PROJECT MUSE®

Hill and Campbell: Separate or symbiotic paradigms?

William H. Yeaton

Observational Studies, Volume 6, Issue 2, 2020, pp. 57-65 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2020.0010>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/793351/summary>

Hill and Campbell: Separate or symbiotic paradigms?

William H. Yeaton
Measurement and Statistics program
Florida State University
Tallahassee, FL 32306, U.S.A.

bill.yeaton@yahoo.com

1. Brief Introduction to Hill

At its heart, A.B. Hill's 1965 Presidential address includes a set of nine "viewpoints" that, if judged to be present, increase the chance that a causal claim is correct. When taken in this light, as Hill cautions, the nine viewpoints are neither necessary nor sufficient for cause. Instead, they might be thought of as attendant features whose presence will enhance causal inference.

What I do not believe – and this has been suggested – is that we can usefully lay down some hard-and-fast rules of evidence that must be observed before we accept cause and effect. None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*. What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question – is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?

(Hill, 1965, p. 299)

2. Introduction to the Campbellian paradigm: Two symbols and the truth¹

A second system of inquiry, elaborated by Donald Campbell and his colleagues (Shadish, Cook and Campbell, 2002), also informs casual inference but does so from what is essentially an opposing perspective. It asks if one or more of a lengthy list of internal validity threats are present, thereby decreasing the probability of a correct causal claim. To be precise, Shadish, Cook, and Campbell list nine internal validity threats (p. 55), though one of these, "ambiguous temporal precedence," is the same as "temporality" in the Hillian paradigm. In addition to these threats to causal inference, four "interaction with selection" threats (selection by history, by maturation, by instrumentation, and by regression) are elaborated, along with four threats to construct validity (compensatory equalization, compensatory rivalry, resentful demoralization, and treatment diffusion) which were previously listed as internal validity threats in Cook and Campbell (1979).

Campbell's primary mechanism for ruling out internal validity threats is an individual research design. While this research design-internal validity threats nexus is rich and de-

¹ "Two symbols" refers to Xs and Os, from which many research designs are created.

tailed, Campbell and his colleagues' discussion of different design options and their role in eliminating validity threats is beyond the scope of this paper. Here, the central point is that the Campbellian approach looks for negative study features and aims to rule them out, while the Hillian paradigm looks for positive study features and tries to rule them in. For Hill, the researcher's task is to demonstrate that favorable conditions exist; for Campbell, the task is to demonstrate that unfavorable conditions do not exist.

3. Combine Hill and Campbell paradigms?

It is difficult to argue that one of these paradigm's approach to truth is intrinsically better than the other. Certainly, researchers' tasks are different, and each approach ultimately requires judgment as neither process provides a "for sure" determination of cause. Classic Campbellian verbiage demands that we "render less plausible" a particular validity threat. A Campbellian might ask if selection bias is likely absent in a particular observational study, while a Hillian might ponder whether sufficient specificity is present to infer cause. Journal reviewers might reasonably disagree about the absence of selection bias or the presence of specificity.

I maintain that these alternative approaches, foundationally based upon very different logical and disciplinary orientations, can advance science more than either, alone. Below, I provide examples that show how each of Hill viewpoints can be amplified using the Campbellian paradigm. The "more is better" argument is not new; as Feyerabend long ago is reported to have said "Let a thousand methods bloom" (e.g., Heit, 2016). Restricting different methodological principles to particular research orientations (isolation) leads to fewer insights. A hybrid orientation (symbiosis) would spark cross-disciplinary collaboration via methodological pluralism, allowing the Hill and Campbell approaches to be simultaneously implemented and enabling each Hill principle to be empirically embellished using elements of the Campbellian approach.

Unfortunately, mutually beneficial symbiosis is not the usual way of science (Campbell, 2005). There are many precedents in which researchers have subscribed to apparently adversarial approaches and acted as rivals. Friction between Bayesians and frequentists is longstanding. Those who utilize the causal diagrams of Judea Pearl (2018) and those who follow the potential outcomes, statistical modelling approach of Don Rubin (1974) are less likely to read the same studies or to conduct research in the same way.

