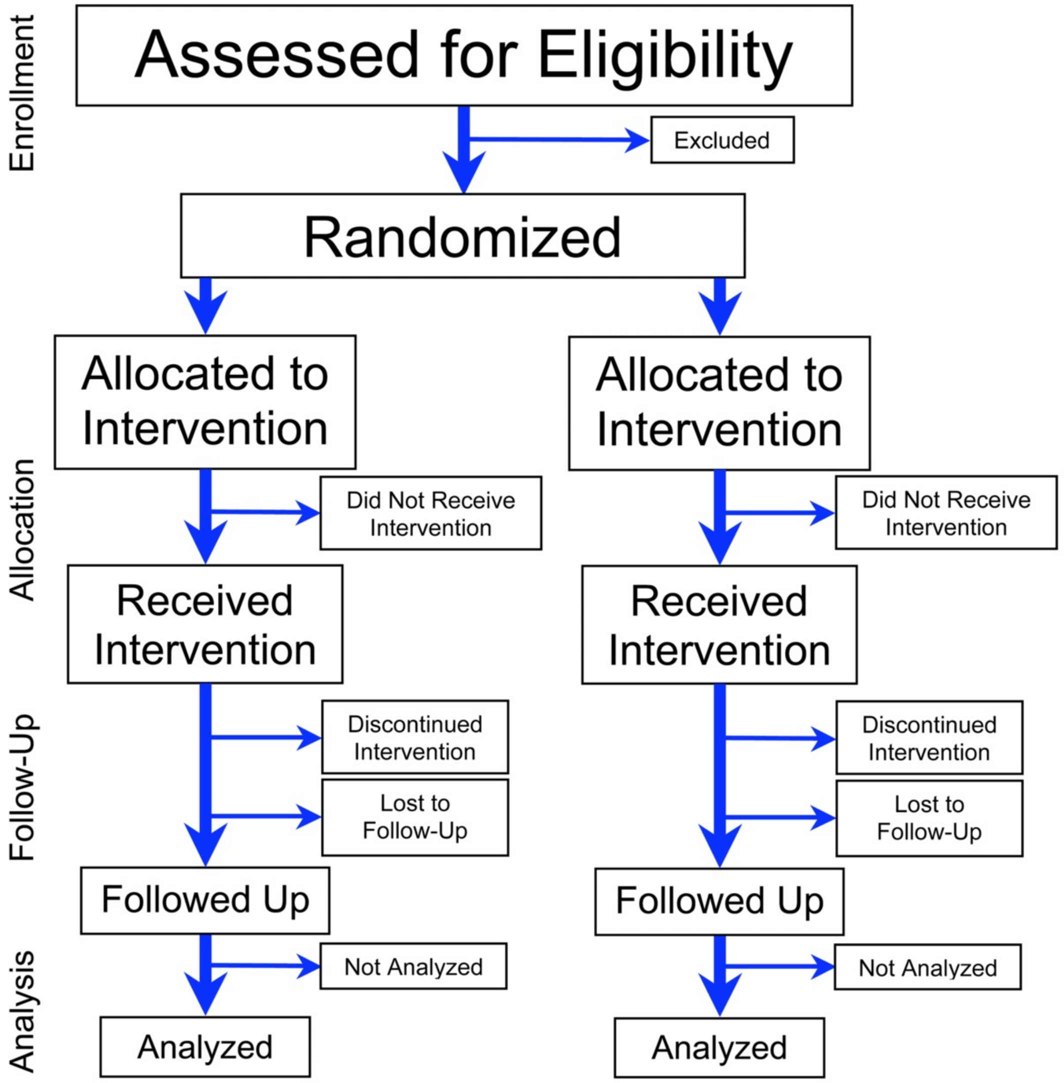
Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences of the United States of America, 113*(27), 7310–7315. http://doi.org/10.1073/pnas.1510479113

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:

?

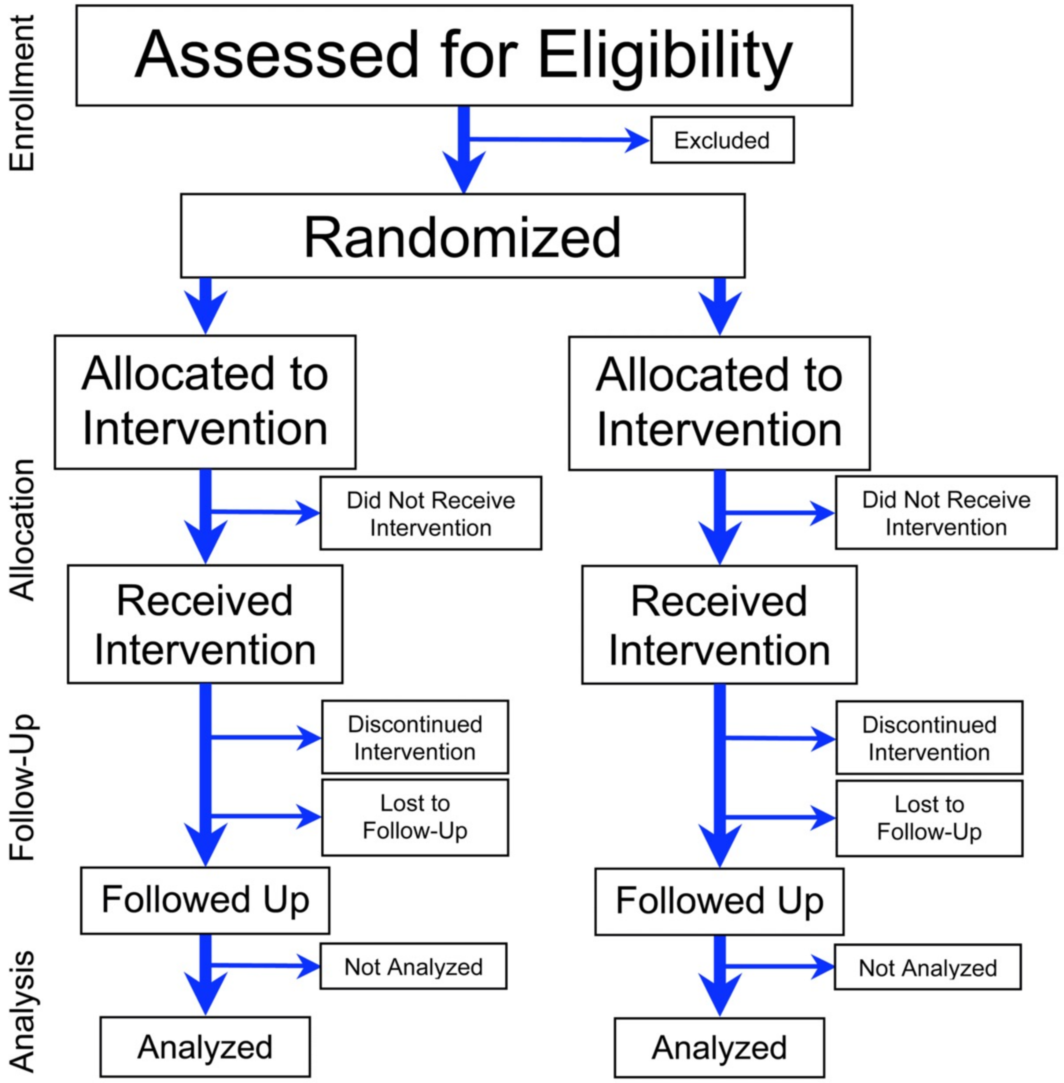
Effectiveness:

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)

