The Curious Case of Casual Connections: Some ruminations on what do and how we do it.

Dick Campbell

IHRP Chalk Talk

December, 2012

What this talk is about

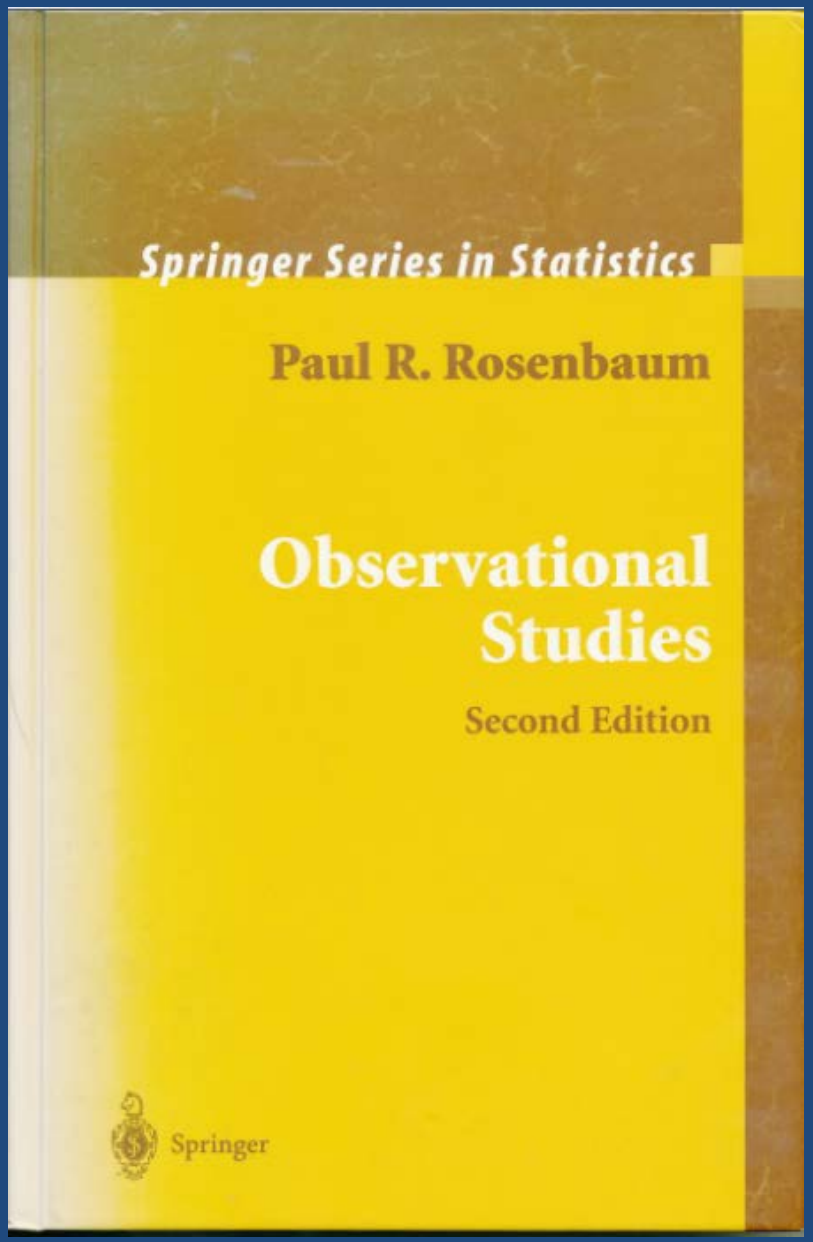
- Like all scientists, we tend to operate on the basis of a set of assumptions and conventions which are usually unstated.
- I am going to review four of them and argue that although they are extremely helpful they have been largely overthrown or at least called into question in such a way that papers and grant proposals based on them may face tough sledding.

Analysis of “observational data”

- A good deal of work at IHRP involves studies in which we wish to draw a causal inference about differences among two or more groups (treatment conditions, races/ethnic groups, clinics) Often, the units of observation (people, classrooms, clinics) have not been randomly assigned to groups.
- Sometimes we have a quasi-experimental design, e.g. interrupted time series from a natural experiment where we can observe an outcome before and after some policy change, e.g. access to free mammograms .
- In the statistical world, such data are referred to as “observational.”

An observational study is one in which “the objective is to elucidate cause and effect relationships [when] it is not feasible to use controlled experimentation … or to assign subjects at random to different procedures. “ William Cochran (1965) as quoted in Paul Rosenbaum’s *Observational Studies* (2002).

There is a well established literature on the design and analysis of observational studies. This book is foundational but at a fairly high mathematical level.

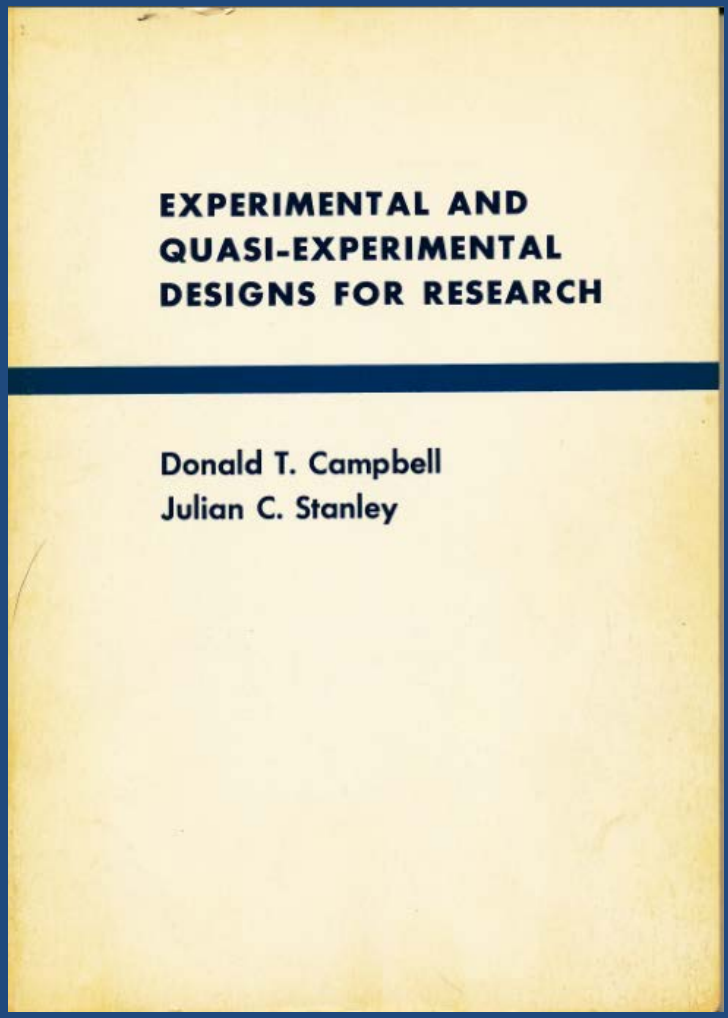


Four ideas which have served us well for nearly forty years

- Quasi-experimental design and evaluation of threats to internal and external validity
- Regression analysis as a general data analytic system
- Structural equation modeling as a way of drawing out the implications of multi-stage theories
- Mediation analysis as a means of estimating direct and indirect effects of presumed causal variables.