I assert that few social scientists are familiar with Hill's paradigm. Consistent with this claim is the fact that neither the prominent, Campbellian-based textbooks (Campbell and Stanley, 1966; Cook and Campbell, 1979; Shadish, Cook, and Campbell, 2002) nor a recently published quasi-experimental design text by a Campbellian disciple (Reichardt, 2019) cite Hill's Presidential address. In a similar vein, Hill would have been unfamiliar with the Campbellian paradigm. Near the time of Hill's 1965 Presidential address, few of Campbell's design-based papers had been published (Campbell, 1957; Campbell and Stanley, 1966). And while Campbell-based methods have increasingly diffused into the medical and health areas, their spread is far from complete.

Consistent with methodological pluralism, I claim that the Campbellian approach, particularly those dimensions that consider the quality of different research designs or implement multiple design elements, can directly inform Hill guidelines. The juxtaposition of

Campbell’s design thinking as a means to rule out threats to internal validity can serve as a powerful adjunct to the nine guidelines that Hill aims to rule in. In addition, to my knowledge, no one has directly addressed the Hillian approach from the Campbellian perspective using research designs as the bridging process. In short, the explicit methods that Campbell uses to rule out internal validity threats provide an empirical basis for gauging the presence of the Hill viewpoints.

4. Using both single and multiple design orientations to embellish Hill’s viewpoints

To illustrate how the Campbellian approach might inform Hill’s viewpoints, I provide examples that follow either the usual practice of incorporating a single research design or a more novel approach that utilizes multiple research designs (Yeaton, 2019). I argue that this context of single and multiple designs creates a better, value-added basis to evaluate Hill’s viewpoints.

While this multiple design strategy is structurally similar to an approach named within-study comparison (WSC), its purpose is fundamentally different. In most WSCs (e.g., Wong and Steiner, 2018), an RCT estimate is used to validate an estimate from a quasi-experiment. In contrast, the primary aim of the multiple design approach is to produce corroborative evidence in which estimates from different designs are comparable in direction and magnitude. Thus, multiple designs *within the same study* are used to produce results aimed to closely agree. With similar design quality, one result from one design does not hold more inferential weight than a second result from a second design.

More formally, this design-in-design approach reflects the principle of “critical multiplism” which Reichardt (2019) defines as “Using multiple methods with different strengths and weaknesses so that, to the extent the results of multiple methods agree, the credibility of results is increased.” Here, “multiple methods” can be thought of as multiple designs. If estimates from different designs within the same study are comparable, then causal inference is enhanced.

I searched for a case in health or medicine in which a multiple-design approach had been used to yield this kind of corroborative evidence, but my efforts were unsuccessful. A recent high quality meta-analytic review (Chaplin et al., 2018) reported 15 WSCs in which RCTs results were used to validate regression discontinuity (RD) results. In a personal communication with one of the authors, (Morris, September 5, 2019), I wondered if the searching procedures these authors had used to identify WSCs might also have identified examples in which more than one design had been used as corroborating evidence. Unfortunately, the Chaplin et al. search strategy did not yield a relevant RCT-RD example in the health or medicine field that had been used in this corroborating way.

5. Examples that relate Campbell’s approach to Hill’s viewpoints

Two fictional but realistic examples apply the Campbellian approach to amplify the nine knowledge probes from Hill. The first example incorporates the regression discontinuity design to assess a drug intervention for hypertension. Imagine a drug intervention aimed to reduce hypertension. Recent rules establish a cut-point for clinically significant systolic

blood pressure (SBP) of 120 mm. of mercury (e.g., Mohammad and Deshpande, 2016). Certainly, family history and other risk factors enter into a clinical decision to prescribe hypertension medication, but, for the sake of presentation brevity, those factors will be ignored. For the RD design, with initial SBP on the x-axis and subsequent SBP on the y-axis, one looks for a discontinuous SBP drop at the cut point in the regression line for patients above 120 who received the drug.

The second example embeds an RCT within the RD design (e.g., Chaplin, 2018). Though also hypothetical, the example’s logic is borrowed from an application of the design-in-design approach used in the social sciences (Yeaton and Moss, 2020). For medical applications, Kaiser (2015) has extolled the virtues of combining RCT and observational studies but does so by incorporating both kinds of information within the same statistical model. In the current study, the combination design is used to corroborate effect estimates.

