Journal of Methods and Measurement in the Social Sciences,
Vol. 7, No. 2., 122-143, 2016

Transforming the canons of John Stuart Mill from philosophy to replicative, empirical research: The Common Cause research design

William H. Yeaton
Florida State University

Christopher G. Thompson
Texas A&M University

When an element or factor is common to a set of circumstances that element may be causal in its relationship to particular dependent variables. This premise was stated by John Stuart Mill more than 170 years ago, and Mill's canon, the Method of Agreement, is used here as a basis to create the "Common Cause" (CC) research design. The CC design is particularly relevant when a set of multiple circumstances can be represented by alternative theories of change or competing explanations. We consider several potential applications of the design and elaborate its structure within the validity framework of Shadish, Cook, and Campbell. We discuss threats to validity controlled by the CC design (e.g., selection bias, the bane of applied researchers, is not relevant) and illustrate possible analytic strategies using simulated data. We explicitly compare the CC design to four quasi-experimental designs in terms of the validity threats that they eliminate. Design weaknesses are addressed and ways to enhance the design's inferential power discussed. The CC design itself represents a proof of concept suggesting that other research designs can be created from philosophical principles.

Keywords: quasi-experiment, research methods, research design, evaluation, John Stuart Mill

On the eve of the twentieth century, Mary Mallon emigrated from Ireland at age fifteen to make her way in New York City. Brave, headstrong, and dreaming of being a cook, she fought to climb up from the lowest rung of the domestic-service ladder... Then, one determined “medical engineer” noticed that she left a trail of disease wherever she cooked, and identified her as an “asymptomatic carrier” of Typhoid Fever. With this seemingly preposterous theory, he made Mallon a hunted woman.

> Internet description of “Fever, a novel” by Mary Beth Keane (2014), describing “Typhoid Mary, the first person in America identified as a healthy carrier of typhoid fever.

[1]Studies that yield causal inference have proven valuable as they often contribute to theory and provide unambiguous evidence of programmatic benefit. The primary aim of this paper is to expand the design options available to multi-disciplinary researchers who regularly confront thorny, applied problems. The new design (termed the “common cause” or CC design) offered here finds its foundational basis in the works of philosophy instead of being created out of “whole cloth.”

Following a long-standing tradition, contemporary methodological researchers have focused on improving existing designs rather than creating new designs. For example, in

[1] The Common Cause (CC) design is the creation of Jared Boyd. It was submitted as a three-page outline to satisfy a class assignment for a research methods class taught in 2005 by the first author. During the last week of class, Mr. Boyd was encouraged to submit a paper for publication based on his work. Several years later, Mr. Boyd and the first author met a second time, and he was again encouraged to publish a paper. On both occasions, Mr. Boyd agreed that this was a worthwhile task. Unfortunately, efforts to communicate with Mr. Boyd in the interim have proved unsuccessful. This paper was written without Mr. Boyd’s collaboration, but both authors wish to be clear that the initial idea for the design is his contribution. In addition, a portion of the current paper follows his class outline.

the last decade or so, estimates from well-established quasi-experimental designs have been empirically validated using so-called "within-studies comparison" (WSC) tactics (e.g., Cook, Shadish, & Wong, 2008). These assessments attempt to establish comparability between the results of randomized studies and those in the three strongest quasi-experiments, regression discontinuity (e.g., Shadish, Galindo, Wong, Steiner, & Cook, 2011), controlled interrupted time series (ITS; e.g., St. Clair 2014), and non-equivalent control group designs (e.g., Shadish, Clark, & Steiner, 2008). (In this paper, the terminology "control group" and "comparison group" are used interchangeably.) To date, inferential improvement has focused upon only these three strong quasi-experiments, leaving intact the inferential quality of less sound quasi-experiments (QEs). The CC design is intended as a viable alternative to these weaker designs; we provide a detailed argument for its inferential advantage compared to four quasi-experiments: case study, single pretest-posttest design, posttest only control group design, and regression point displacement design.

Philosophers have written at considerable length about the conditions under which causal inference can be justified. For example, Aristotle, Plato, Descartes, and Hume each considered causal inference to be an integral element of their particular philosophies (Copri & Cohen, 1990). John Stuart Mill's works are particularly rich and detailed in their focus on cause. In Mill's "A system of logic" (1843), several canons were introduced and procedures for establishing causal inference elaborated. While the Method of Agreement is the focus here, we will also discuss the important role of Mill's Method of Difference.

Mill defines the Method of Agreement as: "If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon" (Mill 1843, p. 454). Symbolically, Mill's Method of Agreement can be stated as

$$\begin{aligned} &\mathrm{A} \rightarrow a \\ &\mathrm{A, B, C} \rightarrow a, b, c \\ &\mathrm{A, D, E} \rightarrow a, d, e \end{aligned}$$

where, following Mill's terminology, A through E can be considered as "circumstances" and a through e as "instances." In more modern terminology (and that used in this paper), "circumstances" are similar to factors and "instances" closely resemble dependent variables.

In less formal terms, if a consistent relationship repeats itself across situations, that replication is suggestive of cause. In the "Typhoid Mary" case cited above, Mary was the common element (A) among a set of circumstances (B, C, D, and E) that represent "other factors" possibly implicated in contracting typhoid (represented by a). Each row of circumstances in the notation example above may reflect a list of people who contracted typhoid. Mary's common presence (and the absence of other common factors associated with a) led public health investigators to conclude that she was the likely cause of the disease.

Related to this canon is Mill's Method of Difference, which Mill articulates as "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former, the circumstance in which the two instances differ, is the cause, or an indistinguishable part of the cause, of the phenomenon" (Mill 1843, p. 455). Symbolically, Mill's Method of Difference can be stated as

$$\begin{array}{l} \text{A} \rightarrow a \\ \text{A, B, C} \rightarrow a, b, c \\ \text{B, C} \rightarrow b, c \end{array}$$

In the context introduced above, one might say: If typhoid (*a*) is consistently present when Mary (A) is present and consistently not present with the absence of Mary, then Mary likely caused many cases of typhoid.

Each method uses a specific kind of process of elimination to arrive at causal inference. With the Method of Agreement, the task is to identify a unique element (in the above case, A) such that "... the different elements have no circumstance in common except A" (Mill 1843, p. 450). As Mill notes, "...*b* and *c* are not effects of A for they were not produced by it in the second line; nor are *d* or *e*, for they were not produced in the first [line]" (p. 451). Because A appears in both lines, its effect must be produced in both lines, and only effect *a* fits this pattern.

The Method of Difference is more familiar to researchers as it is the logical motivation for establishing many possible control groups (typically, a group that has all the features of the treatment group, save one). "Instead of comparing different instances of a phenomenon, to discover in what ways they agree, this method compares an instance of its occurrence with its non-occurrence, to discover in what they differ" (Mill 1843, p. 455). While the Method of Difference is fundamental to causal inference, the Method of Agreement serves as the foundation for creation of the Common Cause (CC) design.