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

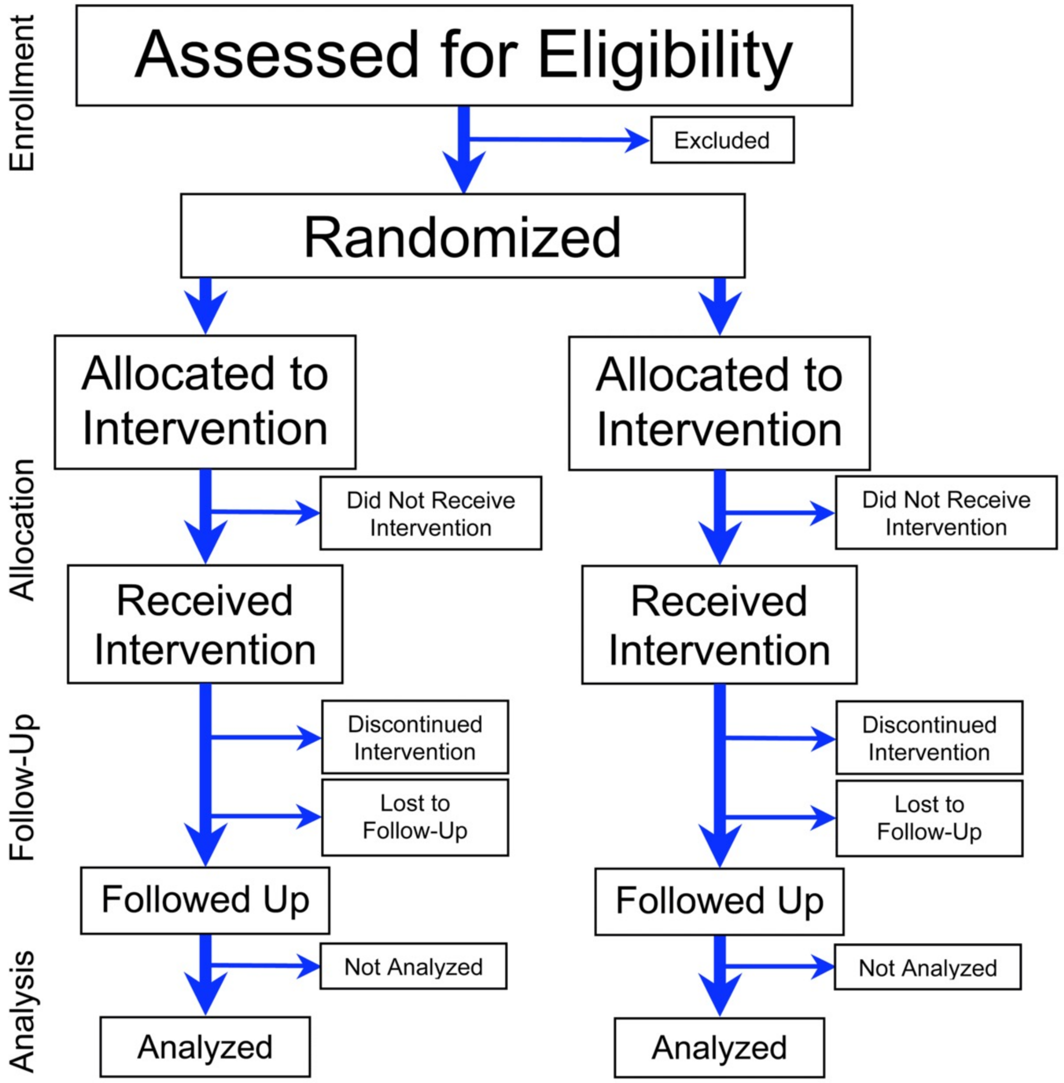
Shorter, E. (2011). A brief history of placebos and clinical trials in psychiatry. *Canadian Journal of Psychiatry, 56*(4), 193–197.

The gold standard…

Experiments/Randomised control trials (RCT)

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.

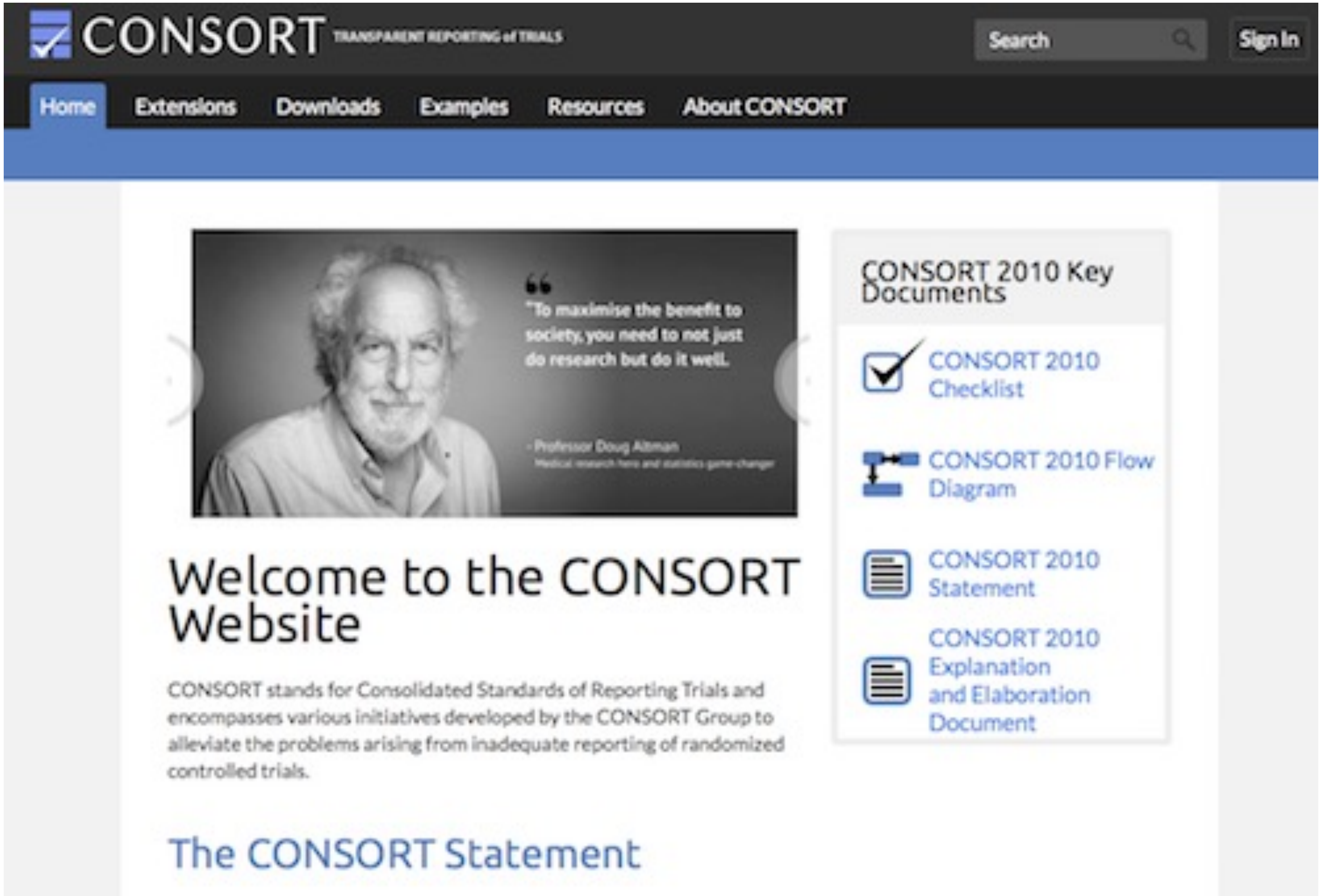
$$Y = B_0 + B_1 \text{group}$$



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

The gold standard…

Consolidated Standards of Reporting Trials



CONSORT TRANSPARENT REPORTING of TRIALS

Search

Sign In

Home | Extensions | Downloads | Examples | Resources | About CONSORT

"To maximise the benefit to society, you need to not just do research but do it well.

- Professor Doug Altman
Medical research hero and statistics game-changer

Welcome to the CONSORT Website

CONSORT stands for Consolidated Standards of Reporting Trials and encompasses various initiatives developed by the CONSORT Group to alleviate the problems arising from inadequate reporting of randomized controlled trials.

The CONSORT Statement

CONSORT 2010 Key Documents

- CONSORT 2010 Checklist
- CONSORT 2010 Flow Diagram
- CONSORT 2010 Statement
- CONSORT 2010 Explanation and Elaboration Document

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 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.