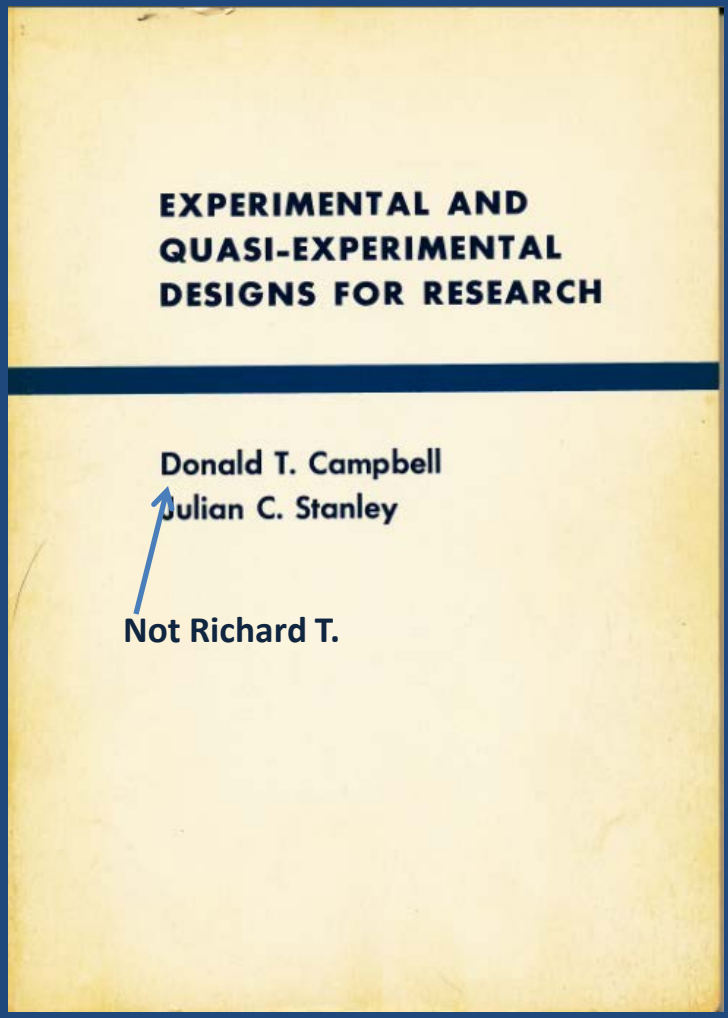
The legacy of Campbell and Stanley

Originally published as a book chapter in 1963 and later as a small book which has gone through numerous printings, this work was profoundly influential. Its impact can be found in methodology texts throughout the social sciences. Many people refer to the concepts without knowing the origin.



The legacy of Campbell and Stanley

Originally published as a book chapter in 1963 and later as a small book which has gone through numerous printings, this work was profoundly influential. Its impact can be found in methodology texts throughout the social sciences. Many people refer to the concepts without knowing the origin.

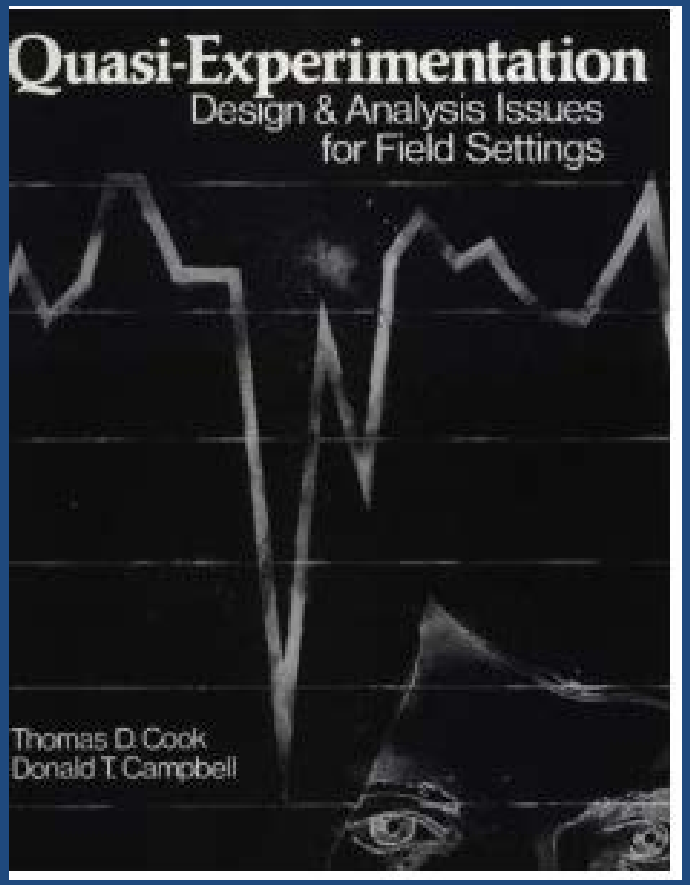


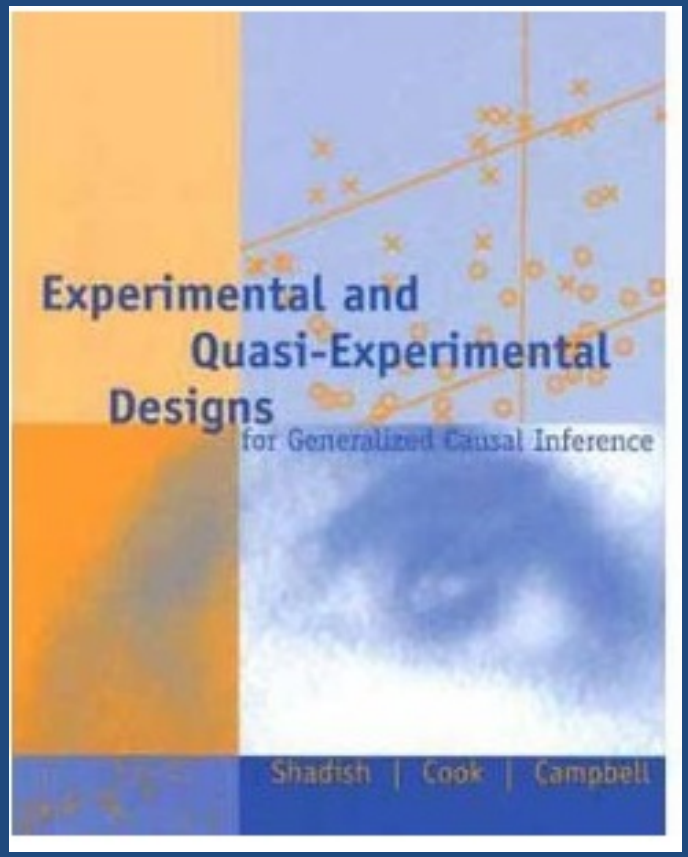
The tradition of laying out threats to validity and designing quasi-experiments to overcome them has provided a way of thinking about design for more than 40 years now. It has, however, been more or less ignored in more recent work by statisticians and economists.