A Brief, Recent History of Research Designs

Research methodologists share the same strong interest in cause as that demonstrated by philosophers. In fact, research designs, as reflected in their various configurations, aim to establish varying degrees of causal inference. The primary social science methods textbook (Shadish, Cook, & Campbell, 2002) notes that when designs reflect temporal precedence of the presumed cause, document covariation of cause and effect, and make alternative explanations implausible, causal inference is enhanced. Utilizing these three features, randomized studies (experiments) allow the strongest causal conclusions (e.g., Fisher, 1935), but more recently developed quasi-experimental designs (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002) can also approximate strong causal inference.

While enhanced techniques to statistically analyze observational data have emerged in the last few decades (e.g., Morgan & Winship, 2007; Rubin 2005), new research designs have been much less frequent. Well after the dissemination of methods for randomized studies by Sir Ronald Fisher, Donald Campbell and his colleagues published a long compendium of quasi-experimental designs (Campbell, 1957; Campbell & Stanley, 1966). With the possible exceptions of the regression-discontinuity design (Cook 2008), given its recent reemergence after an initial appearance in the 1960's (Thistlewaite & Campbell, 1960), the regression point displacement design (Trochim &

Campbell, 2014), and the Sequential Multiple Assignment Randomized Trial (SMART) design (e.g., Lei, Nahum-Shani, Lynch, Oslin, & Murphy, 2012), there has been a dearth of new research designs in the last half century.

Conceptual Examples Using the Common Cause Design

The logic of the CC design is not unfamiliar to those who make or assess claims in everyday life. Imagine that a prosecutor wishes to establish guilt for a series of robberies. A police force begins to investigate and soon uncovers what is believed to be a "signature" hypothesized to represent the alleged criminal's presence at each crime scene. If the crook's *modus operandus* has been consistent, the finding of guilt is enhanced. As in the typhoid Mary example, consider the field of epidemiology where researchers often ask "What element is common to each person who has the disease (e.g., common exposure to a food pathogen or environmental risk factor)?" To "discover" successful paths to academic jobs for graduate students, faculty may look for a consistent pattern of scholarship in recent graduates (e.g., all took classes in research methodology, all published a substantial number of peer-reviewed articles, and each had experience teaching).

Further imagine that a governor wishes to reduce the incidence of drunk driving in her state. You believe that a substantially larger fine for drinking and driving would reduce the rate of intoxicated driving. Suppose that there are three prominent theories with accompanying research evidence for reducing drunk driving. One approach increases the saturation of police at night and on weekends. A second strategy disseminates Public Service Announcements (PSAs) advocating use of a designated driver. The last approach limits the number of alcoholic beverages that bars may serve an individual, after midnight.

To substantiate the claim that more severe punishment (a bigger fine) is effective, you choose three cities in your state whose initial levels of drunk driving are similar. In the first city, police patrolling levels are relatively stable, and you apply the more severe punishment. In the second city, there are no PSAs regarding designated drivers (the level is constant, at zero), and you administer harsher punishments. Finally, the bigger fine occurs in a city where the number of drinks served in bars after midnight, as reflected by legal statute, remains the same. In each instance where the policy reflected by the competing theory is constant, the intervention is applied. And in each case, you find a statistically significant effect.

This replication of impact provides convincing evidence that the intervention may be causal. The reduction in drunk driving was not due to the level of police patrolling, the presence of PSAs, or the policy of serving drinks after midnight, since in each instance these policies were held constant, yet favorable effects consistently occurred. It was the element in common to all three cites, the larger fine, which produced benefit.

One way to marry the verbal approach taken by philosophers and that taken by developers of designs is to translate concepts reflected in philosophical writing into the structure of empirical research reflected by the unique configuration of different research designs. This is precisely the strategy used to develop the philosophy-based CC design. In homage to the Method of Agreement canon espoused by Mill, this new design is termed the "Common Cause" design.

Figure 1: Mill’s method of agreement within the context of the Common Cause design.

Figure 1a: The letter A represents Theory A, for some dependent variable *a*. Rival theories are represented as B, C, D, E, each with respective dependent variables *b*, *c*, *d*, *e*. Among all theories, Theory A is unique in that its effect *a* is consistently present.

Mill’s method of agreement

$$A \rightarrow a$$
$$A, B \rightarrow a, b$$
$$A, C \rightarrow a, c$$
$$A, D \rightarrow a, d$$
$$A, E \rightarrow a, e$$

Figure 1b: Observations are Os. Intervention of interest is X_A, for Theory A.

A representation of Mill’s method of agreement using Xs and Os

$$OB_1 \quad OB_2 \quad X_A \quad OB_3 \quad OB_4$$
$$OC_1 \quad OC_2 \quad X_A \quad OC_3 \quad OC_4$$
$$OD_1 \quad OD_2 \quad X_A \quad OD_3 \quad OD_4$$
$$OE_1 \quad OE_2 \quad X_A \quad OE_3 \quad OE_4$$

Figure 1c: Theory A is the treatment of interest (i.e. A = X). Interventions B-E are constant or absent (constant at level zero) during times t_1-t_4. All dependent variables *a-e* are the same, and levels of interventions for Theories B-E are assumed to be consistent. The Common Cause X for Theory A is absent in baseline (denoted as ~X) and present during treatment for Theories B-E.

An elaboration of the CC design

		Baseline		Intervention	
	Observation	t_1	t_2	t_3	t_4
Theory Dependent Variable	A	~X	~X	X	X
	DV	*a*1	*a*2	*a*3	*a*4
	B	B	B	B	B
	DV	*b*1	*b*2	*b*3	*b*4
	C	C	C	C	C
	DV	*c1*	*c2*	*c3*	*c4*
	D	D	D	D	D
	DV	*d*1	*d*2	*d*3	*d*4
	E	E	E	E	E
	DV	*e*1	*e*2	*e*3	*e*4

Figure 1 is displayed in three parts (Figures 1a, 1b, and 1c). In Figure 1a, we present a symbolic representation of Mill’s Method of Agreement. In Figure 1b, the schematic shows that an intervention (X) based on Theory A (X_A) is implemented within each of the four units of the design, in the context of several Os (observations). Each of the four units (there are four competing theories, in this case) receive the intervention of interest when one wishes to make the claim that the particular X based on Theory A is likely to cause change. One or more pretest and posttest measures are taken in each design line to reflect a unique theory or explanation of change. The idea: If introduction of the X based on Theory A is associated with consistent change in each line, a causal relationship is likely. Since individual theories in other design units are held constant, the consistent co-incidence of change is attributed to the X from Theory A.

Lastly, in Figure 1c we elaborate upon important dimensions of the CC design. Two baseline measures (t_1 and t_2) are made during "~ X" (which can be read as "not X"), prior to initiation of an intervention (noted as "X"), based on Theory A. After X is initiated, two follow-up measures are taken (t_3 and t_4) for the same dependent variable used at baseline. After baseline, in remaining units of the CC design (B-E), that same X is implemented. Because only the X for Theory A is common to each unit, and because the level of the four rival theories has been held constant during observations t_1-t_4, change is attributable to the X.