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

→ 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)…

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 (no-schooling condition)

On the horizon: Autonomous Scientific Agents

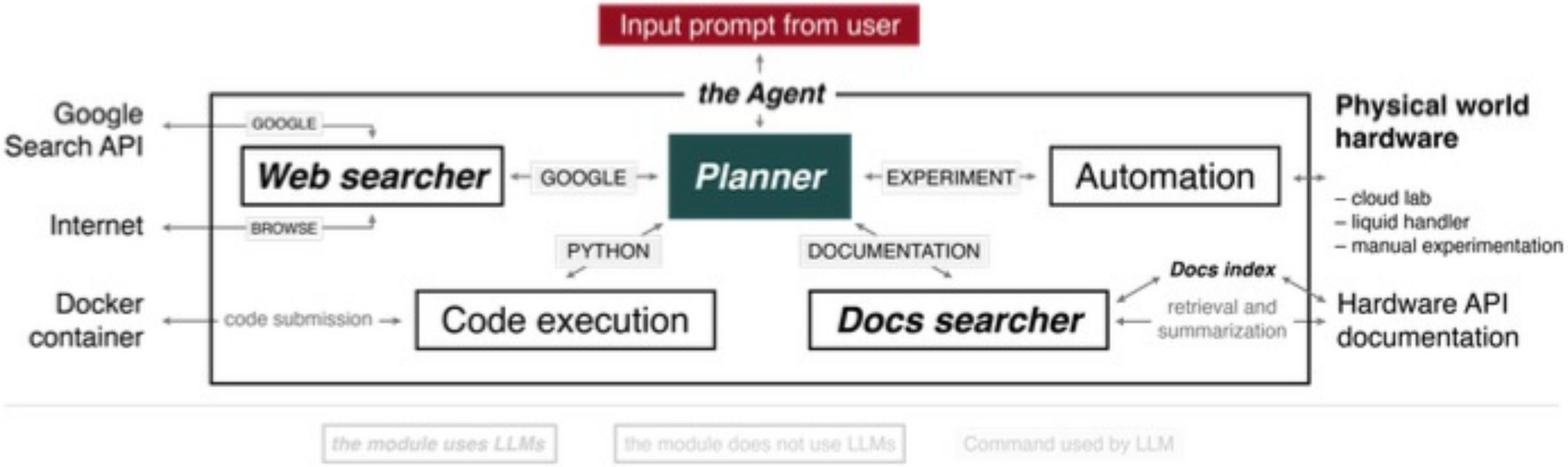
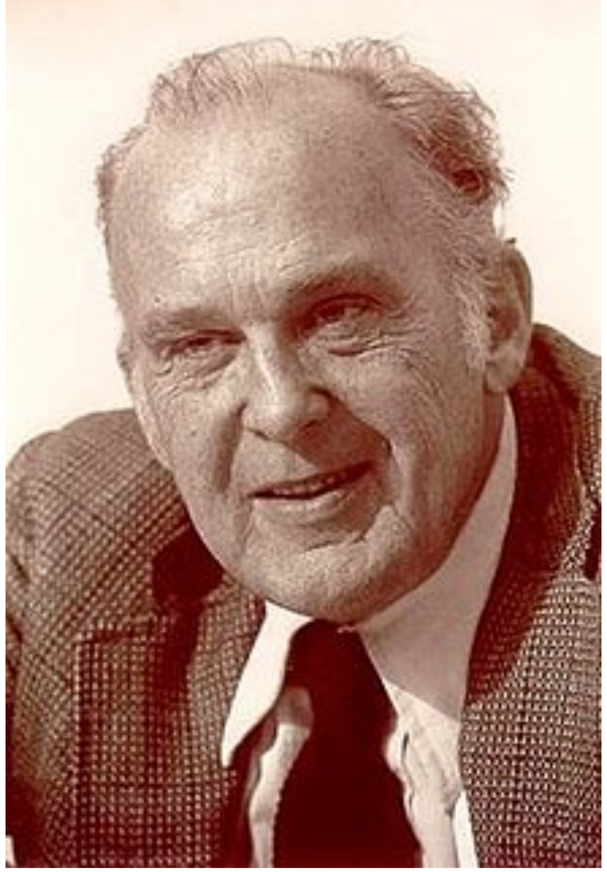


Figure 1. Overview of the system architecture. The **Agent** is composed of multiple modules that exchange messages. Some of them have access to APIs, the Internet, and Python interpreter.

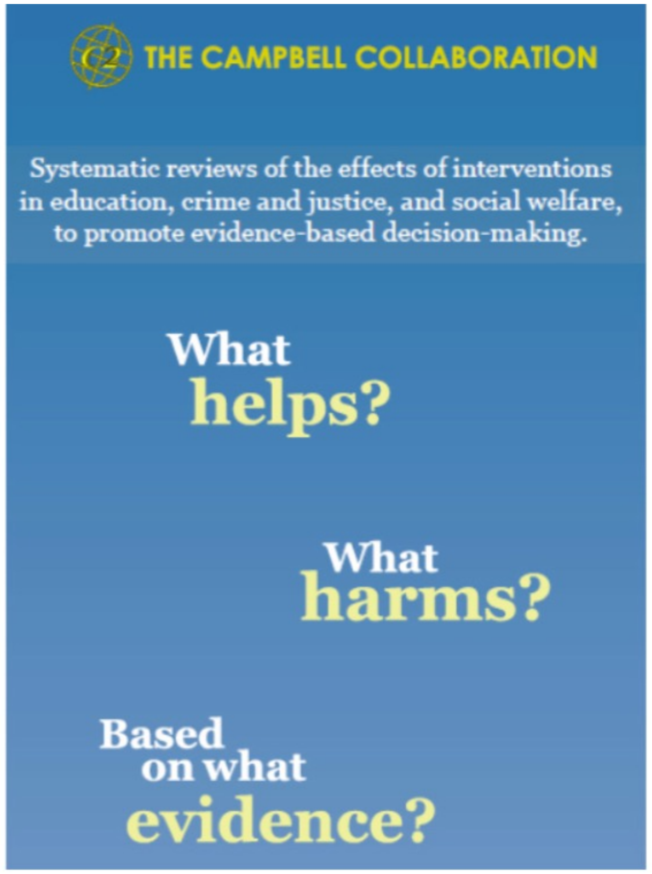
In this paper, we presented an Intelligent **`Agent`** system capable of autonomously designing, planning, and executing complex scientific experiments. Our system demonstrates exceptional reasoning and experimental design capabilities, effectively addressing complex problems and generating high-quality 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.

Boiko, D.A., MacKnight, R. & Gomes, G. L. (2023). Emergent autonomous scientific research capabilities of large language models. https://doi.org/10.48550/arXiv.2304.05332

There are alternatives…



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?

THE CAMPBELL
COLLABORATION

Does education work?

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

YOUR TURN!

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

Quasi-Experimental Designs: Educational effects on intelligence

control 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 did not attend "Kindsgi" vs. 4.0 year-olds that did)