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	+	+	+	+	+	+	+	+	+	?	?	

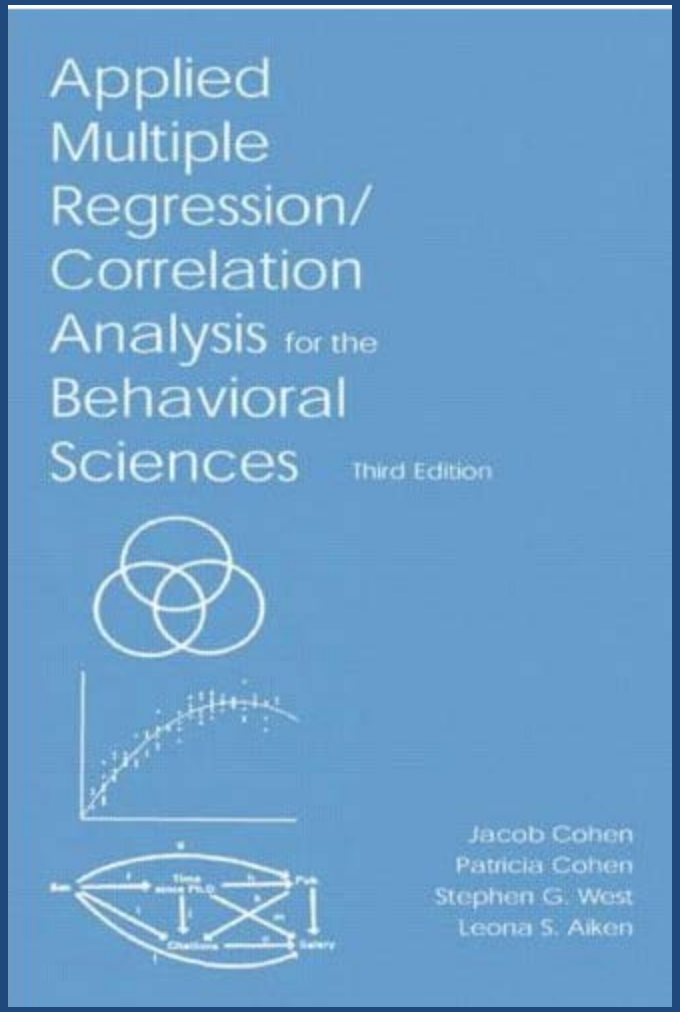




Subsequent editions in 1979 and 2000 greatly expanded on the list of threats and the sophistication of the arguments. The most current edition lists some 37 threats of various kinds. In general, even in the latest edition, the emphasis has been on design and not analysis. This is not where you want to go to learn how to actually do, say, interrupted time series analysis

Multiple regression as a data analytic system

In 1965 Jacob Cohen published a paper with the above title which subsequently became a best selling book. He showed that any analysis of variance or covariance could be written as a multiple regression model. The method deals easily with "unbalanced designs" in which independent variables (factors) are correlated.



The generalized linear model, which extends this notion to outcome variables of virtually any kind is now the way that most data analysis, including longitudinal analysis, is done. Most standard biostatistics texts are now organized on this principle.

It is now routine to run models with more than a score of variables. Indeed, most students these don't learn anything about ANOVA at all. Thus, controlling on multiple variables is trivial.

A simple version of a standard data analysis is shown at the right. We look at a treatment effect, controlling for one or more covariates. But the notion that you can determine a causal impact via regression if you just have enough variables in the model to adjust for pre-intervention group differences is no longer defensible.

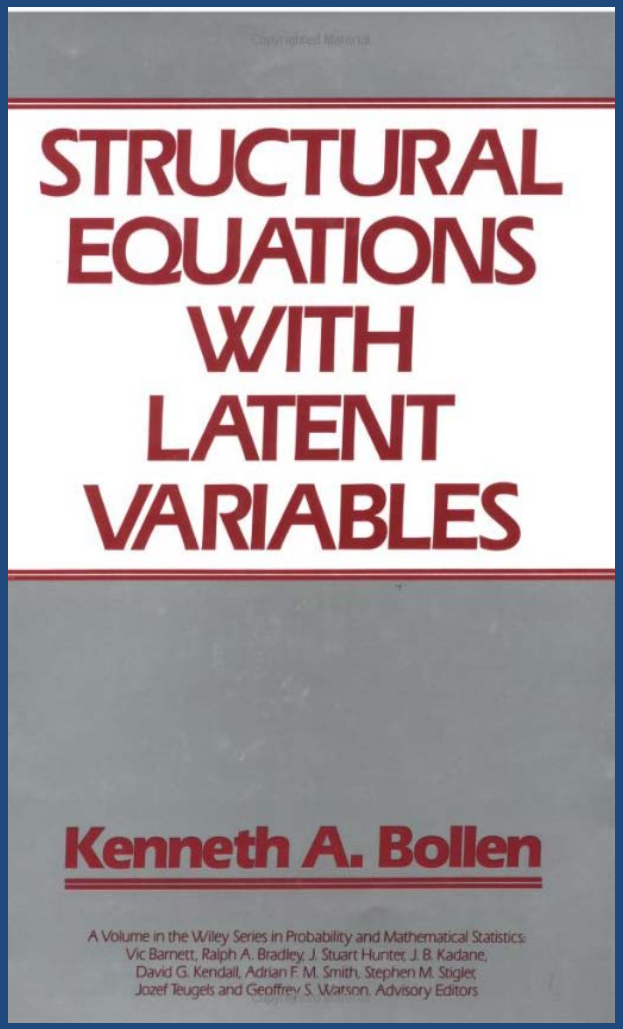
$$\hat{Y}_i = \beta_0 + \beta_1 Treat + \beta_2 Control_1 + \beta_3 Control_2 + ... + \beta_K Control_K + e_i$$

Path analysis and structural equation models as heuristic devices

SEM's allow one to express some very complex ideas with a few boxes, circles and arrows. The approach has enormous heuristic value.