Distinguishing Common Cause from Other Quasi-experimental Designs

Unlike contrasts in most experimental or quasi-experimental designs, the CC design lacks a no-treatment control. All units receive treatment. This fundamental difference is not only critical to a sound understanding of the CC design's logic of inference but also represents a stark advantage. Since there is no control group in the CC design, researchers avoid the often arduous task of demonstrating that selection bias (the presence of initial, between-group differences) is absent. Statistical adjustment procedures such as propensity scores become irrelevant.

To further clarify the logical basis of the CC design, consider the nonequivalent control group (NECG) design which compares results in the treatment group (first row) to those in a control group (second row) that typically receives no treatment:

```
NR  O  X  O
NR  O     O
```

Whether as a difference-in-differences or as a between-group difference at posttest, the inferential logic of the NECG design utilizes the Method of Difference. To the extent that the two groups are otherwise initially similar, the unique treatment element in the first row explains resulting outcome differences. However, while the logical underpinnings of the two designs appear fundamentally distinct, those between-groups elements equalized in the NECG design might be the same elements made consistent within lines of the CC design, and these elements might be held constant one- or many-at-a-time. The Method of Difference also applies to the ITS with comparison group design (below), because the primary difference between groups is presumed to be the existence of treatment in the first row:

```
NR  OOOOOO  X  OOOOOO
NR  OOOOOO     OOOOOO
```

However, the CC design, which uses the Method of Agreement, is based on an inferential foundation that is fundamentally different than both of these quasi-experiments! In the CC design, the counterfactual can be regarded as a within-unit pattern of no pre-post change, across study units. A pattern of consistent change in the presence of Theory A (multiple instances of line 1) versus no consistent change in these multiple instances, will imply cause.

The Strength of Multiple Comparisons in the Common Cause Design

As a potential remedy for mis-estimation of effects in observational studies, Rosenberg (1987) argued that a second control group offers several advantages.

> The value of a second control group depends on the supplementary information that is available about unobserved biases that are suspected to exist. A second control group provides a test of the assumption that conventional adjustments for observed covariates suffice in estimating treatment effects… two control groups can yield consistent and unbiased estimates of bounds on the treatment effect when conventional adjustments fail. (p. 292)

Rosenbaum notes that a second control is particularly apt when statistical adjustment may be inadequate (with known covariates) or when "unobserved bias" is suspected (the problem of important, omitted variables). A second control group allows one to bracket the magnitude of change by determining the degree of overlap of the confidence interval created from the second control group with that created by the first control group.

Following this line of thought, the CC design requires multiple comparisons, but each pre-post comparison is an estimate of the intervention's effect. Within each estimate, a particular covariate may occur to some varying degree (or may exist and not be measured). However, the more units present in the CC design, the greater the prospect that one or more design units include this omitted variable.

To illustrate, if the degree of initial student technology experience is critical to the demonstration of a favorable impact of a computer-based intervention, students in some schools will almost certainly have considerably differing degrees of exposure than others. This omitted variable does not confound between-group differences that hound observational studies based on the Method of Difference; these comparisons are not made in the CC design. Instead, the extent to which this omitted variable is present or absent across study units actually enhances external validity and allows a more reliable bracketing of intervention effect. Thus, while a NECG researcher would try to show no-difference on a multitude of covariates to avoid confounding, a CC design researcher might be motivated to introduce a multitude of different kinds of units on which to replicate the intervention. However, the reason underlying their motives will be quite different (avoid confounding in the first case and enhance external validity or establish a more reliable estimate in the second).

Implementing the Common Cause Design

To facilitate implementation of the CC design, we provide a set of guiding steps. While this proxy for a "user's manual" occurs in the context of a single example, interventions in other disciplines can easily be substituted (e.g., exercise for heart health, shared activities for quality of relationship status, medication for illness, removal of disincentives to enhance voter turnout).

Continuing the example noted above, assume you wish to assess the benefits of new computers on learning in poorly performing schools. Further assume that multiple years of previous test scores are available in each school. These achievement scores exist in both mathematics and English content areas. The times during which computers are introduced are different for each school. Is it possible to determine that ensuing increases in learning are attributable to the new computers?

In the logical framework of the CC design, we seek a consistent pattern of increase in both mathematics and English, across schools. When individual schools with new computers have bigger gains than those without new computers, and when these consistent benefits occur soon after computer introduction, this pattern of agreement will lead us to conclude that the relationship between computer introduction and greater learning is likely to be causal.

A critical, first step in the implementation process for the CC design is to establish a list of all theories that would lead to changes in the dependent variable. The adequacy of the list depends on the thoroughness of the set of possible explanations that have been identified. This list is analogous to the correct identification of the set of all important baseline covariates in between-groups designs, where selection bias will occur if the list in not exhaustive. For the above, computer-based intervention, researchers must protect against the presence of coincident interventions based on other theories.

The idea is straightforward: demonstrate that the presence of X, the claimed theory of change, leads to statistically significant differences for the multiple contexts in which each rival theory has not varied. In many research domains such as psychology, sociology, public health, and communications, well-known, rival theories have a long-standing history of being pitted against one other, so the process of identification should be relatively straightforward. Incomplete vigilance can undermine correct inference with all research designs.

It is worth re-emphasizing that the presumed causal mechanism of *each* rival theory must be held constant at some particular level (that level can be zero) during the duration of the study. Otherwise, change in the rival theory may compete with the X as an explanation for the change in the dependent variable. As an example, suppose one wishes to demonstrate that a legal change in the speed limit is instrumental in the decrease in traffic fatalities when the new law is passed. In three different states, poor weather conditions might be consistent in a one state, the level of automobile safety features (e.g., better tires) constant in the second, and safe driver PSAs absent in the third. Should one or more of the three counterfactual states exhibit a change in the level of the independent variable of its relevant theory during the study's duration, a different state should be found to test the intervention of interest.

The CC design requires that at least one pretest and one posttest measure be collected within each unit. The X, the intervention based upon the theory of interest, is implemented in each unit. That implementation can occur at a single point in time or can be staggered over time, within the differing study units. As noted in the later section on threats to validity, however, there may be substantial advantages in eliminating validity threats within each unit when multiple pretest and posttest measures have been recorded.

At the data analysis stage, the CC design is treated as if there were multiple independent studies, one study per each sample. One wishes to establish that, in the presence of the X, there is a statistically significant change in one or more dependent variables of interest. Thus, the critical demonstration is within-units; CC design researchers seek a consistent pattern of pre-post changes. It is best when baseline levels of the dependent variable are reasonably comparable across units, and the magnitude of the change relatively large (and statistically significant) and comparable, across units, to clearly indicate that X is causal. However, as with inference in all research designs, heterogeneity of results (e.g., across subsets, across follow-up periods, and across measures) may occur, and judgment regarding beneficial patterns will be required.

