PROJECT MUSE®

Reflections on “Observational Studies”: Looking Backward and Looking Forward

Stephen G. West

Observational Studies, Volume 1, Number 1, 2015, pp. 231-240 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2015.0026>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/793425/summary>

Reflections on “Observational Studies”: Looking Backward and Looking Forward

Stephen G. West

sgwest@asu.edu

*Arizona State University and Freie Universität Berlin
Tempe, AZ 85287, USA*

Abstract

The classic works of William Cochran and Donald Campbell provided an important foundation for the design and analysis of non-randomized studies. From the remarkably similar perspectives of these two early figures, distinct perspectives have developed in statistics and in psychology. The potential outcomes perspective in statistics has focused on the conceptualization and the estimation of causal effects. This perspective has led to important new statistical models that provide appropriate adjustments for problems like missing data on the outcome variable, treatment non-compliance, and pre-treatment differences on baseline covariates in non-randomized studies. The Campbell perspective in psychology has focused on practical design procedures that prevent or minimize the occurrence of problems that potentially confound the interpretation of causal effects. It has also emphasized empirical comparisons of the estimates of causal effects obtained from different designs. Greater interplay between the potential outcomes and Campbell perspectives, together with consideration of applications of a third perspective developed in computer science by Judea Pearl, portend continued improvements in the design, conceptualization, and analysis of non-randomized studies.

1. Reflections on “Observational Studies”: Looking Backward and Looking Forward

William G. Cochran in statistics and Donald T. Campbell in psychology provided much of the foundation for the major approaches currently taken to the design and analysis of non-randomized studies in my field of psychology. The initial similarity of the positions taken by Cochran (1965, 1972, 1983) and Campbell (1957, 1963/1966) in their early writings on this topic is remarkable. Their work helped define the area and raised a number of key issues that have been the focus of methodological work since that time. Truly significant progress has been made in providing solutions to several of these key issues. Going forward to the present, work in statistics by Cochran's students (particularly Donald Rubin and his students) and in psychology by Campbell's colleagues (particularly Thomas Cook and William Shadish) have diverged in their emphases. Reconsideration of the newer work from the foundation of Cochran and Campbell helps identify some persisting issues.

2. Looking Backward

Cochran (1972) defined the domain of observational studies as excluding randomization, but including some agents, procedures, or experiences...[that] are like those the statistician

would call treatments in a controlled experiment... (p. 1). This definition reflects a middle ground between experiments and surveys without intervention, two areas in which Cochran had made major contributions (Cochran, 1950; 1963). The goal and the challenge of the observational study is causal inference: Did the treatment cause a change in the outcome? The domain established by Cochran's definition can still be considered relevant today. There is continued debate over exactly what quantities should be called "treatments in a controlled experiment" (e.g., Holland, 1986; Rubin, 2010). And some authors (e.g., Cook, Shadish & Wong, 2008; Rosenbaum, 2010; Rubin, 2006) appear to have narrowed Cochran's more inclusive definition of observational study to focus only on those designs that include baseline measures, non-randomized treatment and control groups, and at least one outcome measure. I will restrict the use of observational study to this narrower definition below, using the term non-randomized design to indicate the more inclusive definition.

Cochran (1972, section 3) discusses several designs that might be used to investigate the effects of a treatment in the absence of randomization. He also provides discussion of some potential confounders that potentially undermine the causal interpretation of any *prima facie* observed effects of the treatment. Campbell (Campbell & Stanley, 1963/1966) attempted to describe the full list of the non-randomized and randomized designs then available, including some he helped invent (e.g., the regression discontinuity design, Thistlethwaite & Campbell, 1957). Associated with each non-randomized design are specific types of potential confounders that undermine causal inference. Campbell attempted to enumerate a comprehensive list of potential confounders which he termed threats to internal validity. These threats represented "an accumulation of our fields criticisms of each other's research" (Campbell, 1988, p. 322). Among these are such threats as history, maturation, instrumentation, testing, statistical regression, and attrition in pretest-posttest designs, selection in designs comparing non-randomized treatment and control groups only using a posttest measure, and interactions of selection with each of the earlier list of threats in observational studies (narrow definition). Both Cochran and Campbell clearly recognized the differential ability of each of the non-randomized designs to account for potential confounders.