If properly drawn, a path diagram allows one to write the corresponding equations and set up the computer analysis on sight. There is a one to one link between the diagram and the estimation equations.

That said, SEM's are often misunderstood. They *do not* allow one to either determine causality or test for it, although they do allow one to refute casual assertions conditional on the correct specification of the rest of the model. Many SEM's make causal assertions which are very difficult to defend.



An example of a SEM from a recent paper in the *Journal of Health and Social Behavior*

Path analysis and structural equation models, and the whole "causal analysis" tradition is under serious challenge. The essential argument is that one can't merely assert a causal relationship, it has to be tested in some way.

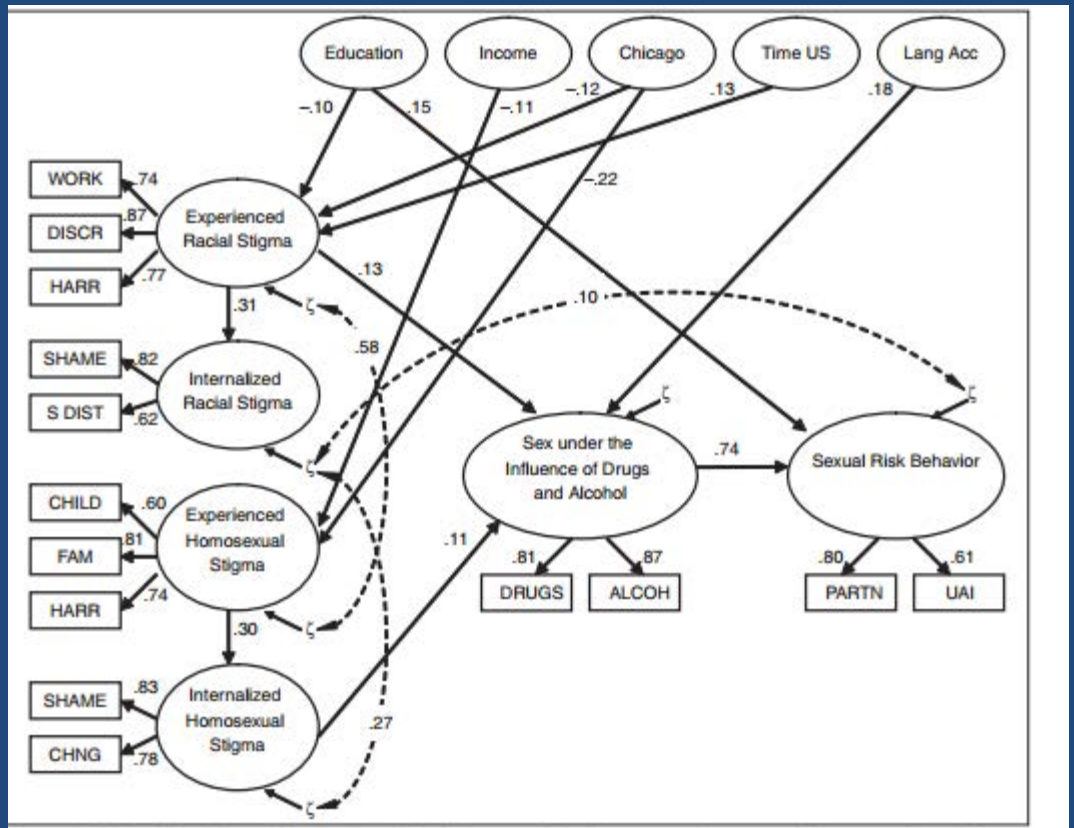


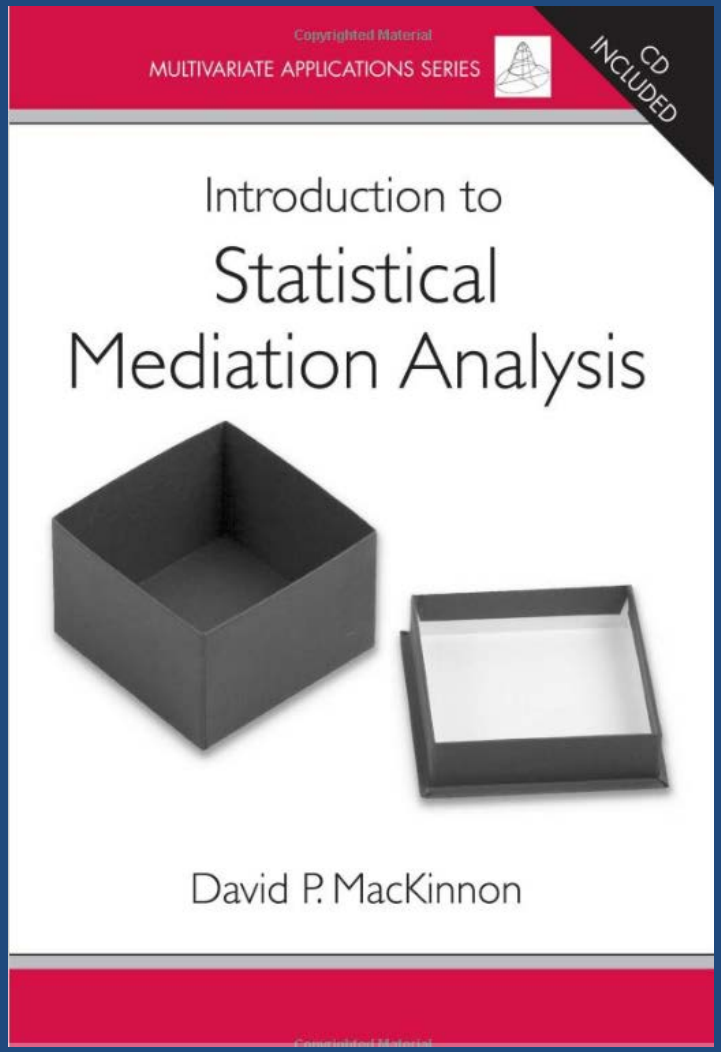
Figure 2. Final Model and Standardized Coefficients of the Association between Homosexual and Racial Stigmatization on Sexual Risk Behavior (all significant paths at $p \leq .05$)[a]

[a]Correlations among exogenous variables not shown.