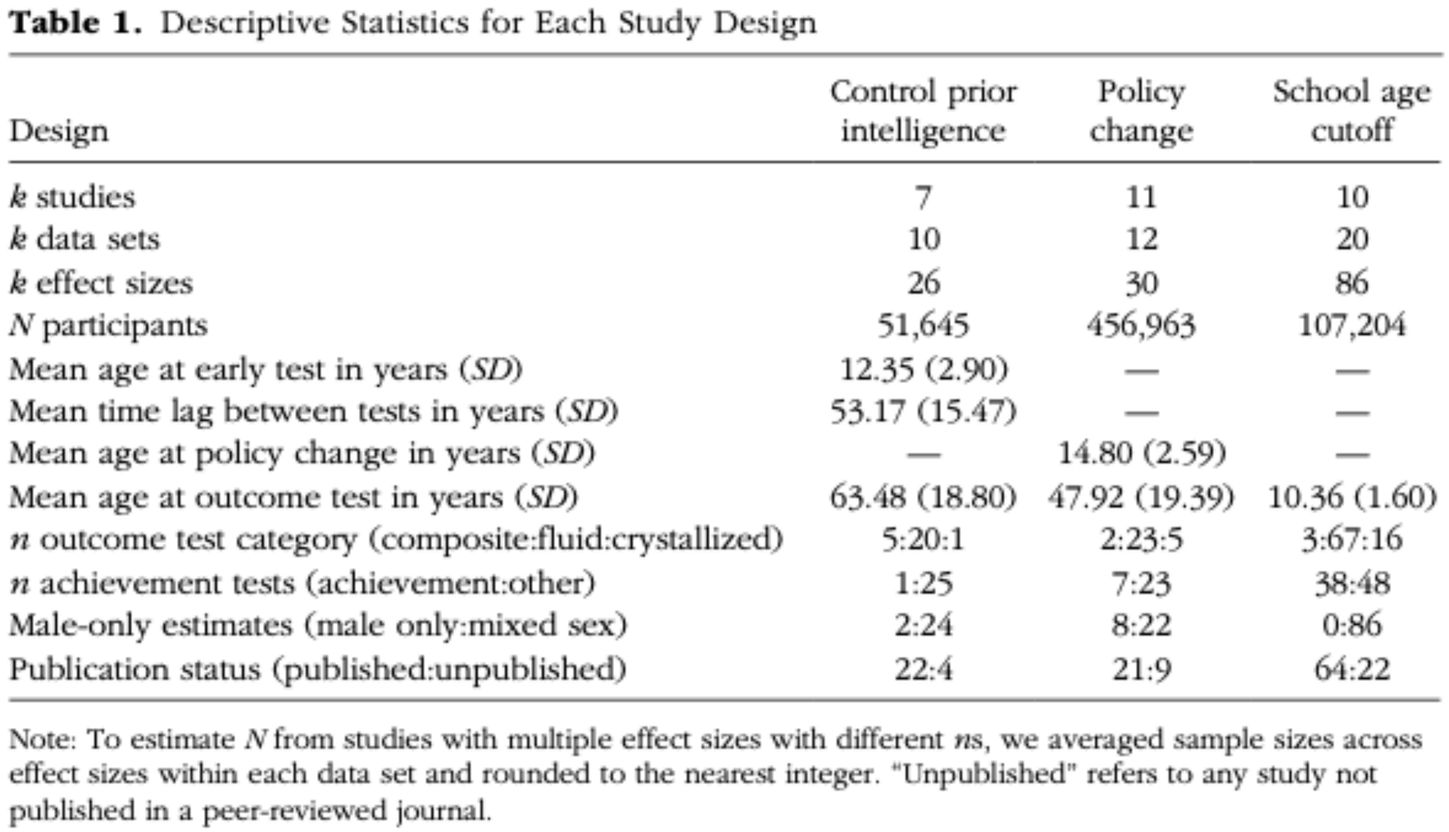


Table 1. Descriptive Statistics for Each Study Design

Design	Control prior intelligence	Policy change	School age cutoff
k studies	7	11	10
k data sets	10	12	20
k effect sizes	26	30	86
N participants	51,645	456,963	107,204
Mean age at early test in years (*SD*)	12.35 (2.90)	—	—
Mean time lag between tests in years (*SD*)	53.17 (15.47)	—	—
Mean age at policy change in years (*SD*)	—	14.80 (2.59)	—
Mean age at outcome test in years (*SD*)	63.48 (18.80)	47.92 (19.39)	10.36 (1.60)
n outcome test category (composite:fluid:crystallized)	5:20:1	2:23:5	3:67:16
n 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: Educational effects on intelligence

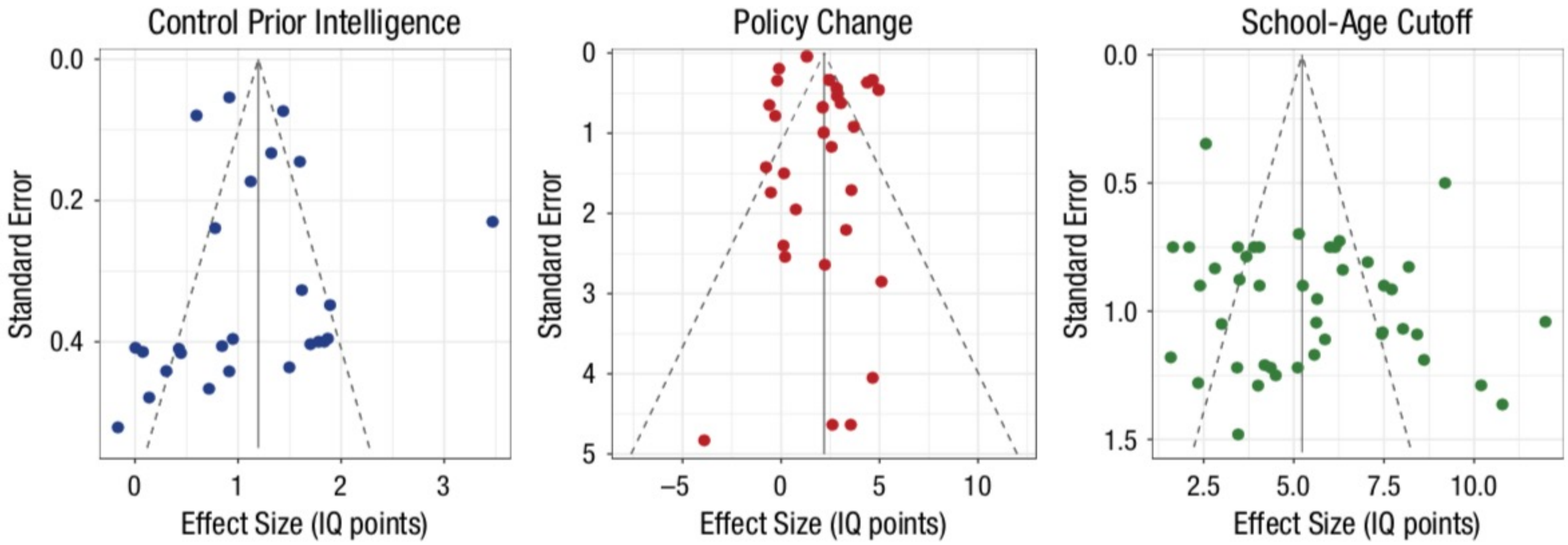


Fig. 2. Funnel plots showing standard error as a function of effect size, separately for each of the three study designs. The dotted lines form a triangular region (with a central vertical line showing the mean effect size) where 95% of estimates should lie in the case of zero within-group heterogeneity 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."