Combining the CC Design with Other Designs

The CC design need not be used in isolation and could be combined with between-group designs to enhance causal inference. In fact, a recent methodological trend encourages multiple designs within the same study (e.g., Shadish, Clark, & Steiner, 2008). Strong quasi-experiments may also be used to rule out additional threats found in weaker quasi-experiments (Kowalski, Yeaton, Kuhr, & Pfaff, in press). Prospective and retrospective data can be combined, and subsets of within- and between-groups data analyzed.

To illustrate, a New York Times Op-Ed contribution (Krugman, 2015) cites a two decades old, between-groups study (Card & Krueger, 1994) used to test orthodox economic theory positing that an increase in the minimum wage would negatively impact employment. Card and Krueger used a single pretest and posttest from each of 410 fast-food restaurants in New Jersey (where the minimum wage was increased), in eastern Pennsylvania (no increase in the minimum wage), and in restaurants in New Jersey (where the raised minimum was previously in effect). Krugman noted that, rather than a negative impact on employment, "…they found, if anything, a positive effect." Card and Kruger relied upon average results in each state and a number of different between-groups regression models to reach their conclusions.

Now, imagine what a CC design would have contributed. Each fast-food restaurant in New Jersey could be examined for employment change after the minimum wage law's implementation, as could each nearby restaurant in eastern Pennsylvania. Rival economic explanations for unemployment such as recession were constant across time, in each state. For more than 400 restaurants, a pattern of no decrease in employment in most New Jersey fast-food restaurants but few changes in Pennsylvania would represent more fine-grained and compelling evidence than conclusions based on averages.

Threats to Validity: Internal and Statistical Conclusion

The inferential strength of the CC design fundamentally depends upon the consistency of outcomes across the multiple units of the design. The logic of the CC design does not require the highest level of causal inference *within* each study unit. Taken individually, pretest-posttest designs are inferentially weak. But it is agreement in the pattern of consistent change *across* multiple units that is critical to determining if an internal validity threat exists.

Within the CC context, history is a threat to internal validity if, despite replication of the intervention and consistent effects across study units, treatment is given at the same time point within each unit. When some external event happens to coincide with each implementation of the intervention for each unit of the CC design, the cogency of the Method of Agreement will be diminished. Fortunately, treatment can be staggered in time so that history can be rendered less plausible (e.g., laws may be initiated at different times in different states).

Testing is a potential threat in the CC design. Fortunately, outside of educational contexts, learning that occurs from multiple testing is less likely to occur. When "tests" are simply observations rather than the paper and pencil variety we quickly imagine, testing does not threaten inference. Prior to intervention, if there is a consistent upward or a consistent downward trend in the level of the baseline variable in one or more study units, use of a single pretest and posttest measure of change may wrongly attribute benefit to the X. But, as it is unlikely that maturation's effects will be similar across units, the presence of consistent outcomes in the CC design makes maturation implausible.

The CC design does not control for instrumentation. If the methods by which the outcomes are measured are shown to consistently vary from pretest to posttest in each study unit, causal inference will be suspect. Protection against this validity threat relies on careful vigilance of the research team to maintain measurement quality or to conduct supplementary data analyses to render this threat less likely (see the example for TV reports of suicide, reported below). The ability to detect a possible regression artifact will not exist when there is a single pretest and posttest measure. However, if the outcome effects of regression are dissimilar in different study units and results remain consistent, regression is implausible.

Differential attrition manifests itself differently in the OXO design, where it can occur within each unit of the CC design (in contrast to the between-groups version). We ask if the characteristics of dropouts are such that pretest covariate averages are different than for covariate averages at posttest, across study units (e.g., 5% fewer males at posttest in unit one, 13% fewer males in unit two, 27% fewer males in unit three). If we find consistency in outcomes across these units, then differential attrition is not a viable threat to causal inference.

As noted above, selection bias (confounding of rival explanatory variables, *between groups*) is not a threat within each unit of the CC design, since such differences (confounds) are not relevant to the logic of the CC design. Interpretation will be considerably more straightforward, however, if baseline levels of the outcome variable are comparable between units. If there is heterogeneity of initial responding between units and yet a consistent impact of the intervention, external validity would be enhanced (the treatment effect would be robust to different, initial levels of responses).

Threats to statistical conclusion validity are also inherent to each individual unit of the CC design. Thus, familiar threats such as adequate sample size, reliability of measures, and treatment integrity naturally depend on the quality of these dimensions in each unit of the design.

Threats to Validity: External and Construct

The CC design possesses an inherent advantage regarding external validity, given that the treatment is replicated across units (the more replications, the better). Should different units also utilize different measures, each tapping similar dependent variable constructs (e.g., injuries or fatalities in a traffic safety intervention, morbidity or mortality following administration of a particular drug) or different outcome variable operationalizations, generalizability will increase.

The CC design naturally protects against several threats to construct validity. Mono-operation bias will be unlikely when the independent variable construct is operationalized differently in each design unit. Perhaps physical exercise, the X, is administered by a trainer in some instances, at the gym with friends during other times, and at home in the third unit. Consistent, positive change in health status makes it likely that the general construct "exercise" is instrumental in improving health. In the CC design, because everyone receives essentially the same X, there will be no resentful demoralization, no compensatory rivalry, and no compensatory equalization (since treatment is not withheld from some units). The design avoids "confounding constructs and level of constructs," if different doses of treatment within each design unit lead to consistent effects. Treatment diffusion will not likely be an issue, since all units receive treatment (unless multiple "doses" are differentially favored by study participants).

Changing OXO to ITS in CC design units: eliminating additional internal validity threats

We might imagine that the most fundamental CC design consists of OXSs in each study unit. However, the researcher may alter a subset of the OXOs to create an ITS (OOOXOOO) design by adding additional pretest and posttests. This alteration will lead to enhanced causal inference within study units, thereby bolstering overall causal inference in the CC design.

For example, with many pretests, incremental learning from the next-to-last to last pretest is minimal, so testing becomes a less reasonable threat. The lengthy baseline period germane to the interrupted times series design makes it possible to assess trend and to gauge maturation's potential impact on the outcome. If there are multiple pretests, the possibility of regression should be apparent to the researcher (determine if the X was given at an atypically high or low baseline level). The presence of multiple pretest and posttest observations has no direct impact on history, instrumentation, selection, or differential attrition (though shorter follow-ups may decrease the risk of dropouts).

Eliminating internal validity threats: Comparing the CC design to other quasi-experiments

In this section, we focus upon the advantages of the CC design compared to a subset of four quasi-experimental designs: case study, the single pretest posttest, regression point displacement, and the posttest only control group. Our design comparisons are limited to internal validity threats. Before elaborating internal validity threats in the four other designs, we briefly review threats germane to the CC design.