In the recent past, the line between normal and elevated systolic blood pressure has vacillated between 120 and 140, and best placement of the distinguishing cut-point still bears research scrutiny (e.g., Flint et al., 2019). Thus, one might imagine an RCT specifically aimed at persons in the range of uncertainty between 120 and 140. We would expect an RCT-based reduction to be comparable to the size of the discontinuous drop for the RD (keeping in mind that the RCT measures average treatment effect and the RD measures average treatment effect at the cut-point).

To better articulate the value-added of combining the Campbell and Hill paradigms, I first note and briefly describe each of Hill’s nine viewpoints. Then, in the context of the two fictional examples introduced above, I elaborate upon additional knowledge gained by adding data from the Campbellian perspective.

1. Strength (size of effect)

For Hill, a strong environment-disease relationship is one that produces large effects; it is easier to discern causal relationships in the context of large rather than small effects. To best assess the comparability of effects from two different designs, we would compare the size of the RD discontinuity to the distance between the RCT’s treatment and control regression lines at the cut-point. If the average treatment effect in the RCT and the average treatment effect at the cut-point in the RD were similar, that corroboration would provide a more robust strength estimate...one at the cut-point and one near the cut-point.

2. Consistency (similar results across different study units)

Hill seeks findings that are similar across “different persons, places, circumstances, times.” The RD assignment variable runs the gamut of baseline SBPs. The RCT is limited to SBPs for persons in the 120-140 range. If the RCT were based on a random sample of participants in the region near the cut, further consistency of study findings would be established.

3. Specificity (asks if the relationship is one-to-one: does a given disease have one kind of outcome and is that disease created by a single cause)

Hill notes that one-to-one, independent-dependent variables relationships are infrequent. To illustrate, he asks whether cancers related to smoking will increase but

cancers and deaths due to other causes (e.g., brain cancer) will not. To probe specificity in the Campbellian approach, both RD and RCT can utilize the so-called “non-equivalent dependent variable” design (e.g., Herman and Walsh, 2011). The idea: the hypertensive drug’s benefit will be reflected by a difference in SBP between the treatment and the control group in both designs, but a second disease outcome, say ulcer prevalence, will be not show such a difference. The multiple design approach allows replication of a no-difference finding in diseases unlikely to change from hypertension medication.

4. Temporality (causes precede effects)

RD and RCT designs are both prospective. The cause necessarily precedes the effect.

5. Biological gradient (dose-response)

In the absence of confounders that covary with dose, if increased doses lead to increased effects, cause is buttressed. In the RD design, one could track different dose levels in different patient subsets to determine if there were different degrees of SPB decrease. In the RCT, one could randomly assign patients different doses.

In an educational version of the multiple design approach, a random sample of college students who had been placed on academic probation with semester grades below C (2.0) but at or above D (1.0) were first chosen (Yeaton and Moss, 2018). Next, students in this 1.0-2.0 region were randomly allocated to receive either a strongly worded, warning letter sent by certified mail (stronger dose) or an email warning letter containing less strong language (weaker dose). In addition, RD results in the higher dose, probation intervention were compared to results for students who scored at or above 2.0. Thus, the dose-response data from the RCT were used in collaboration with the higher dose treatment versus no-treatment results of the RD design.

6. Experiment (random allocation)

A small sample RCT was used in the design-in-design approach. This hybrid design approach with experimental and quasi-experimental evidence has coincident advantages; RD had larger sample size (enhanced statistical power) and enrolled patients with a wider range of pretest scores.

Each of the remaining three Hillian viewpoints addresses the credibility of the independent-dependent variable relationship. In the Campbellian paradigm, these three viewpoints relate to causal attribution but are not direct threats to internal validity. Rather they address construct validity, the representations and labels given to the cause and effect constructs and the causal mechanism itself. Empirical study of these three viewpoints using the Campbellian paradigm would allow researchers to address the nature of the presumed cause. Investigation of these three Hillian viewpoints requires elaborated Campbellian designs (e.g., additional design elements such as comparative groups or supplementary data that establish causal links) to inform each viewpoint.