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

Before-and-after measures

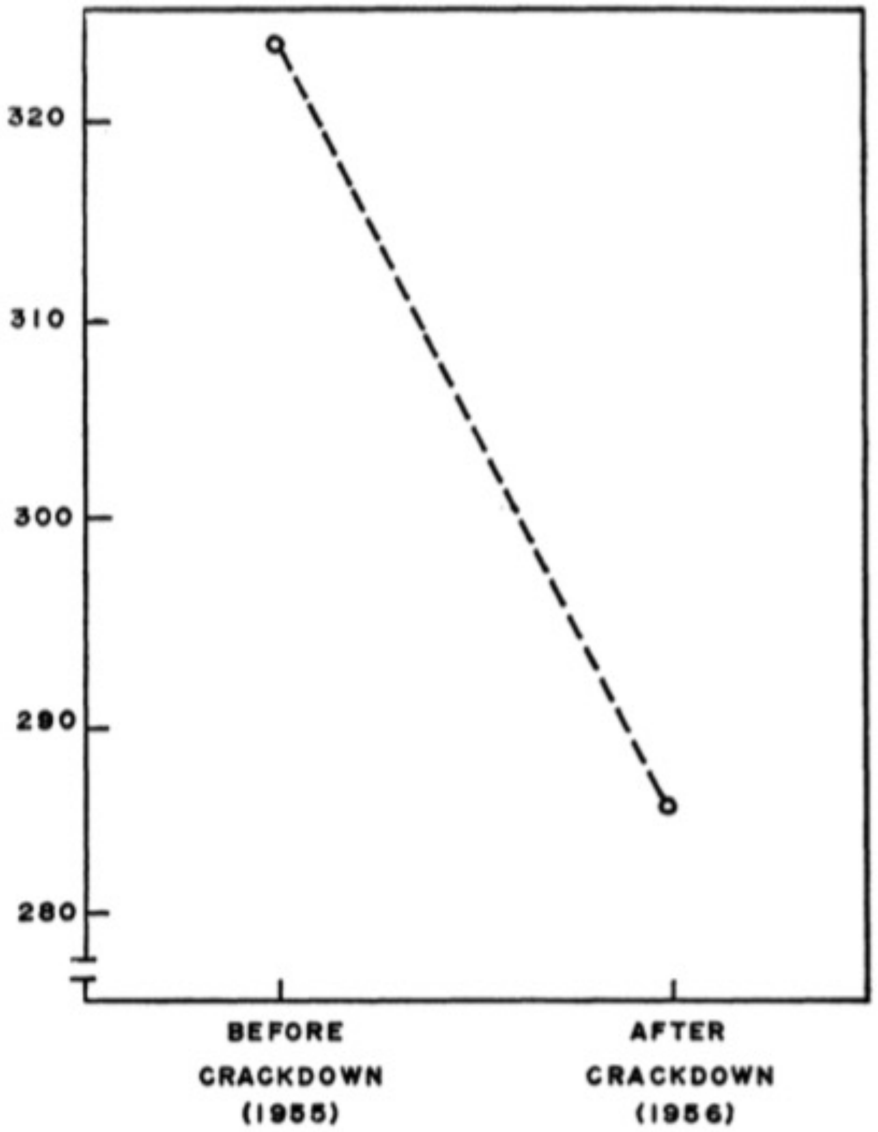


Figure 1. Connecticut Traffic Fatalities, 1955-1956

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

Quasi-experimental designs

Multiple time series

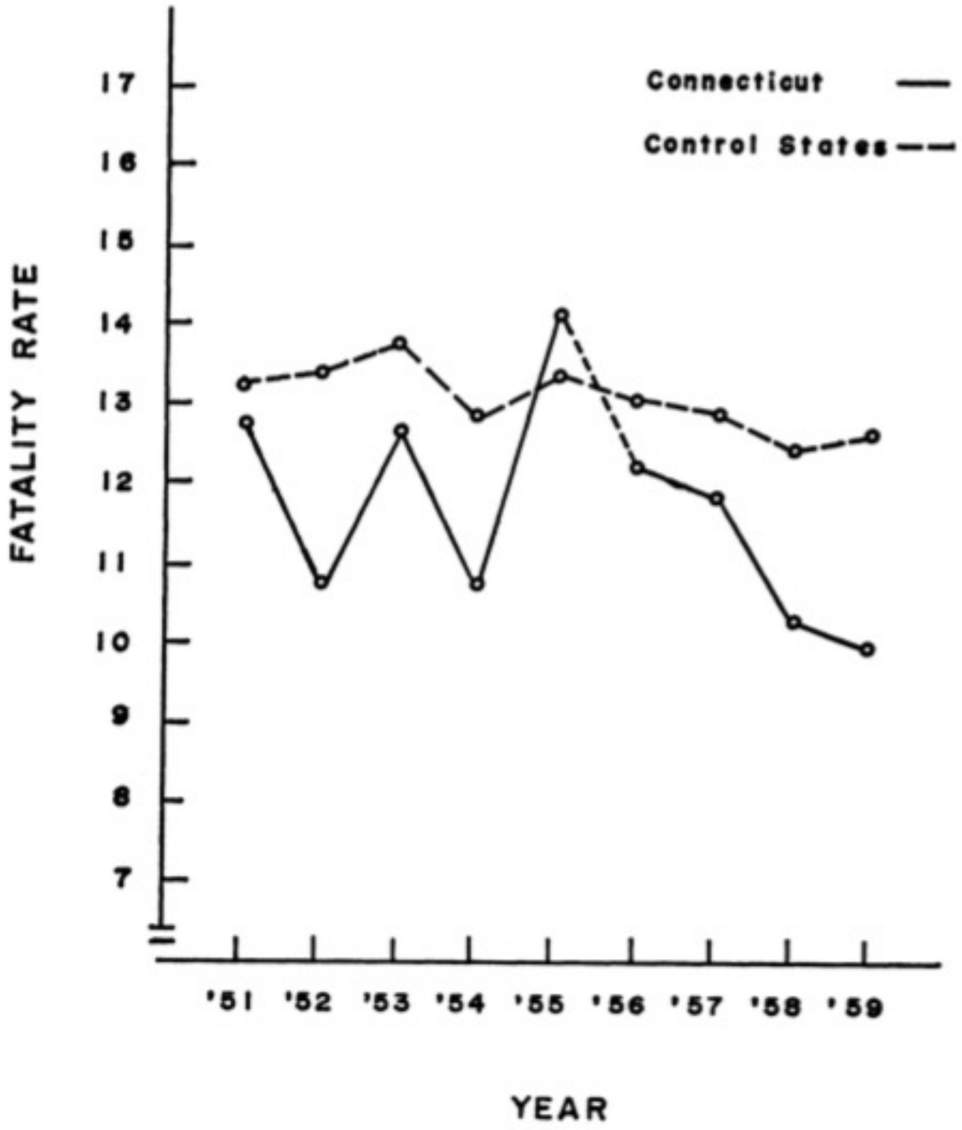


Figure 3. Connecticut and Control States Traffic Fatalities, 1951-1959 (per 100,000 population)

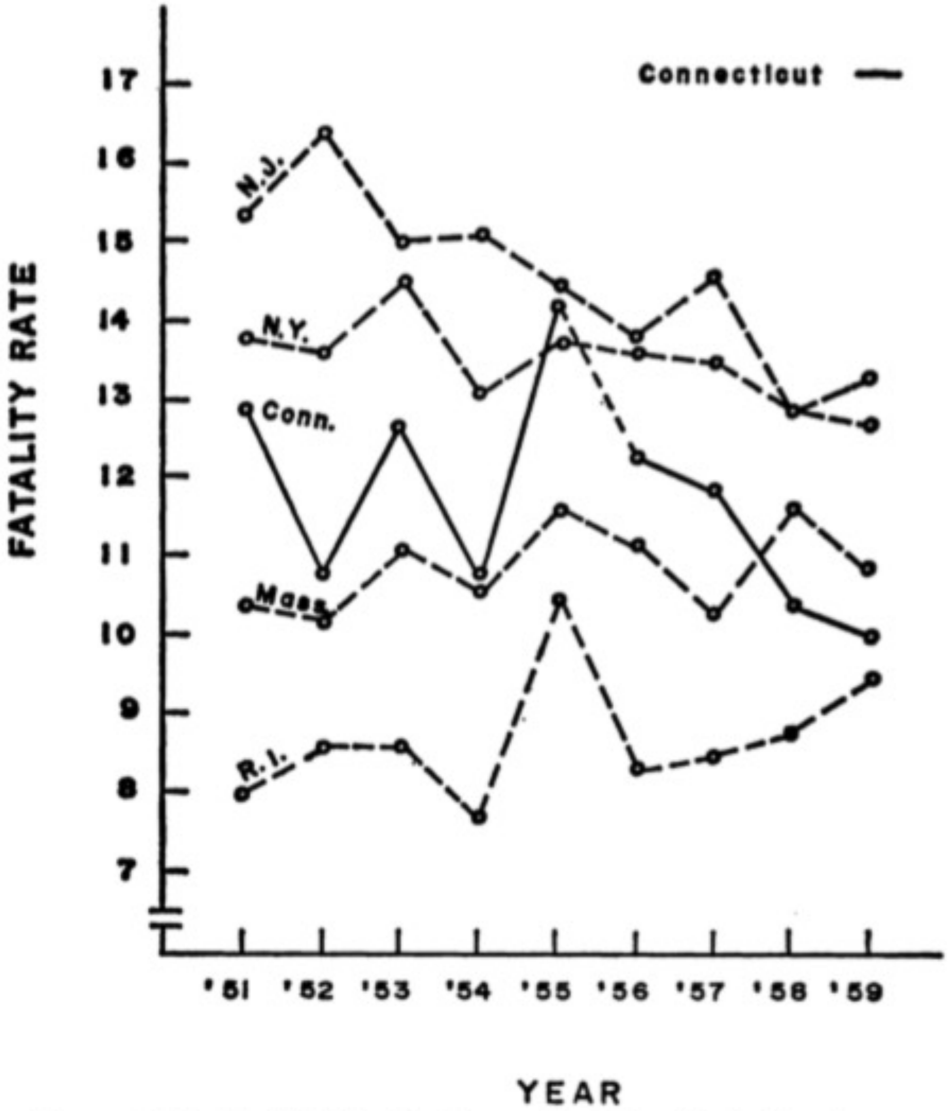


Figure 4. Traffic Fatalities for Connecticut, New York, New Jersey, Rhode Island, and Massachusetts (per 100,000 persons)