In the CC design, temporal staggering of the multiple intervention instances controls history, while testing will not be a threat for some kinds of outcomes (e.g., number of bankruptcies before and after a law change does not lead to increased "learning" at posttest). Instrumentation poses a possible threat but supplementary data analysis may enable elimination. Maturation will not be a problem unless the intervention is systematically applied across replications when outcome data in each unit are consistently trending in the same direction. Regression seems implausible when the magnitude of regression is different in different units (consistent size of outcomes across units in the face of different amounts of regression reduces the threat of regression). If covariates are substantially changed due to dropouts from pre to post but outcomes consistently favor the intervention, differential attrition is implausible. It is worth repeating... one of the major strengths of the CC design is that neither selection bias nor the interaction threats with selection threaten causal inference--the logic of inference does not rely on the method of difference.

We now consider the internal threats likely present in four other QEs. Campbell and Stanley (1966) characterize the "one-shot case study" as a "pre-experimental design" and note in Table 1 (p.8) that it does not control most of the listed threats to internal validity. The single pretest-posttest design is said to only control selection and differential attrition (here, each threat refers to possible differences in the characteristics of participants before and after intervention--but these two validity threats are most relevant to designs comparing treatment and no-treatment groups). Thus, when used without the multiple replication feature of the CC design, a single OXO is inferentially quite weak, and neither of these designs fare well in comparison to CC.

After languishing for decades, the regression point displacement (RPD) design has begun a resurrection (e.g., Yeaton & Moss, under review). Typically, a single group receives an intervention and its pretest-posttest result is compared to the regression line of pretest-posttest results for a set of control units. If displacement from the regression line is statistically significant and if other internal validity threats have been controlled, causal inference is enhanced. History and testing are controlled (the impact of both threats will be equalized in the intervention unit and the control units), but instrumentation will be a threat if the measurement method in the treatment group is altered during posttest assessment and no such alteration occurs for posttest measures in control units. Regression will be a threat when there is measurement error or if the intervention group has been chosen based on an atypical extremeness. Random assignment of the single treatment group is recommended, not to achieve pretest comparability of intervention and control groups but rather to avoid choosing the intervention unit to favorably slant results towards a desired benefit. Thus, selection bias is not central to the RPD design's logic of causal inference. However, all of the "selection by interaction" threats are possible.

The posttest only control group design eliminates many internal validity threats (e.g., history, testing, regression, and instrumentation). This design configuration hints that the tactic of adding a posttest group to one or more of the multiple groups of the CC design could be used to eliminate particular validity threats and to strengthen inference. However, the control group tactic does not necessarily eliminate maturation and, more importantly, selection bias, differential attrition, or any of the interaction threats with selection.

```
NR  X  O
NR     O
```

The CC design compares favorably with these last two QEs but has a decided advantage in that selection and its interactions are not potential threats for CC. Like the regression point displacement and posttest only control group designs, history and testing are typically implausible. Unlike the posttest only control group design, maturation is usually not a threat for the CC design and, unlike the two-group, regression point displacement design, regression is unlikely unless the intervention in each OXO consistently follows such atypical periods and leads to inconsistent outcomes across study units. Unfortunately, none of the other designs meaningfully addresses differential attrition (short duration pretests and posttests and outcome consistency can further minimize its likelihood in CC).

Analytic Strategies

Causal claims for the CC design depend on the consistent demonstration of impact, for each line of the design. Thus, inference requires that a statistically significant pre-post difference is shown for most of the individual OXO designs (or ITS designs: OOOXOOO). Fortunately, there are several analytic tactics available to the researcher depending on the available number of pre- and post-measures. With single pretests and posttests, a mixed-design ANOVA may be utilized. If there are several pretest and posttest measures, a repeated measures ANOVA or growth curve analysis would be appropriate. It may also be possible to correctly model a time series analysis when there is a large number of pre and posttest measures (weekly or monthly). Recently developed single-case design analytic procedures (Shadish, Hedges, & Pustejovsky, 2014; Shadish, Zuur, & Sullivan, 2014) are also applicable.

Enhancing the Applicability of the Common Cause Design

This section of the paper is intended to provide a user's guide for analysis and interpretation of CC design results. First, we present several, plausible scenarios users are likely to encounter. We then show patterns of findings indicating when causal inference is warranted, when it is clearly not warranted, and when causal inference is less certain. Second, we provide appropriate tests of statistical significance for these scenarios. In each case, we create simulated data, allowing us to elaborate realistic scenarios for its use.

We explore the applicability of the CC design using four hypothetical research scenarios. Each scenario considers four competing "theories" reflected by three, separate groups. Our choice for the number of groups and theories is arbitrary, though a minimum of three theories is required--a theory of interest and two additional theories (groups) are needed to ascertain a pattern of agreement. Observations are made at four time points for each scenario, and a single intervention occurs between the second and third observations. Using Campbell and Stanley (1966) notation (where "NR" means non-random allocation):

$$\begin{array}{cccccc} NR & OA_1 & OA_2 & X_D & OA_3 & OA_4 \\ NR & OB_1 & OB_2 & X_D & OB_3 & OB_4 \\ NR & OC_1 & OC_2 & X_D & OC_3 & OC_4 \end{array}$$

We consider the case in which each group A-C receives an identical intervention (X_D), constructed from Theory D.

Scenarios are characterized by a feature we term “consistency” (the Mill-ian feature of “agreement”): the uniformity of the presence of an effect and the uniformity of the magnitude of that effect, among study groups. The first, less stringent standard requires consistent presence or consistent absence of an effect across all theories. We call this “effect consistency.” The more stringent standard of consistency requires that the magnitude of effect be uniform, across theories (“magnitude consistency”). Effect consistency does not imply magnitude consistency. For those scenarios in which an effect was always absent, magnitude consistency is vacuously satisfied. This kind of consistency includes the case in which all theories show no effect (i.e., each effect magnitude is essentially equal to zero).

In Table 1, we note which combinations of effect and magnitude consistency were captured within each scenario and provide relevant statistical analyses. In addition, we describe how each scenario was created and how data were generated.