Source: Valles, J. R., Kuhns, L. M, Campbell, R. T. and Diaz, R. M. 2010. Social Integration and Health: Community Involvement, Stigmatized Identities, and Sexual Risk in Latino Sexual Minorities. *Journal of Health and Social Behavior 51*; 30-47.

Mediation analysis

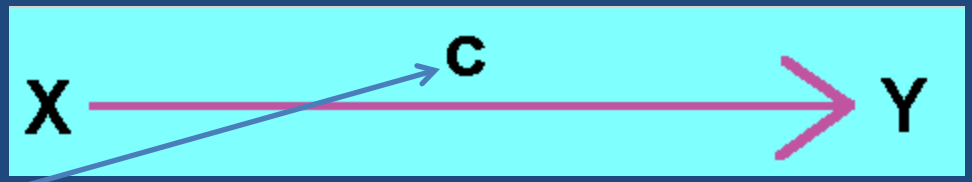
A good deal of what we do at IHRP involves attempting to change some belief, attitude or behavior with the intent of influencing some downstream health outcome, e.g. dietary behavior and obesity. We frequently ask if the variable we are attempting to change directly serves as a mediator between an intervention (either randomly or non-randomly allocated) and the outcome we are interested in.

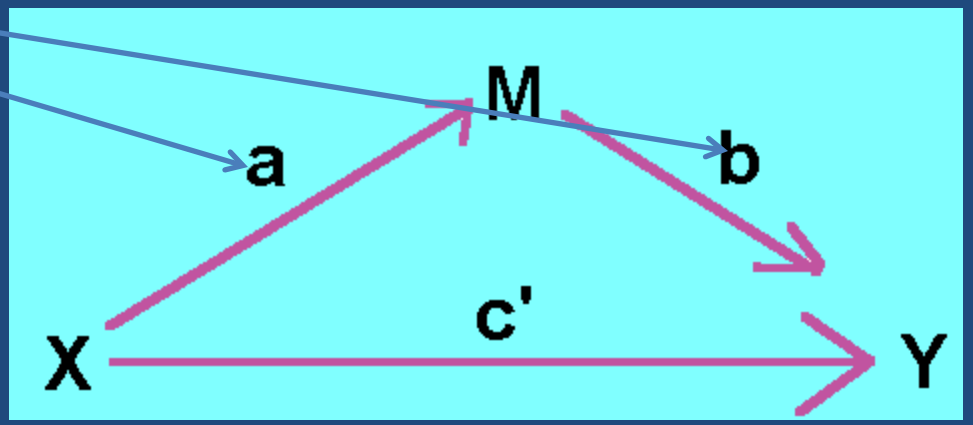


The diagrams at the right shows the classic mediation set up. According to David Kenny and others, you can estimate the proportion of the effect of X on Y which is mediated by M as:

c – ab.

This approach is certainly causal, but many influential commentators (e.g. Bengt Muthén, who created MPlus) now agree that even in randomized designs this decomposition is faulty.





I will return to this issue later.

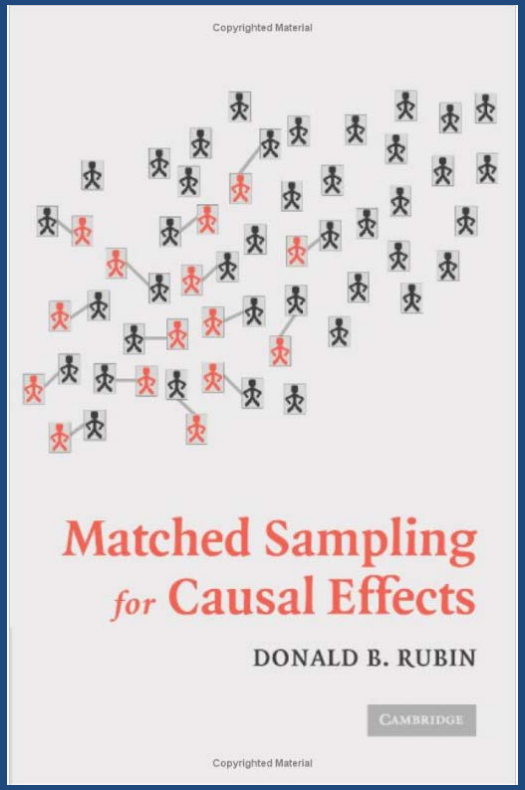
Source: http://davidakenny.net/cm/mediate.htm#IE

The times they are a-changin

- Thus, we are in the midst of a multifaceted revolution, by no means complete, which is changing what is considered acceptable by grant and journal reviewers. Almost every aspect of what we have taken for granted for a long time has come into question.
- In particular, traditional ways of inferring and estimating causal effects are under serious challenge. Our four pillars have become somewhat shaky.

Rubin's approach

- A great deal of what might be called "new causal analysis" flows from the work of Donald B. Rubin.
- What has come to be known as "Rubin's causal model", also known as the "potential outcomes approach" or the "counter factual approach," appeared in the early 1970's and finally gained traction in the past five to ten years .



The Rubin perspective on an observational study

- Imagine two groups – treated and control – to which individuals have been randomly assigned.
- We administer some treatment to one of the groups.
- We know that the effect of treatment will vary across individuals. We know that within each group there is variance in the outcome variable.
- It may be that some of the variation is systematic in that persons with particular characteristics might react differently to the treatment.
- The fact that we have randomized means that the difference, whatever the source, washes out and the difference between the two group means is an unbiased estimate of the treatment effect.

- Suppose, however, that treatment assignment is non-random. To consider the implications, Rubin tells us to think of each individual as having two potential outcomes, one in the treatment group and one in the control. But a person can only be in one group, hence one of the outcomes is "counterfactual."
- For a given person, we can think of $Y_i^{(T)} - Y_i^{(C)}$, the difference between that person's score in the two conditions. But unless we have a cross over design, which presents its own problems, that is impossible. We can only observe each person under one condition as shown in the table on the next slide.

The “fundamental problem of causal inference.”

Table 1
Potential and Observed Outcomes

	Potential outcomes		Observed outcomes	
Participant	*T*	*C*	*T*	*C*
1	10	10	10	■
2	11	13	■	13
3	11	11	■	11
4	12	16	■	16
5	12	12	12	■
6	12	15	12	■
7	12	13	■	13
8	13	15	13	■
9	13	17	■	17
10	14	18	14	■
	True average treatment effect: 2.0		Prima facie average treatment effect: 1.8	

Note. The columns labeled “Potential outcomes” illustrate the true responses of the 10 participants under the treatment (*T*) and control (*C*) conditions. The columns labeled “Observed outcomes” illustrate the responses of the same 10 participants in a randomized experiment or an observational study. A ■ indicates that the response was not observed. Half of the potential outcomes are not observed. The prima facie average treatment effect is the simple (possibly biased) difference between the observed means in the *T* and *C* groups.