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

Quasi-experimental designs

Interrupted time series

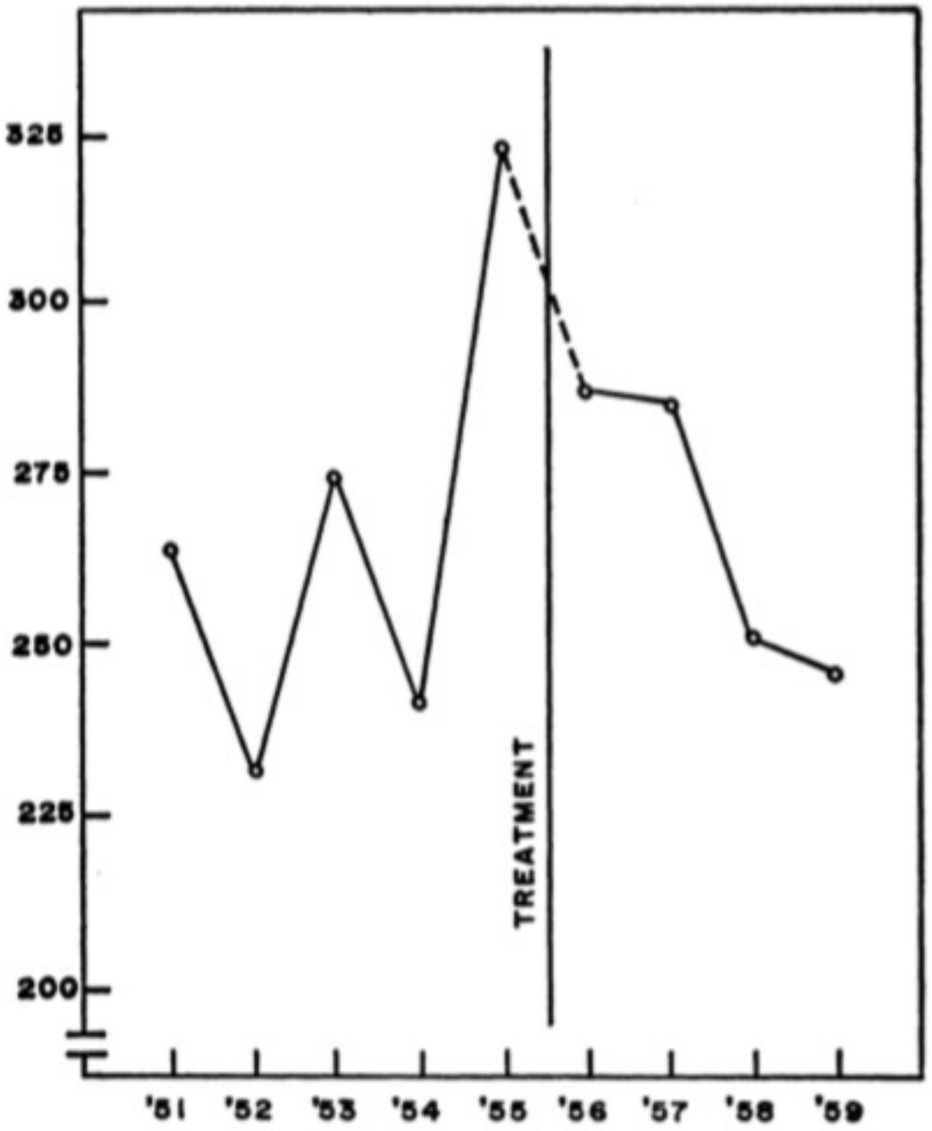
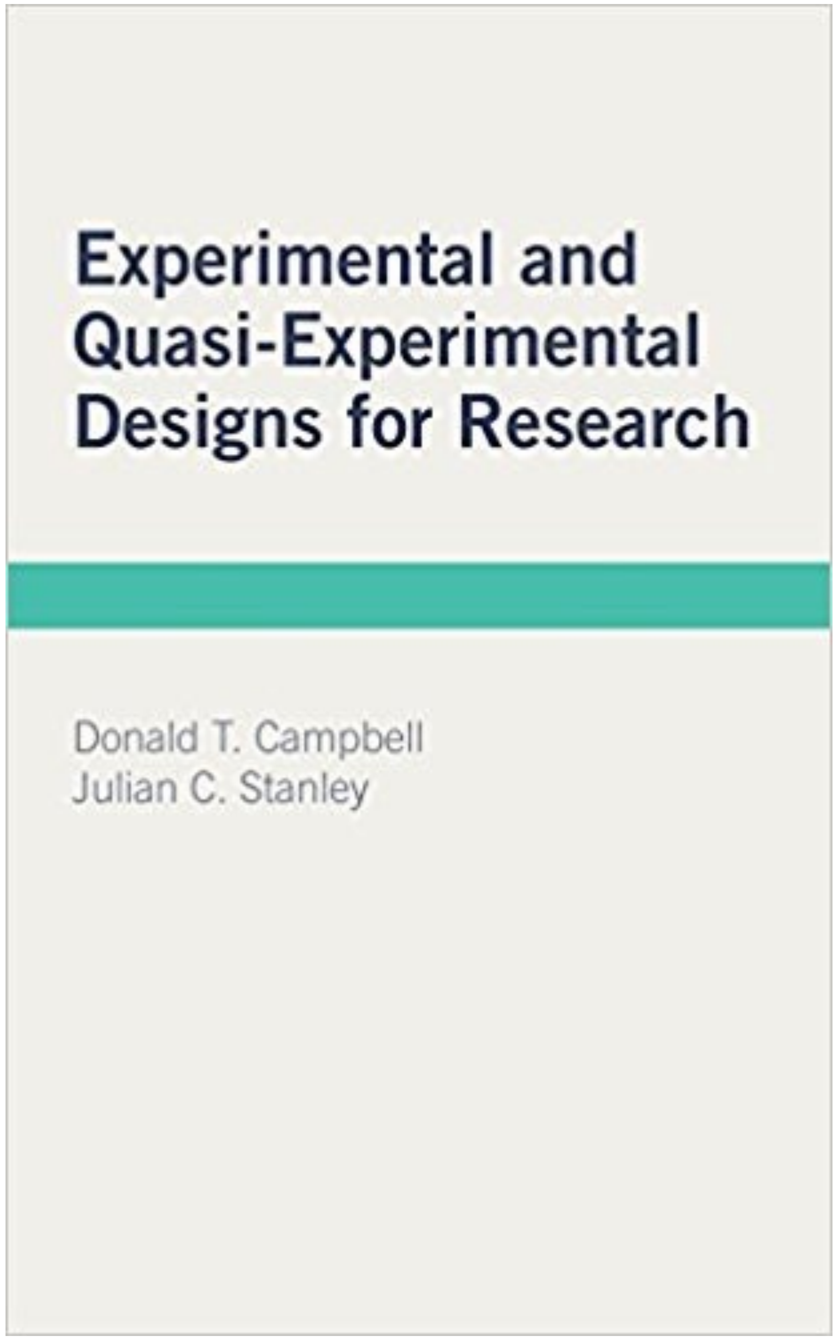