Echoing his famous earlier quoting of Fisher to "Make your theories elaborate" (Cochran, 1965, p. 252), Cochran (1972, p. 10) stated that "the investigator should think of as many consequences of the hypothesis as he can and in the study try to include response measurements that will verify whether these consequences follow." Campbell (1968; Campbell & Stanley, 1963/1966; Cook & Campbell, 1979) emphasized the similar concept of pattern matching in which the ability of the treatment and each of the plausible confounders to account for the obtained pattern of results is compared. Campbell emphasized that both response variables and additional design features (e.g., multiple control groups having distinct strengths and weaknesses; repeated pre-treatment measurements over time) be included in the design of the study to distinguish between the competing explanations.

Cochran (1972) also considered several of the ways in which measurement could affect the results of observational studies, notably the effects of accuracy and precision of measurement on the results of analyses and the possibility that measurements were non-equivalent in the treatment and comparison groups. Campbell and Stanley (1963/1966) included measurement-related issues prominently among their threats to internal validity and Campbell and Fiske (1957) offered methods of detecting potential biases (termed "method effects") associated with different approaches to measurement (e.g., different types of raters;

different measurement operations). Based on his experiences attempting to evaluate compensatory education programs, Campbell particularly emphasized the role of measurement issues, notably unreliability and lack of stability over time of baseline measurements, in producing artifactual results in the analysis of observational studies (Campbell & Boruch, 1975; Campbell & Erlebacher, 1970; Campbell & Kenny, 1999).

3. Looking Forward to the Present

From the initial similarity of the perspectives of Cochran and Campbell on non-randomized studies, their followers have diverged in their emphases. In statistics, Donald Rubin, one of Cochran’s students, has developed the potential outcomes approach to causal inference (Rubin, 1978; 2005; Imbens & Rubin, 2015), which provides a formal mathematical statistical approach for the conceptualization and the analysis of the effects of treatments. In psychology, Campbell’s colleagues have continued to develop aspects of his original approach focusing on systematizing our understanding of design approaches to ruling out threats to internal validity. I highlight a few of these differences below (see West & Thoemmes, 2010 for a fuller discussion). Table 1 summarizes the typical design and statistical analysis approaches to strengthening causal inference associated with some randomized and non-randomized designs that come out of the Rubin and Campbell perspectives, respectively.

3.1 Rubin’s Potential Outcomes Model

The potential outcomes model has provided a useful mathematical statistical framework for conceptualizing many issues in randomized and nonrandomized designs. This framework starts with the (unattainable) ideal of comparing the response of a single participant under the treatment condition with the response of the same participant under the control condition at the same time and in the same setting. Designs that approximate this ideal to varying degrees can be proposed including the randomized experiment, the regression discontinuity design, and the observational study. The potential outcomes framework forces out the exact assumptions needed to meet the ideal and defines mathematically the precise causal effects that can be achieved if these assumptions can be met. The framework draws heavily on Rubin’s (1976; Little & Rubin, 2002) seminal work on developing unbiased estimates of parameters when data are missing. Randomized experiments can be conceived of as a design in which no observations are available for treatment group participants in the control group and no observations are available for control group participants in the treatment group, but data are missing completely at random.

The potential outcomes framework permits the unbiased estimation of the magnitude of the *average* causal effect in experiments given that four assumptions are met: (a) successful randomization, (b) full compliance to the assigned treatment condition, (c) full measurement of the outcome variables, and (d) the stable unit treatment value assumption (SUTVA: participants outcomes are unaffected by the treatment assignments of others; no hidden variations of treatments). In cases when these assumptions break down (“broken randomized experiments”), new approaches requiring additional assumptions have been developed from the foundation of the potential outcomes model. Angrist, Imbens, and Rubin (1996) developed an approach that provides unbiased estimates of the causal effect for those participants *who would take the treatment in a randomized experiment if assigned*

Table 1: Key Assumptions/Threats to Internal Validity and Example Remedies for Randomized Experiments and Non-randomized Alternative Designs