Table 1
Consistency Combinations for Four Scenarios

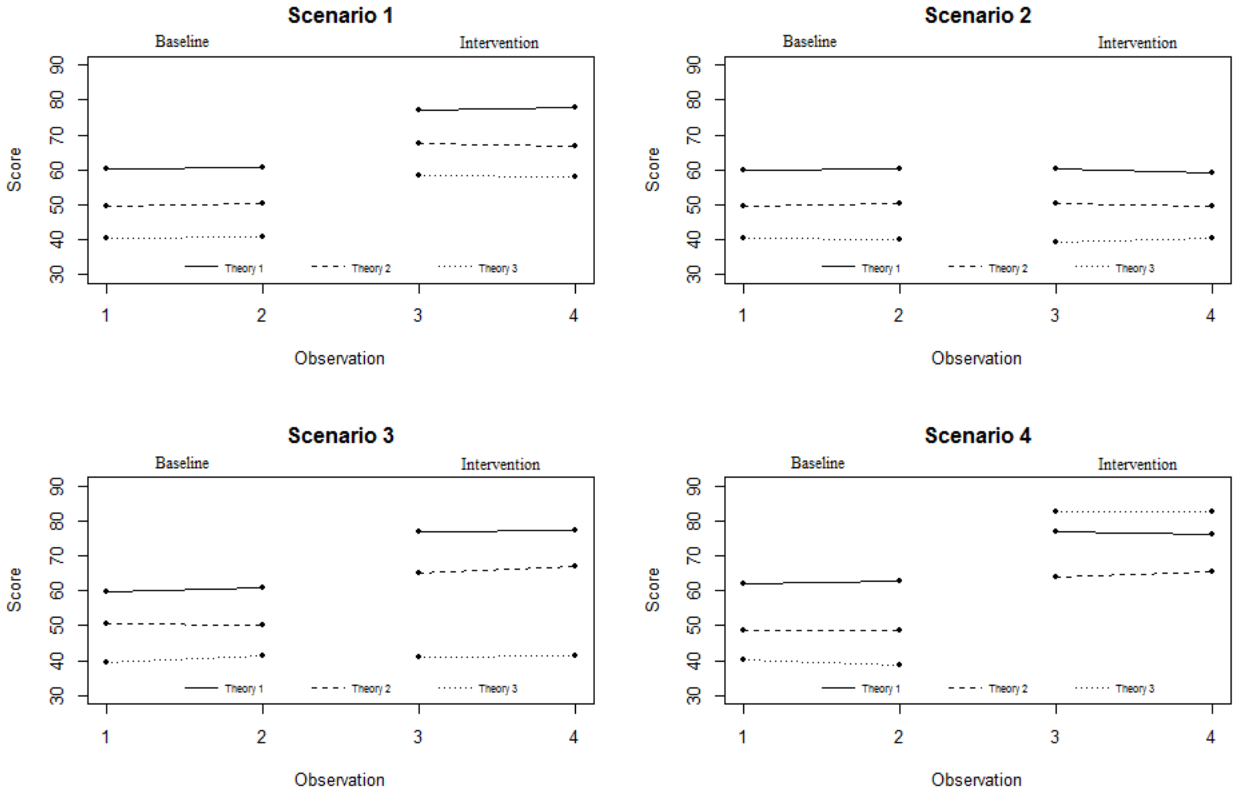
	Effect Consistency (Presence/Absence)	Magnitude Consistency
Scenario 1	Yes (Presence)	Yes
Scenario 2	Yes (Absence)	Yes
Scenario 3	No (a group with no effect)	No
Scenario 4	Yes (Presence)	No

Note. When effect consistency was present, the manner in which it was established (i.e. presence of effect across all theories or absence of an effect across theories) is noted in the text.

For each scenario, data were randomly generated[2] as mean outcomes from a normal distribution with arbitrary parameters that included a sample size of 100, a standard deviation of 5, and a specified mean. Across scenarios, theory-based means were established from a range of outcome values, roughly from 30 to 90, with relatively small, within-observation variability. These parameters were used to create four observations (i.e., time points) for each theory, for all scenarios. Differences between scores at time one and time two or at times three and four were attributable to chance.

[2] Data generation and graphics were completed using R (R Development Core Team, 2016), and all analyses were completed using IBM SPSS (IBM Corp. Released, 2012).

Figure 2: Results for four simulated scenarios of the CC design. Different theories are distinguished by solid, dashed, and dotted lines. In each instance, Theory 4, the theory of interest, is included along with each of the other three Theories (Theory 1 and Theory 4 appear together, etc.).



Descriptions and Analyses of Four Results Scenarios

Here, we describe individual scenarios and their corresponding analyses. Within the repeated-measures framework, we considered cases in which pre-intervention measures and post-intervention measures were stable. This configuration allowed us to average pre-intervention and post-intervention measures (rather than address four, distinct observations per theory) without substantially altering the scenarios themselves. This simplifying assumption made the discussion of outcome combinations much more straightforward. With scenarios involving unstable baselines, a repeated-measure ANOVA would be more applicable. Given these pre-intervention and post-intervention means, analyses were conducted using mixed-design ANOVA. If an interaction between the time and theory variables was statistically significant, multiple comparison tests (Tukey HSD) were conducted to determine which theories had different mean change.

Scenario 1. In Figure 2, we show results for each of four CC design scenarios. For the scenario in the upper-left corner of Figure 2, we have effect consistency and magnitude consistency. From the graphic, both types of consistency are visually evident. The implementation of X using the theory of interest (Theory 4) appears to produce consistently higher means, post-intervention. For many situations in which the CC design is applied, researchers will test interventions among theories yielding outcomes distinct from one another. These claims were statistically verified using ANOVA and multiple comparisons.

The main effect of time was statistically significant ($F(1, 297) = 3669.2,\ p < .001$, $\eta_{\mathrm{p}}^2 = .93$). This finding indicates that pre-intervention and post-intervention averages, for at least one of the theories, were different (i.e., there was a statistically significant mean change). However, there was no statistically significant interaction between time and theory ($F(2, 297) = 0.342, p = .711$). In other words, the statistically significant mean change did not vary by theory. From the combination of a statistically significant time main effect and a non-statistically significant interaction, we can conclude that both effect consistency and magnitude consistency exist. Because of the non-statistically significant interaction, we have magnitude consistency. And because time was statistically significant, it follows that mean differences among theories were essentially the same and that they differed from zero.

This scenario represents an ideal result pattern for the CC design. We have established a consistent, statistically significant effect (effect consistency) whose magnitude is uniform across all theories (magnitude consistency), with the intervention of interest being a common element among theories. Causal inference is quite plausible.

Scenario 2. In Scenario 2, we have both effect consistency and magnitude consistency. In this scenario, however, effect consistency is established as a result of the uniform *absence* of an effect across theories. It naturally follows that magnitude consistency will also be present.

In contrast to Scenario 1, the main effect of time in Scenario 2 was not statistically significant ($F(1, 297) = 0.903, p = .343$). This result indicates that, for all theories, means at pre-intervention and post-intervention measurements were essentially the same (i.e., no effect was present). From this single result, we can determine that, while magnitude consistency occurred, there was no consistent effect. This pattern of outcomes represents a worst case scenario for the CC design. Causal inference is not plausible.

Scenario 3. In Scenario 3, we do not have effect consistency, and as a direct result, do not have magnitude consistency. This conclusion follows as both Theories 1 and 2 appear to show a pre-post effect, whereas Theory 3 seems to not produce an effect. While the magnitude of the effect for Theories 1 and 2 appears roughly the same, because the effect of Theory 3 is essentially zero, the magnitude consistency standard is not met. As with Scenario 1, the main effect of time in Scenario 3 was found to be statistically significant ($F(1, 297) = 1529.0, p < .001$, $\eta_{\mathrm{p}}^2 = .84$). Again, this result indicates it is likely that pre-intervention and post-intervention means, for at least one of the theories, were different. Furthermore, there was a statistically significant interaction between time and theory ($F(2, 297) = 328.8, p < .001, \eta_{\mathrm{p}}^2 = .69$).