7. Plausibility (asks if the relationship between the independent and dependent variable “makes sense,” from the standpoint of contemporary knowledge)

8. Coherence (asks if different levels of evidence “tell a similar story” regarding cause and effect. As Hill notes, that evidence may be more expansive; women begin to suffer the same lung cancer fate as men when women begin to smoke more, or the pattern of harm is similar when the biology of smoking is examined in rats and man).
9. Analogy (asks if another drug or disease would have similar biological effects)

To address the plausibility of benefit of a particular hypertensive drug, each design might include measures of the biological markers known to be associated with that hypertensive drug. If the predicted causal links occur within such a mechanism study, then the drug’s link with benefit is enhanced. To address coherence, inference from the RD and RCT designs might be buttressed by using multiple study conditions (e.g., different drug treatment groups in both the RCT and the RD...perhaps those with long and short histories of hypertension). To address analogy, different hypertensive drugs of the same class, say a different beta blocker, might be used, in both designs, to better probe the “logic” of the disease.

6. Cochran’s causal crossword in the context of Hill and Campbell

Rosenbaum (2015) reflected upon Cochran’s idea of a “causal crossword” to better articulate an intricately woven argument showing how cigarette smoking might be causally related to lung cancer. Below, Rosenbaum’s words appear in italics but are interrupted to allow relevant Hillian and Campbellian commentary. Each element of Rosenbaum’s argument reflects a Hillian viewpoint. And each Hillian viewpoint can be directly or indirectly related to design thinking via Campbell.

A controlled randomized experiment shows that deliberate exposure to the toxin causes this same cancer in laboratory animals.

Experiment (with rats, not humans)

Consider the same ideas in a biological context. A high level of exposure to a toxin, such as cigarette smoke, is associated with a particular disease, say a particular cancer, in a human population, where experimentation with toxins is not ethical.

Consistency (with humans) in an observational design (smokers versus non-smokers)

A DNA-adduct is a chemical derived from the toxin that is covalently bound to DNA, perhaps disrupting or distorting DNA transcription. Exposure to the toxin is observed to be associated with DNA-adducts in lymphocytes in humans.

Plausibility (addresses mediator/moderator issues by providing data on biological mechanism)

A further controlled experiment shows that the toxin causes these DNA adducts in cell cultures.

Coherence (via experiment, using evidence at the cellular level)

A case-control study finds cases of this cancer have high levels of these DNA-adducts, whereas non-cases (so-called “controls”) have low levels.

Temporality (observational data with humans, but the data are retrospective)

A pathology study finds these DNA-adducts in tumors of the particular cancer under study, but not in most tumors from other types of cancer.

Specificity (idea is reflected in the non-equivalent dependent variable design)

Certain genes are involved in repairing DNA, for instance, in removing DNA adducts; see Goode et al. (2002). In human populations, a rare genetic variant has a reduced ability to repair DNA, in particular a reduced ability to remove adducts, and people with this variant exhibit a higher frequency of this cancer, even without high levels of exposure to the toxin.

Analogy (idea of an observational study for human cancer rates with and without this variant)

Each of these entries in the larger puzzle is quite tentative as an indicator that the toxin causes cancer in humans, and some of the entries do not directly intersect; e.g., the rare genetic variant is not directly linked to the toxin. Yet, the filled in puzzle with its many intersections may be quite convincing.

This summary statement nicely reflects the Hillian perspective; only the interwoven tapestry argument becomes definitive. Note that Rosenbaum does not include the viewpoints strength or dose-response. However, these two viewpoints could easily be incorporated using relevant design elements via Campbell (by adding one pack-a-day and three pack-a-day smokers to an observational study and by reporting the large expected difference in lung cancer rates for these two smoking conditions).

7. What is missing from Hill

The Hillian paradigm is incomplete. The Campbellian perspective considers dimensions not directly discussed by Hill that are critical to correct causal inference. For example, differential attrition, treatment compliance, and diffusion from the treatment to control group are not explicitly addressed in Hill's viewpoints. The construct validity of cause threats mentioned above (compensatory equalization compensatory rivalry, and resentful demoralization) directly relate to the one of the Stable Unit Treatment Value Assumption requirements of Rubin's (1974) potential outcomes approach—that administering treatment to intervention group members does not change the likelihood that control group participants receive treatment. In addition, subsequent to Hill's 1965 address, design advancements such as instrumental variable analysis (Angrist, Imbens, and Rubin, 1996) have been applied to provide high quality outcome estimates in the face of incomplete treatment compliance.