Source: West and Thoemmes, 2010

- An important component of Rubin's argument is that a given case might respond differently to being in a particular group than another case, as a function of particular unmeasured covariates. For example, persons who self select into a non-treated group might not do as well in the treated group, if they were in it, as those who self-select into it.

- If we are willing to suspend disbelief for a moment, the best estimate of casual effect in a non-randomized study would be:

$$(\overline{Y}^{t}_{i\in t} - \overline{Y}^{c}_{i\in t}) + (\overline{Y}^{t}_{i\in c} - \overline{Y}^{c}_{i\in c})$$

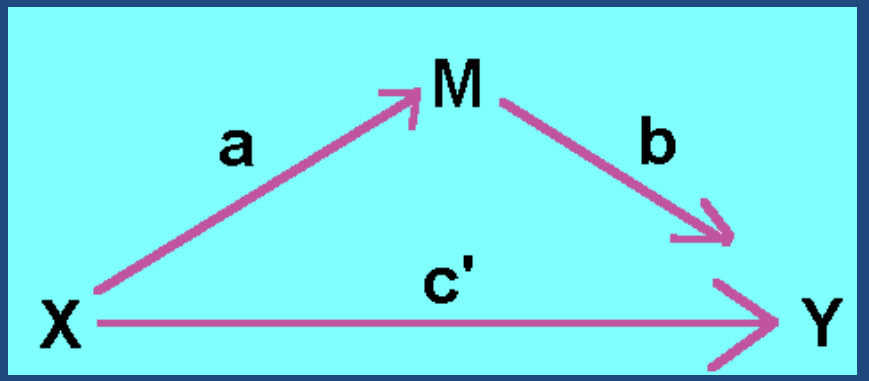
Counterfactual

Stable unit treatment value (SUTVA) assumption

- SUTVA is the a-priori assumption that the value of Y for unit *i* when exposed to treatment *t* will be the same no matter what mechanism is used to assign treatment t to unit *i* and no matter what treatments the other units receive.
- The assumption may seem innocuous, but it has wide ranging implications.

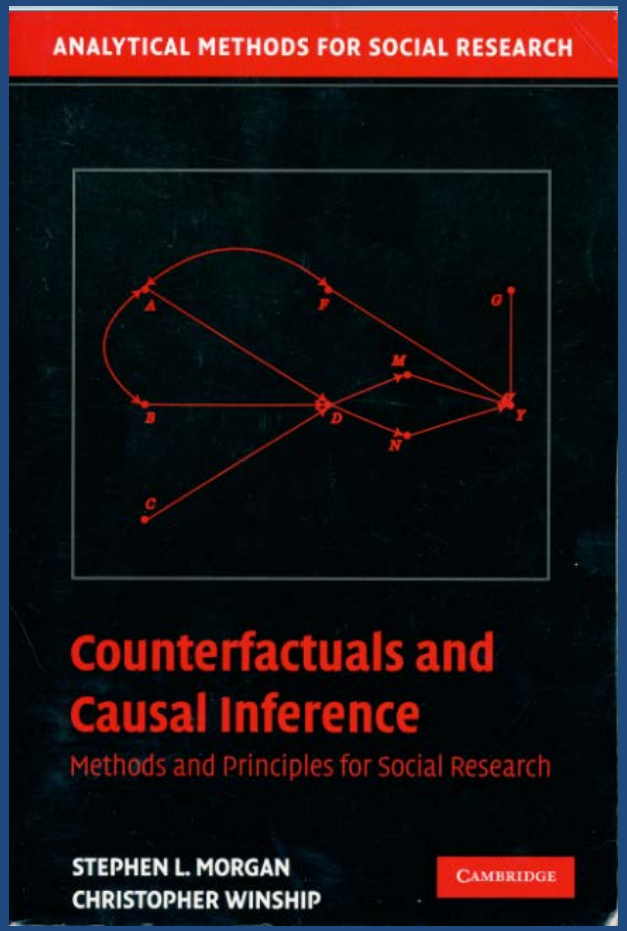
Back to mediation

- In the classic mediation model, shown previously, the same issues apply if X is a non-randomly assigned treatment. Again, you can't just dump a load of covariates into your model.
- Some critics, particularly Michael Sobel (2008), argue that even if X *is* randomly assigned, M, the mediator, is not. Hence, mediation and the usual effect decomposition, can not be interpreted in casual terms.
- This issue is by no means resolved but if you do classical mediation analysis you may well be challenged by reviewers.

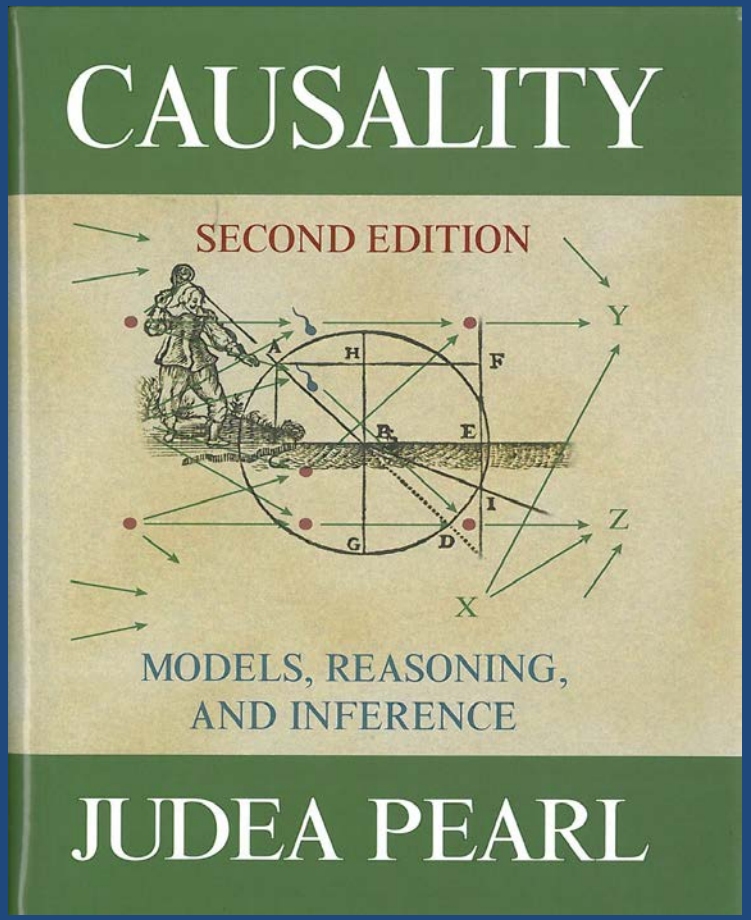


Two very important books

2007



2010 (2000)