Assumption or Threat to Internal Validity ^a	Typical Approaches to Mitigating the Threat	
	Design Approach	Statistical Approach
<u>Randomized Experiment</u> Independent units	temporal or geographical isolation of units	multilevel analysis; other statistical adjustment for clustering
Stable Unit Treatment Value Assumption (SUTVA): Other treatment conditions do not affect participant's outcome; No hidden variation in treatments	temporal or geographical isolation of treatment groups	statistical adjustment for measured exposure to other treatments
Full treatment adherence	incentive for adherence	instrumental variable analysis (assume exclusion restriction)
No attrition	sample retention procedures	missing data (assume missing at random)
<u>Regression Discontinuity Design</u> Functional form of relationship between assignment variable and outcome is properly modeled	Replication with different cutpoint; Nonequivalent dependent variables	Nonparametric regression; Sensitivity analysis
<u>Interrupted Time Series Analysis</u> Functional form of the relationship for the time series is properly modeled; another historical event, a change in population (selection), or a change in measures coincides with the introduction of the intervention.	nonequivalent control series in which intervention is not introduced; switching replication in which intervention is introduced at another time point; nonequivalent dependent measure.	Diagnostic plots (autocorrelogram; spectral density). Sensitivity analysis
<u>Observational Study</u> Measured baseline variables equated Unmeasured baseline variables equated Differential maturation	Multiple control groups; Nonequivalent dependent measures Additional pre-and-post intervention measurements	Propensity score analysis; Sensitivity analysis; Subgroup analysis

^a Note. The list of assumptions/threats to internal validity identifies issues that commonly occur in each of the designs. The alternative designs may be subject to each of the issues listed for the RE in addition to the issues listed for the specific design. The examples of statistical and design approaches for mitigating the threat to internal validity illustrate some commonly used approaches and are not exhaustive. For the observational study design, Rubin's and Campbell's perspectives differ so that the statistical and design approaches do not map 1:1 onto the assumptions/threats to internal validity that are listed. Reprinted from West, S. G. (2009). Alternatives to randomized experiments. *Current Directions in Psychology*, 18, 299–304.

to the treatment condition but take the control if assigned to the control condition (a.k.a., complier average causal effect; see Sagarin, West, Ratnikov, Homan, Ritchie, & Hansen, 2014 for a recent review of approaches to treatment non-compliance). Little and Rubin

(2002) and Yang and Maxwell (2014) offer methods for estimating unbiased causal effects when there is attrition from measurement of the outcome variable.

In the context of observational studies, Rosenbaum and Rubin (1983) developed propensity score analysis as a vehicle to adjust for the effect of a large number of covariates (potential confounders) measured at baseline (see Austin, 2011 and West, Thoemmes, Cham, Renneberg, Schultze, & Weiler, 2014 for recent reviews). Hong and Raudenbush (2005; 2013) extended propensity score analysis to provide proper estimates of the average causal effect when the treatment was delivered to a pre-existing group (e.g., treatments delivered to an existing classrooms of students) in group-based observational studies. Thoemmes and West (2011) developed approaches to the analysis of group-based observational studies in which the treatment is delivered to originally independent individuals who are constituted into groups for the purpose of the study. In each case, the potential outcomes perspective provides the foundation for conceptualizing the analysis, helps identify the necessary assumptions, and specifies the exact causal effect that may be estimated.

The potential outcomes approach has been particularly fruitful for the analysis of designs in which there are baseline covariates, outcome measures collected at a single time point, treatment and control conditions, and in which treatment assignment is assumed to be either independent of potential confounders or to be independent of all potential confounders after conditioning on covariates (ignorable). It has provided a valuable tool for analyzing randomized experiments, broken randomized experiments, the regression discontinuity design, and the observational study (narrow definition). However, the potential outcomes approach becomes more challenging to apply in designs in which measurement of the outcome variable is extended over time (e.g., interrupted time series designs, see Imbens & Wooldridge, 2009) or there are time-varying treatments (see Hong & Raudenbush, 2006). In addition, assumptions underlying the application of the potential outcomes approach (e.g., ignorability—no other unmeasured confounders exist) may not be testable, an important limitation.