From this result, we conclude that the statistically significant mean change likely varied by theory. This pattern could be a combination of: 1) not all mean changes for theories were significantly different from zero or 2) magnitudes of mean change were not equal. Multiple comparison tests show that while theories 1 and 2 had statistically significant mean changes which were essentially the same, the third theory did not exhibit a mean change which statistically differed from zero. Now, for a different reason (effect inconsistency), causal inference is less plausible than in Scenario 1. This pattern of evidence may prompt researchers to investigate possible reasons for finding a single instance of no effect, within Scenario 3.

Scenario 4. Finally, in Scenario 4 we present a case with effect consistency but not magnitude consistency. As with Scenario 1, all theories appear to have some degree of effect. However, Theory 3 seems to have a much larger effect than Theories 1 and 2 (those with similar effects).

As with Scenario 3, both the main effect of time and the interaction of time and theory were statistically significant ($F(1, 297) = 6996.7, p < .001,\ \eta_{\mathrm{p}}^2 = .96$ for time and $F(2, 297) = 1025.8, p < .001, \eta_{\mathrm{p}}^2 = .87$ and for the interaction term). Multiple comparison tests again show that Theories 1 and 2 had mean changes which were statistically the same. As with Scenario 3, it was Theory 3 which differed from the other two theories. However, Theory 3 showed a mean change which was significantly larger than the other two theories. Thus, although we have evidence of effect consistency, we fail to meet the criteria of magnitude consistency. Causal inference is more certain than in Scenario 3 but *less* certain than in Scenario 1. Since each pre-post difference was statistically significant, a causal claim is enhanced. However, the more demanding standard that effect sizes be consistent did not occur.

Summarizing Conclusions Stemming from the Four Scenarios

These four scenarios are meant to inform readers of typical patterns for which the CC design is applicable. Graphical examples and corresponding statistical analyses were presented for each scenario. Some scenarios clearly establish or fail to establish causal inference; others represent varying degrees of ambiguity. It is not our intention to convince readers that perfect-like conditions (e.g., similar baseline levels, large and statistically significant effects in each unit) are intrinsic to the CC design. Careful judgment of a pattern of beneficial change is required.

An Application of CC Design Logic

When a version of the CC design has been implemented in the past, it appears to have been viewed as an analytic tactic rather than as a design. Phillips and Carstensen (1986) examined the possibility that nationally televised news or feature reports of suicide increased subsequent suicide rates among American teens. They examined 38 instances of such televised stories to determine if the number of suicides consistently increased one week before compared to the period one week after each TV report. The example provides an excellent opportunity to examine causal inference in a CC design applied within an existing research context.

From a threats to internal validity point of view, history is controlled as the TV reports are staggered across time. Testing is not relevant given the nature of the dependent variable. Maturation is implausible but may be problematic if news reports were systematically introduced during an uptrend in teen suicide rates. Regression is unlikely if the reports did not systematically occur soon after an atypically high period of teen suicides. Instrumentation is a potential problem in the CC design but was discounted by the *NEJM* study's analysis of "misclassification" (the authors used supplementary data to demonstrate that subsequent suicide rate increases were not an artifact of a coincident decrease in cases labelled "ambiguous").

The study offers a clear example of several inherent advantages of the CC design. Suicide rates were generally greater in the multiple instances in which TV stories were carried by a greater number of networks (application of a dose-response relationship). In addition, this multiple replication feature allowed researchers to test theory, as female teenagers had considerably greater suicide rates than males, and teen rates of suicide increased substantially while those for adults did not (since imitation is less plausible, for adults).

Weaknesses of the Common Cause Design

The CC design is not without its weaknesses. Like the often arduous task of identifying important covariates and demonstrating their baseline equivalence in non-randomized, between-groups designs, in the CC design an exhaustive search for alternative theories of change is the responsibility of the researcher. As a practical matter, the more groups utilized in the CC design (the more explanations that have been held constant), the greater its likely inferential power (the greater the chances that all relevant explanations have been identified).

The CC design relies on researcher's ability to demonstrate that competing explanations have remained constant during the period when data were collected. In the case of a possible legislative change (e.g., reduction in speed limits), mere knowledge that existing law has not been altered will be sufficient (though police enforcement of the speed limit could vary). But demonstration that other theories have remained constant will presuppose both availability and careful analysis of relevant data (e.g., archival records of monthly saturation levels of police patrolling). In addition to the consistency of competing explanations *within* each line of the CC design, researchers may also want to demonstrate that the pre-treatment, average rate of arrests for drunk driving was generally consistent *between* lines of the design.

While the CC design allows one to identify a plausible causal construct, that construct may have been confounded when given as an intervention. As Mill notes (Mill 1843, p. 458) "… the effect may have been produced not by the change, but by the means we employed to produce the change." For example, if a ticket followed by a fine for failure to stop at a red light was directly administered by police rather than by mail (perhaps after camera surveillance at dangerous intersections), the face-to-face confrontation with law enforcement rather than the loss of dollars may be the more critical component of change. The CC design does not protect against such misnaming.

It is possible that the competing explanations of each line of the CC design are not independent. Such complicated dynamics among factors may partially explain why competing explanations are functional in producing change. After all, the fundamental purpose of the CC design is to demonstrate that the intervention of interest has a causal impact rather than to rule out causal impacts of other potential explanations. From this perspective, exactly as Mill posited, the commonality of change in each line of the design is integral to the logical conclusion that the presence of the common element only helps to establish causal inference.

Ways to Enhance the Inferential Power of the Common Cause Design

Some weaknesses of the CC design can be easily rectified. As the credibility of change within each study unit is its most important, logical underpinning, any methodological tactic that strengthens this credibility will also enhance causal inference. Each unambiguous demonstration of change due to the theory of interest contributes to overall, causal inference.

As noted above, there are many validity threats intrinsic to the OXO design. However, a substantial increase in the number of pretests and posttests will enable the researcher to render many of these threats (testing, maturation, and regression) less plausible (Campbell & Stanley, 1966). The OXO design can also be strengthened by adding a control group to one or more units. If each control adequately matches baseline covariates, selection bias will be diminished, and pre-post change within lines of the design can be more clearly attributed to the theory being tested. But as Mill (1843) noted, utilizing a control group that lacks only the X actually incorporates the Method of Difference. While the conditions applicable to the Method of Difference occur less frequently than those applicable to the Method of Agreement, the simultaneous combination of the Method of Difference with the Method of Agreement ultimately seems best suited to demonstrate causal inference.

One may further imagine that the CC design has been embedded *within* a second design, say the non-equivalent control group (NECG) design. Now, instead of a single treatment condition in the NECG design, one finds the multiple treatment replications characterizing the CC design. Analogously, even a single, relatively weak, normative control group will add credence to conclusions stemming from the CC design. Multiple counterfactuals are now present (the possibility that all lines of the CC design show an effect *and* that there is no change from pre-intervention to posttest in the no-treatment control group).