Figure 2. Connecticut Traffic Fatalities, 1951-1959

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

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

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

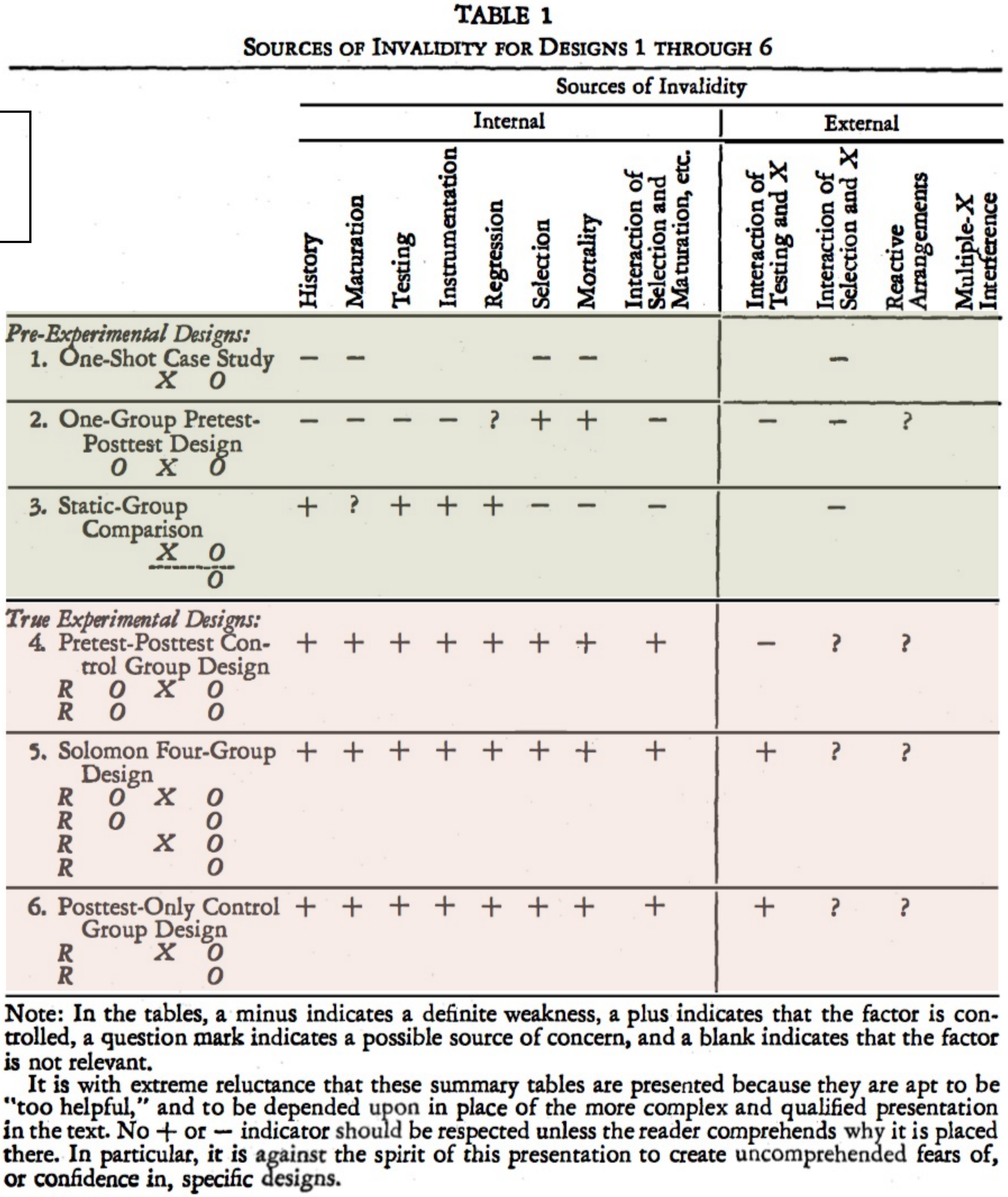


TABLE 1

Sources of Invalidity for Designs 1 through 6

	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
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.

Campbell & Stanley (1963)

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

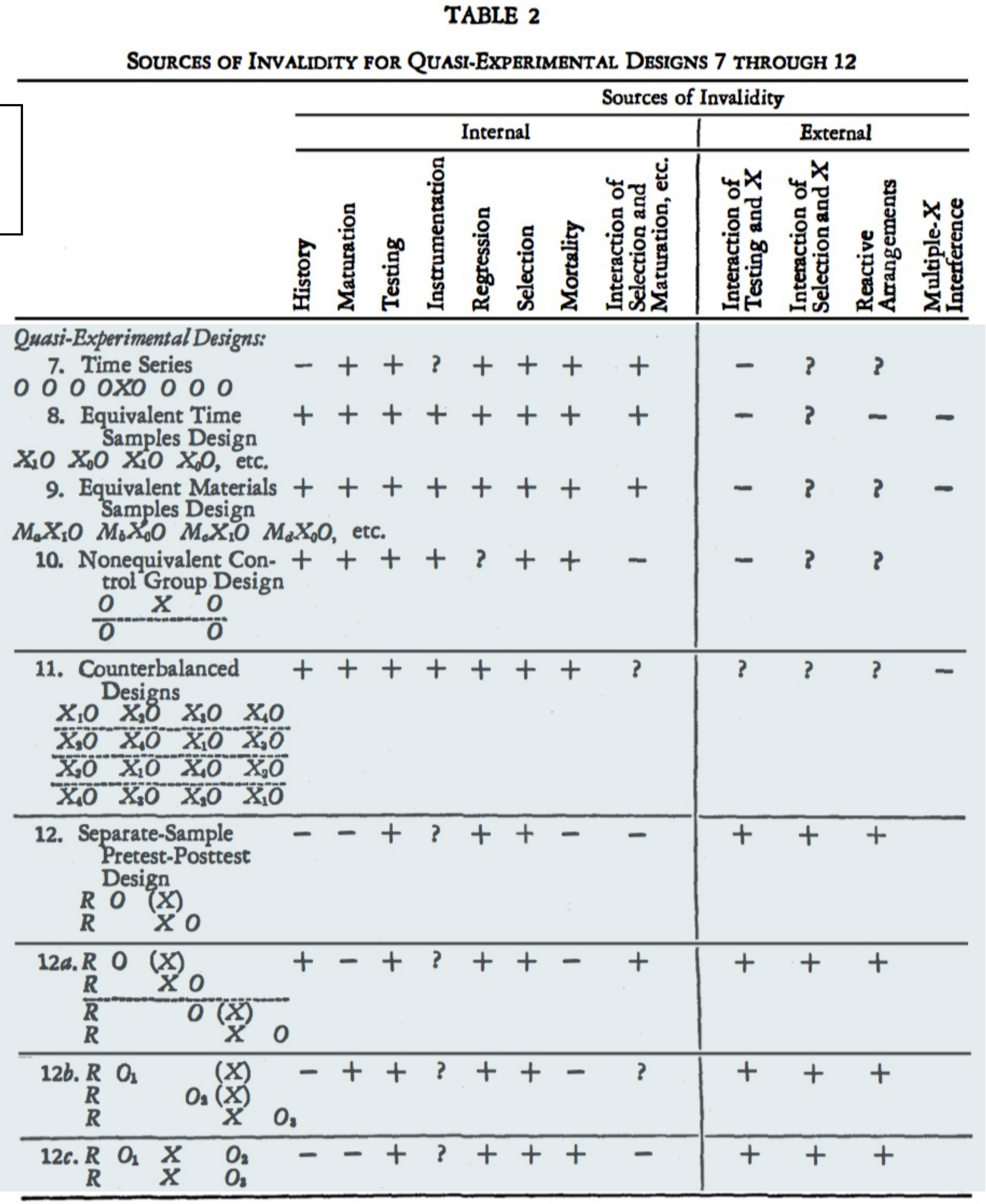


TABLE 2

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_0O X_1O X_0O, etc.	+	+	+	+	+	+	+	+	−	?	−	−
9. Equivalent Materials Samples Design M_aX_1O M_bX_0O M_cX_1O M_dX_0O, 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	−	−	+	?	+	+	−	−	+	+	+	
12*a*. R O (X) R X O - - - - - - - R O (X) R X O	+	−	+	?	+	+	−	+	+	+	+	
12*b*. R O_1 (X) R O_2 (X) R X O_3	−	+	+	?	+	+	−	?	+	+	+	
12*c*. R O_1 X O_2 R X O_3	−	−	+	?	+	+	+	−	+	+	+	