3.2 Campbell’s Practical Working Scientist Approach

In psychology, design-based approaches have been given priority in the Campbell tradition over statistical adjustment of treatment effects: “When it comes to causal inference from quasi-experiments, design rules, not statistics” (Shadish & Cook, 1999, p. 300). The preference is to use the strongest design that can be implemented in the research context (Shadish, Cook, & Campbell, 2002). Advice is given to researchers about justifications that may be given to individual participants, communities, and organizations so that they will permit the strongest possible design to be implemented. Each of the potential threats to internal validity associated with the specific design is then carefully evaluated for plausibility in the research context. Attempts are then made to prevent or minimize the threat. For example, Ribisl et al. (1996) developed a valuable compendium of the then-available strategies of minimizing attrition in longitudinal designs (which needs updating to incorporate new technologies for tracking and keeping in contact with individual participants). Efforts to prevent the threat are supplemented by the identification of design elements that specifically address each identified threat (e.g., multiple pretests; multiple control groups; see Shadish et al., 2002, p. 157). Then, the pattern of obtained results is compared to the results predicted by

the hypothesis and by each of the threats to internal validity (confounders). Shadish et al. present illustrations of the use of this strategy and numerous research applications in which it has been successful. Campbell's approach does not eschew the elegant statistical analysis approaches offered by the potential outcomes framework (indeed, it welcomes them), but it gives higher priority to design solutions.

There are two key difficulties in applying Campbell's approach to non-randomized designs. First, the researcher's decision to rule out or ignore certain threats to internal validity *a priori* may be incorrect. In Campbell's approach, there are suggestions to use common sense, prior research, and theory to eliminate some potential threats. While often a good guide, common sense, prior research, and theory might be misleading. Second, although the pattern matching strategy can be compelling when the evidence supporting one hypothesis is definitive, more ambiguous partial support for several competing hypotheses can also occur. Rosenbaum (2010) offers some methods of integrating results across multiple design elements, but methods of formally assessing the match of the obtained results to the hypothesis and each of the plausible threats to validity need further development.

Campbell emphasized a practical approach that mimics the approach of the working scientist. One intriguing development within the Campbell tradition is the empirical comparison of the results of overlapping study designs in which a randomized and non-randomized design are both used. A series of papers have compared non-randomized designs to the gold standard of the randomized experiment under conditions in which the two designs share a common treatment group and ideally sample from the same population of participants. No difference was found between the estimates of the magnitude of the treatment effect from the randomized experiment and the regression discontinuity design (Cook & Wong, 2008; see also Shadish et al., 2011 for a randomized experiment comparing the estimates of treatment effects from the two designs). Similarly, no differences were found between the estimates of the treatment effect from a randomized experiment and interrupted time series design (St. Clair, Cook, & Halberg, 2014). In contrast, syntheses of existing comparison studies (Cook, Shadish, & Wong, 2008) as well as randomized experiments comparing the effect sizes for participants randomly assigned to treatment and control conditions or randomly assigned to self-select between the same treatment and control conditions (Shadish, Clark, & Steiner, 2008) identified cases in which the two designs produced comparable and non-comparable results. Several factors facilitated obtaining comparable results (Cook, Shadish, & Wong, 2008; Cook, Steiner, & Pohl, 2009; Steiner, Cook, Shadish, & Clark, 2010): (a) a known selection rule that determines assignment to treatment and control conditions, (b) inclusion of measures of all relevant confounders in the statistical adjustment model, (c) inclusion of a pretest measure of the outcome variable in the statistical adjustment model, (d) reliable measurement of the covariates, and (e) control participants being selected from the same population as the treatment participants (originally descriptively termed a "focal, local" control group). Generalization of the results of these studies was originally limited by the small number of research contexts in the existing studies upon which the conclusions were based. However, in an ongoing project more than 65 studies of this type have been identified and compiled thus far and this database continues to expand. Synthesis of these studies is increasingly providing a practical basis for the design of non-experimental studies that can help minimize bias in the estimates of causal effects of treatments.