Morgan and Winship (2007) provide a convenient list of strategies to enhance inference when an ITS occurs in each line of the CC design. To illustrate, the researcher might include multiple outcomes that change due to the presumed cause (e.g., drunk driving, speeding tickets, vehicular homicides) when the harsher punishment is administered. Rather than utilize a control *group*, a control *variable* serves as the counterfactual. The idea is straightforward; we hypothesize that the variable of interest will change but further contend that other, conceptually relevant dependent variables will not. In the above description, we claimed that instances of drunk driving would decrease in the face of larger fines. But, we do not expect that arrests for robbery or for tickets due to improper vehicle equipment would change.

In the drunk driving example, in those cities for which rates are higher to begin, we might expect larger decreases in drunk driving rates. Pre-intervention trend should be considered and several of the suggested statistical analyses allow for appropriate adjustment (e.g., growth curve analysis, repeated measures ANOVA). As discussed previously, the point in time when intervention is given can be systematically varied across cities (to control for history). It is also important to note if (and why!) any reduction occurred in non-intervention cites. Finally, if the fine for drunk driving returns to its previous level, one can determine if drunk driving averages revert to their pretreatment level.

The importance of establishing a dose-response relationship to enhance causal inference (Hill, 1965) has long been recognized as an important tool in public health research (Mill calls this the "Method of Concomitant Variation"). This methodological tactic would utilize different intervention doses within different lines of the CC design. For example, inference would be enhanced should aspirin be given at successively higher doses, each sub-grouping leading to proportionally more positive outcomes when other plausible contributors to the problem of interest (e.g., average age, treatment history, and duration of prior headache symptoms) are held constant in each line of the CC design.

When the X in each unit has been determined by a cut-point, the CC design can be considered as embedded within the regression discontinuity (RD) design. If many schools or classes receive an innovation, each determined by a cut-score, a consistent discontinuity in each unit would provide stronger evidence that the X was causal. The same logic can be applied to the RPD (regression point displacement design; Trochim & Shadish, 2014). If several cities within a state received extra funds for computers followed by a consistently positive, pre-post impact on achievement, the CC design suggests benefit. In addition, if these upward displacements in standardized achievement scores occurred for each intervention city when compared to the regression line established by the control cities in that same state not receiving funds, this pattern would represent a very favorable impact from the perspective of the RPD design. Finally, if the favorable pattern was replicated across states, for those cities receiving support for computers, we would conclude that the funds were well spent.

Potential Contributions of the Common Cause Design

There are numerous, positive elements germane to this new design. Unlike many quasi-experiments, the CC design does not require a control group. Researchers need not expend efforts to rule out selection bias germane to between-groups designs, which represents a substantial advantage to the methodologist's toolkit. With effect consistency, differential attrition becomes implausible. The CC design eliminates many threats to validity, especially those relevant to construct validity. The multiple replication feature of the design naturally enhances external validity. Well known and easily implemented data analysis procedures exist to determine if significant change occurred within each line of the CC design.

The CC design is intended to be used when it is possible to show agreement (consistency) in the effects (both existence and size) in each study unit. If individual study units allow strong tests of the claim of a theory (if individual units have strong research designs), then the CC design becomes less relevant. However, as many applied settings may not present opportunities for strong tests *in each unit*, the CC design allows one to patch together relatively weak, individual fabrics of evidence (e.g., those from a pretest posttest design) to bolster causal claims.

As noted above, the CC design can be embedded within quasi-experiments such as NECG, RD, and RPD designs. By adding and noting the incremental effects of increasing doses, the CC design might also be embedded within randomized controlled trials given in sequential fashion, an approach termed SMART (Sequential Multiple Allocation Randomized Trials) (Lei, Lynch, Oslin, & Murphy, 2012). The CC design is applicable to many different disciplines, especially those where competing theories are frequently encountered. Lastly, the translation of philosophical principles to new design structures offers considerable promise as a template to create other novel research designs.

Author Notes: William H. Yeaton currently teaches a research methods class at Florida State University. His most recent academic interests focus upon the conceptualization and application of embedded research designs (designs-within-designs) to better establish causal inference. Christopher G. Thompson is an Assistant Professor of Research, Measurement, and Statistics at Texas A&M University. His research interests are meta-analysis, Bayesian data analysis, quasi-experimental design, and applications of fuzzy sets to the social sciences.

References

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*, 297-312.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, *84*, 772-793.

Cook, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics, and economics. *Journal of Econometrics*, *142*(2), 636-654.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cook, T. D., Shadish, W. J., & Wong, V. C. (2008). Three conditions under which observational studies produce the same results as experiments. *Journal of Policy Analysis and Management*, *27*, 724-750.

Copri, I. M., & Cohen, C. (1990). *Introduction to logic* (8th ed.). New York: Macmillan.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyle.

Hill, A. B. (1965). The environment and disease: Association or causality? *Proceedings of the Royal Society of Medicine*, *58*, 295-300.

IBM Corp. Released (2012). IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Keane, M. B. (2014). *Fever: A novel*. New York, NY: Scribner.

Kowalski, C., Yeaton, W. H., Kuhr, K., & Pfaff, H. (in press). Helping hospitals improve patient centeredness: Assessing the impact of feedback following a best practices workshop. *Evaluation & the Health Professions*.

Krugman, P. (2015, July 17). Liberals and wages, *New York Times,* p. A27.

Lei, H., Nahum-Shani, I., Lynch, K., Oslin, D., & Murphy, S. D. (2012). A "SMART" design for building individualized treatment sequences. *Annual Review of Clinical Psychology, 8*, 21-48.

Mill, J. S. (1843). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. London: Harrison and Company.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. New York: Cambridge University Press.

Phillips, D. P., & Carstensen, L. L. (1986). Clustering of teenage suicides after television news stories about suicide. *New England Journal of Medicine*, *315*, 685-689.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.R-project.org

Rosenbaum, P. R. (1987). The role of a second comparison group in an observational study. *Statistical Science*. *2*, 292-316.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association 100*, 322-331.

Shadish, W. R. (2013). Propensity score analysis: promise, reality, and irrational exuberance. *Journal of Experimental Criminology*, *9*, 129-144.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association*, *103*, 1334-1343.

Shadish, W. R., Cook, T. D., & Campbell, T. D. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random to cutoff-based assignment. *Psychological Methods*, *16*, 179-191.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*, *52*, 123-147.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology 5*, 149-178.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation*, *35*, 311-327.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods*, *15*(3), 250-267.

Thistlewaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*, 309-307.

Trochim, W. M. K., & Campbell, D. T. (n.d.). The regression point displacement design for evaluating community-based pilot programs and demonstration projects. Retrieved January, 15, 2014, from http://www.socialresearchmethods.net/research/RPD.RPD.pdf.

Yeaton, W. H., & Moss, B. G. (*under review*). Below 2.0: Estimating the impact of academic probation on college student achievement by combining the randomized, regression discontinuity, and regression point displacement designs.