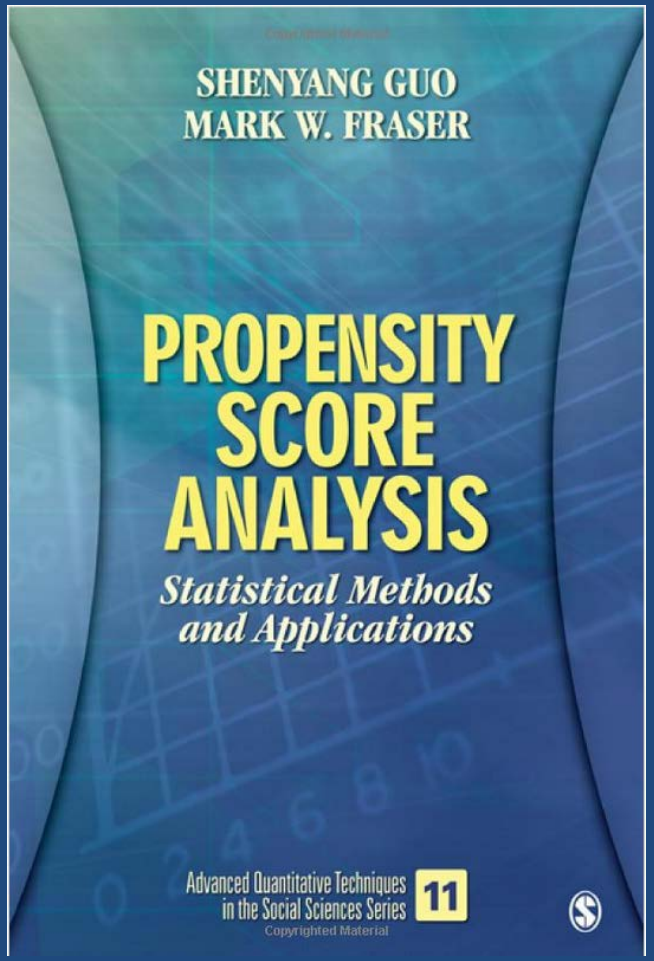


Rubin uses this approach to derive ways of estimating causal effects under various assumptions. For example, one can estimate, in some case ATT, the treatment affect among the treated. We will skip all of this.

Matching and propensity scores

- Rubin's much preferred solution to the problem of non-equivalent groups is matching.
- Campbell and Stanley explicitly rejected matching primarily because of the difficulty of matching on multiple variables.
- Rubin's solution is to estimate the probability of each case being in the treatment group as a function of as many covariates as he can get. You then end up with a single covariate that carries information from a lot of variables.
- The estimated probability is referred to as a propensity score.
- It is crucial that the p-scores be balanced across groups, i. e., that the distribution of p scores be the same in both groups.

- There are numerous ways to do propensity score analysis and a good deal of dispute as to which is best. This 2010 book is a very good summary of current work. There is a web site that shows many worked examples.



An example based on work in progress

- Dick Warnecke et al are interested in assessing the effects of living in a medically underserved area (MUA) on late stage breast cancer diagnosis.
- An MUA gets various kinds of federal support for access to medical care. Eligibility is determined based on certain criteria.
- Some areas of the city are eligible in that they meet the criteria but not designated because designation is not automatic; it requires local action. They are the comparison group.
- Persons living in the two areas may differ on various covariates so we used a propensity score to match them.

Matching variables thus far

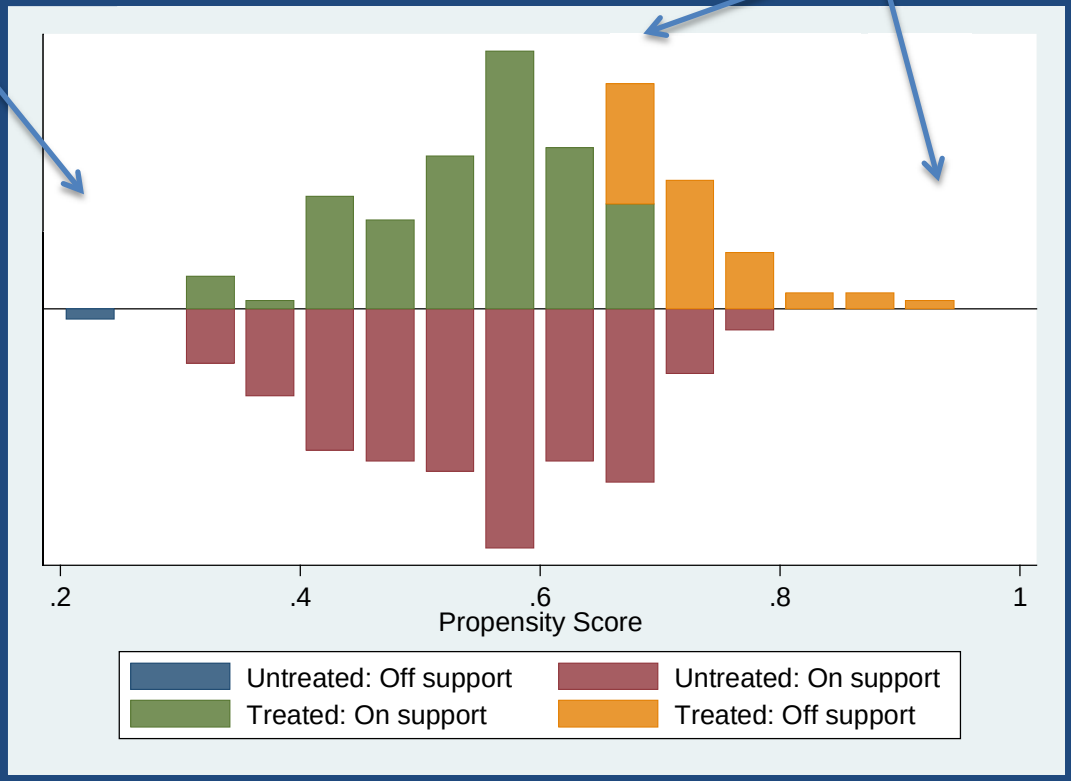
- Education
- Income
- Age
- Prior anomalous mammogram findings
- Number of co-morbidities

- Numerous other variables, e.g. network structure will be added

Matching Results

Unmatched cases eliminated from comparison group

Unmatched cases eliminated from treatment group



Results For Late Stage Dx

Logistic regression

Number of obs	= 229
LR chi2(1)	= 4.24
Prob > chi2	= 0.0395
Log likelihood = -156.24199	Pseudo R2 = 0.0134

pstages2xneo	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_treated	.5782732	.1545645	-2.05	0.040	.3424669	.9764446
_cons	1.169811	.2188421	0.84	0.402	.8107324	1.687929

Computations and graphs done using PSMATCH2 in Stata.