Campbell & Stanley (1963)

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

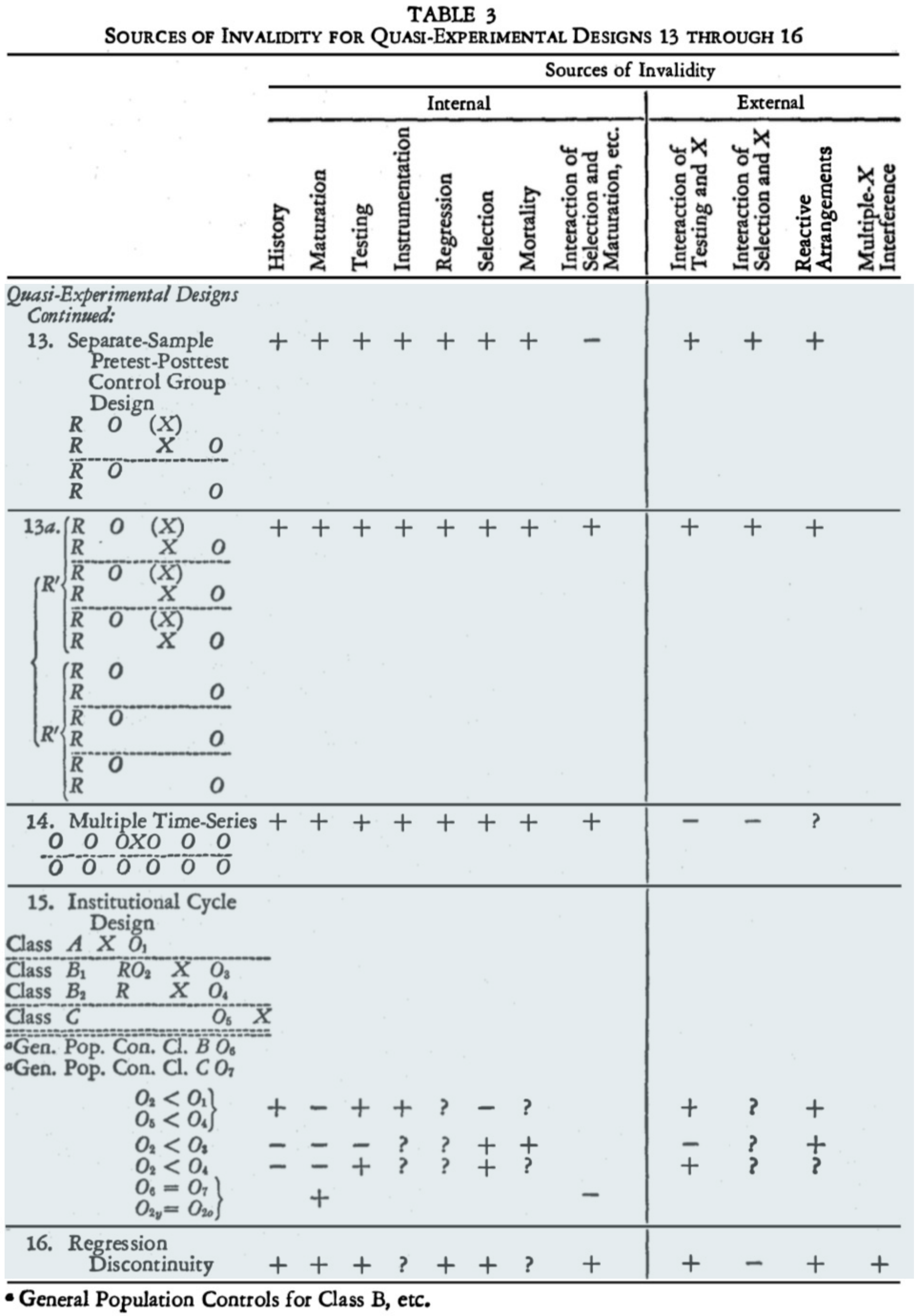


TABLE 3

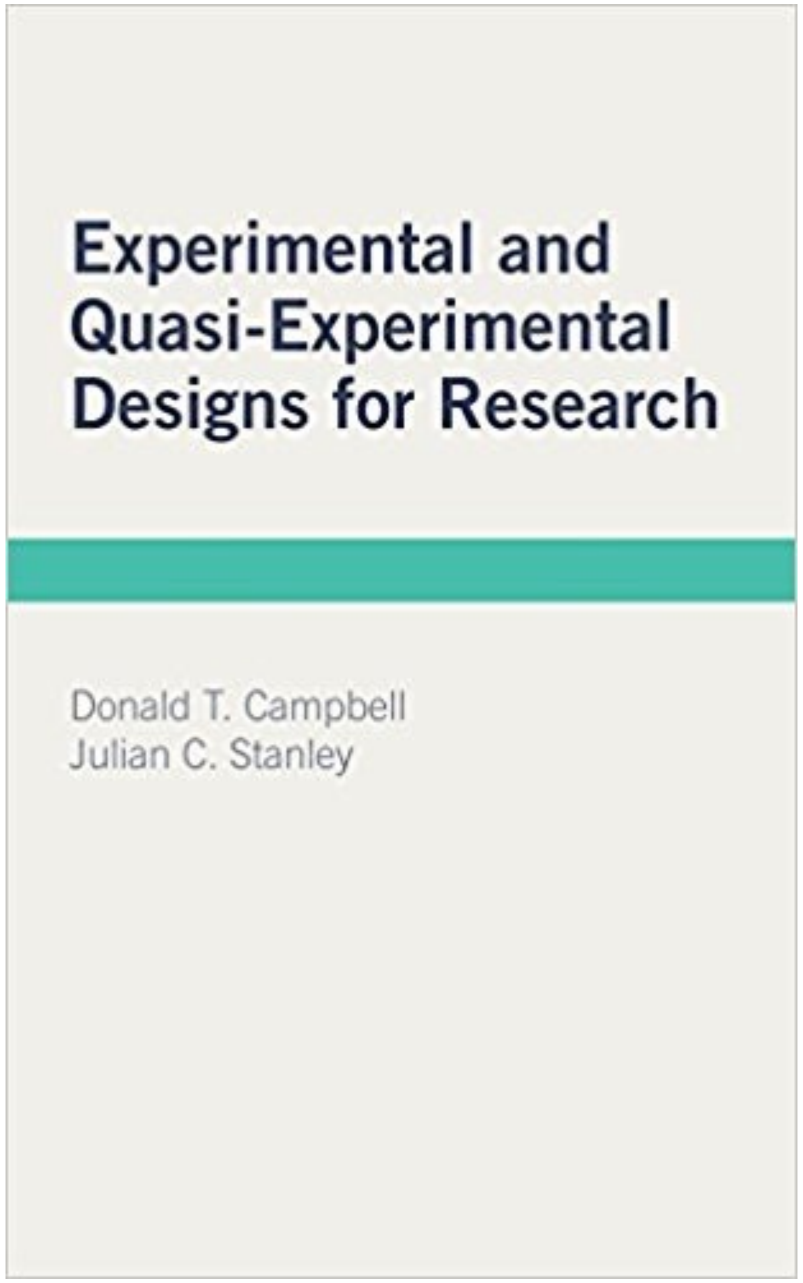
SOURCES OF INVALIDITY FOR QUASI-EXPERIMENTAL DESIGNS 13 THROUGH 16

	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*	+	+	+	+	+	+	+	−	+	+	+	
13*a*. *R O (X)* *R X O* *R′* *R O (X)* *R X O* *R O (X)* *R X O* *R O* *R O* *R′* *R O* *R O* *R O* *R O*	+	+	+	+	+	+	+	+	+	+	+	
14. Multiple Time-Series O O OXO 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)

Experimental and Quasi-experimental Designs



Experimental and Quasi-Experimental Designs for Research

Donald T. Campbell
Julian C. Stanley

“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**.”

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

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.