8. Value-added considerations via Campbell

Five of Hill's viewpoints are directly informed by Campbellian, design-based approaches, while three Hill viewpoints can be informed using empirical evidence of a surrogate nature. (Recall, one Hillian viewpoint, temporality, is redundant with a Campbellian internal validity threat.) Multiple research designs used in the same context can better inform the Hillian paradigm, since one design informs fewer viewpoints. For example, an estimate of effect from a single RD design would not inform consistency. However, one might use a single, prospective design (e.g., RD only) to address temporality, specificity, and dose-response, then add a second design, say an RCT that incorporates a design-in-design approach, to enhance both strength and consistency (to judge both size and similarity of effect sizes). As an extra bonus, with evidence from multiple designs within the same study, consistent with Hill's values, the need to rely upon statistical significance is further reduced.

By itself, Hill’s paradigm provides a rich, principle-driven perspective that contributes important theoretical perspectives. By itself, Campbell’s approach provides design-contingent, empirical evidence allowing one to investigate the pattern of multiple study findings and to elaborate Hill’s nine queries. Together, the empirically- and conceptually-based probes of causal truth contribute complementary rule-in and rule-out arguments on which to reduce uncertainty and to enhance inference. Though sometimes non-intersecting, the resulting Hill-Campbell methodological combinations allow researchers to form a more complete causal mosaic.

References

- Angrist, J., Imbens, G. and Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91, 444–472.
- Campbell, D.T. (1957). Factors relevant to validity of experiments in field settings. *Psychological Bulletin*, 54, 297–312.
- Campbell, D.T. (2005). Ethnocentrism of disciplines and the fish-scale model of omniscience. In S.J. Derry, C.D. Schunn, and M.A. Gernsbacher, (Eds.), *Interdisciplinary Collaboration: An Emerging Cognitive Science*, pp. 3–21. Lawrence Erlbaum, Mahwah, NJ.
- Campbell, D.T. and Stanley, J.C (1966). *Experimental and quasi-experimental designs for research*. Rand McNally, Chicago.
- Chaplin, D.D, Cook, T.D., Zurovac, J., Coopersmith, J.S., Finucane, M.M., Vollmer, L.N., and Morris, R.E., (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, 37, 403–429, <https://doi.org/10.1002/pam.22051>
- Cook, T.D. and Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Rand McNally, Chicago.
- Flint, A.C. et al. (2019). Effect of systolic and diastolic blood pressure on cardiovascular outcomes. *New England Journal of Medicine*, 381, 243–251.
- Heit, H. (2016). Reasons for relativism: Feyerabend on the ‘rise of rationalism’ in ancient Greece. *Studies in History and Philosophy of Science*, 57, 70–78.
- Herman, P.M. and Walsh, M.E. (2011). Hospital admissions for acute myocardial infarction, angina, stroke, and asthma after implementation of Arizona’s comprehensive statewide smoking ban. *American Journal of Public Health*, 101, 491–496.
- Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58, 295–300.
- Kaisar, E.E. (2015). Incorporating both randomized and observational data into a single analysis. *Annual Review of Statistics and its Application*, 2, 49–72.
- Pearl, J. (2018). *The book of why: The new science of cause and effect*. Basic Books, New York.
- Reichardt, C.S. (2019). *Quasi-experimentation. A guide to design and analysis*. Guilford, New York.
- Rosenbaum, P.R. (2015). Cochran’s causal crossword. *Observational Studies*, 1, 205–211.
- Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

- Saklayen, M.G. and Deshpande, N.V. (2016). Timeline of history of hypertension treatment. *Frontiers in Cardiovascular Medicine*, 3, 1-14.
- Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Houghton Mifflin, Boston.
- Wong, V.C. and Steiner, P.M. (2018) Designs of empirical evaluations of nonexperimental methods in field settings. *Evaluation Review*, 42, 176-213.
- Yeaton, W.H. and Moss, B.G. (2020). A multiple-design, experimental strategy: Academic probation warning letter's impact on student achievement. *Journal of Experimental Education*, 88, 123-144.
- Yeaton, W. H. (2019). Experimental and quasi-experimental designs. In P. Brough (Ed.), *Advanced Research Methods for Applied Psychology*, pp. 107-123. Routledge, New York.