Other analytic approaches

- Primarily within economics, but also in policy analysis and epidemiology there has been a lot of work using methods other than matching to deal with observational data.
  - The regression discontinuity design (#16 in C&S) has been formalized and used extensively.
  - The instrumental variables method has been used to model selection into groups.
  - Difference in differences, a kind of interaction test, has been applied to interrupted time series and other designs.

Coming full circle

- New developments in statistical approaches to observational data have mostly ignored the Campbell-Stanley tradition.
- Rubin cites C&S but as a pro-forma exercise.
- Economists tend to ignore it entirely.
- But recently, there has been a bit of a rapprochement.
- Schadish has published an important paper with an excellent commentary by West and a student of his.

Psychological Methods
2010, Vol. 15, No. 1, 3–17

1082-989X/10/$12.00 DOI: 10.1037/a0015916

Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings

William R. Shadish
University of California, Merced

This article compares Donald Campbell's and Donald Rubin's work on causal inference in field settings on issues of epistemology, theories of cause and effect, methodology, statistics, generalization, and terminology. The two approaches are quite different but compatible, differing mostly in matters of bandwidth versus fidelity. Campbell's work demonstrates broad narrative scope that covers a wide array of concepts related to causation, with a powerful appreciation for human fallibility in making causal judgments, with a more elaborate theory of cause and generalization, and with a preference for design over analysis. Rubin's approach is a more narrow and formal quantitative analysis of effect estimation, sharing a preference for design but best known for analysis, with compelling quantitative approaches to obtaining unbiased quantitative effect estimates from nonrandomized designs and with comparatively little to say about generalization. Much could be gained by joining the emphasis on design in Campbell with the emphasis on analysis in Rubin. However, the 2 approaches also speak modestly different languages that leave some questions about their total commensurability that only continued dialogue can fully clarify.

Keywords: Rubin's causal model, causal inference, validity, SUTVA, propensity score

A response by Rubin is almost a *mea culpa*

Psychological Methods
2010, Vol. 15, No. 1, 38–46

1082-989X/10/$12.00 DOI: 10.1037/a0018537

Reflections Stimulated by the Comments of Shadish (2010) and West and Thoemmes (2010)

Donald B. Rubin
Harvard University

This article offers reflections on the development of the Rubin causal model (RCM), which were stimulated by the impressive discussions of the RCM and Campbell's superb contributions to the practical problems of drawing causal inferences written by Will Shadish (2010) and Steve West and Felix Thoemmes (2010). It is not a rejoinder in any real sense but more of a sequence of clarifications of parts of the RCM combined with some possibly interesting personal historical comments, which I do not think can be found elsewhere. Of particular interest in the technical content, I think, are the extended discussions of the stable unit treatment value assumption, the explication of the variety of definitions of causal estimands, and the discussion of the assignment mechanism.

But there's more to do

American Journal of Epidemiology

DOI: 10.1093/aje/kwq084
Advance Access publication: June 14, 2010

Previous

Practice of Epidemiology

Generalizing Evidence From Randomized Clinical Trials to Target Populations

The ACTG 320 Trial

Stephen R. Cole* and Elizabeth A. Stuart

* Correspondence to Dr. Stephen R. Cole, Department of Epidemiology, Gillings School of Global Public Health and Center for AIDS Research, CB7435, University of North Carolina, Chapel Hill, NC 27599 (e-mail: cole@unc.edu).

Initially submitted October 21, 2009; accepted for publication March 24, 2010.

Properly planned and conducted randomized clinical trials remain susceptible to a lack of external validity. The authors illustrate a model-based method to standardize observed trial results to a specified target population using a seminal human immunodeficiency virus (HIV) treatment trial, and they provide Monte Carlo simulation evidence supporting the method. The example trial enrolled 1,156 HIV-infected adult men and women in the United States in 1996, randomly assigned 577 to a highly active antiretroviral therapy and 579 to a largely ineffective combination therapy, and followed participants for 52 weeks. The target population was US people infected with HIV in 2006, as estimated by the Centers for Disease Control and Prevention. Results from the trial apply, albeit muted by 12%, to the target population, under the assumption that the authors have measured and correctly modeled the determinants of selection that reflect heterogeneity in the treatment effect. In simulations with a heterogeneous treatment effect, a conventional intent-to-treat estimate was biased with poor confidence limit coverage, but the proposed estimate was largely unbiased with appropriate confidence limit coverage. The proposed method standardizes observed trial results to a specified target population and thereby provides information regarding the generalizability of trial results.

bias; bias (epidemiology); causal inference; external validity; generalizability; randomized trials; standardization

American Journal of Epidemiology

DOI: 10.1093/aje/kwq084
Advance Access publication:
June 14, 2010

Practice of Epidemiology

Generalizing Evidence From Randomized Clinical Trials to Target Populations

The ACTG 320 Trial

Stephen R. Cole* and Elizabeth A. Stuart

* Correspondence to Dr. Stephen R. Cole, Department of Epidemiology, Gillings School of Global Public Health and Center for AIDS Research, CB7435, University of North Carolina, Chapel Hill, NC 27599 (e-mail: cole@unc.edu).

Initially submitted October 21, 2009; accepted for publication March 24, 2010.

Properly planned and conducted randomized clinical trials remain susceptible to a lack of external validity. The authors illustrate a model-based method to standardize observed trial results to a specified target population using a seminal human immunodeficiency virus (HIV) treatment trial, and they provide Monte Carlo simulation evidence supporting the method. The example trial enrolled 1,156 HIV-infected adult men and women in the United States in 1996, randomly assigned 577 to a highly active antiretroviral therapy and 579 to a largely ineffective combination therapy, and followed participants for 52 weeks. The target population was US people infected with HIV in 2006, as estimated by the Centers for Disease Control and Prevention. Results from the trial apply, albeit muted by 12%, to the target population, under the assumption that the authors have measured and correctly modeled the determinants of selection that reflect heterogeneity in the treatment effect. In simulations with a heterogeneous treatment effect, a conventional intent-to-treat estimate was biased with poor confidence limit coverage, but the proposed estimate was largely unbiased with appropriate confidence limit coverage. The proposed method standardizes observed trial results to a specified target population and thereby provides information regarding the generalizability of trial results.

bias; bias (epidemiology); causal inference; external validity; generalizability; randomized trials; standardization