4. Conclusion

Cochran and Campbell helped define the domain of non-randomized studies and the threats to internal validity (potential confounders) that may compromise the interpretation of observed treatment effects. Subsequent work in statistics and psychology has taken complementary paths. Work in the potential outcomes model tradition in statistics has substantially improved the conceptualization of causal effects and has provided improved estimates of their magnitude. Work in psychology following in the tradition of Campbell has emphasized the development of practical methods for researchers to improve the design of non-randomized studies. Some of this work has built on the insights of Cochran and Campbell to make theories complex and to design studies so that pattern matching may be used to distinguish between potential explanations. Some of this work has re-emphasized the importance of some of Cochran and Campbell’s views about the key status of baseline measures of the outcome variable and the importance of highly reliable measurement of baseline measures in non-randomized designs, features that have received less attention in the potential outcomes approach. Within the potential outcomes framework, less attention has been given to the development of new methods in which the observations are measured repeatedly over time or time-varying treatments are implemented (but see e.g., Robins & Hernán, 2009). Within the Campbell framework, practical methods for strengthening causal inferences in the interrupted time series have been developed (Shadish et al., 2002) and new work has focused on improving the design and analysis of the single subject design (Shadish, 2014; Shadish, Hedges, Pustejovsky, Rindskopf, Boyajian, & Sullivan, 2014). And within both approaches the development of formal methods for assessing Cochran’s elaborate theories or Campbell’s pattern matching has received relatively little attention. This situation may be beginning to change. Judea Pearl’s (2009) approach to causal inference developed in computer science is being applied to systems of variables. Although Pearl’s approach can be seen as having its own limitations (Shadish & Sullivan, 2012; West & Koch, 2014), it has helped sharpen our conceptualization of some causal inference problems (e.g., mediation in which treatment causes changes in intermediate variables [mediators], which, in turn, produce changes in the outcome variable; which confounders should and not be controlled in observational studies). It has also provided challenges to the potential outcomes approach given its alternative (but overlapping) approach to the conceptualization and estimation of causal effects. Although important advances have occurred since the foundational work of Cochran and Campbell, greater interplay among the potential outcomes, Campbell, and Pearl perspectives portends continued improvements in our design, conceptualization, and analysis of non-randomized studies.

Acknowledgments

I thank Thomas D. Cook, William R. Shadish, and Dylan Small for their comments on an earlier version of the manuscript.

References

- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, 91, 444–472.
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399–424.
- Campbell, D. T. (1957). Factors affecting the validity of experiments in social settings. *Psychological Bulletin*, 54, 297–332.
- Campbell, D. T. (1966). Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of Egon Brunswick* (pp. 81–106). Holt, Rinehart, & Winston. New York.
- Campbell, D. T. (1988). Can we be scientific in applied social science? In E. S. Overman (Ed.), *Methodology and epistemology for social science: Selected papers of Donald T. Campbell*. University of Chicago Press, Chicago.
- Campbell, D. T., and Boruch, R. F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiments: Some critical issues in assessing social programs* (pp. 195–296). Academic Press, New York.
- Campbell, D. T., and Fiske, D. W. (1957). Convergent and discriminant validation by the multrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Campbell, D. T., and Erlebacher, A. E. (1970). How regression artifacts can mistakenly make compensatory education programs look harmful. In J. Hellmuth (Ed.), *The disadvantaged child: Vol. 3, Compensatory education: A national debate* (pp. 185–210). Brunner/Mazel, New York.
- Campbell, D. T., and Kenny, D. A. (1999). *A primer on regression artifacts*. Guilford, New York.
- Campbell, D. T., and Stanley, J. C. (1963/1966). *Experimental and quasi-experimental designs for research*. Rand McNally, Chicago. Originally published in N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Rand McNally, Chicago.
- Cochran, W. G. (1965). The planning of observational studies (with discussion). *Journal of the Royal Statistical Society, Series A*, 128, 234–266.
- Cochran, W. G. (1972). Observational studies. In T. A. Bancroft (Ed.) (1972), *Statistical papers in honor of George W. Snedecor* (pp. 77–90). Ames, IA: Iowa State University Press. Reprinted in *Observational Studies*, 1, 126–136.
- Cochran, W. G. (1983). *Planning and analysis of observational studies*. Wiley, New York.
- Cochran, W. G., and Cox, G. M. (1950). *Experimental designs*. Wiley, New York.
- Cochran, W. G. (1953). *Sampling techniques*. New York, NY: Wiley.
- Cook, T. D., and Wong, V. C. (2008). Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique*, 91–92, 127–150.
- Cook, T. D., Shadish, W. R., and Wong, V. C. (2008). Three conditions under which observational studies produce the same results as experiments. *Journal of Policy Analysis and Management*, 27(4), 724–750.

- Cook, T. D., Steiner, P. M., and Pohl, S. (2009). Assessing how bias reduction is influenced by covariate choice, unreliability and data analytic mode: An analysis of different kinds of within-study comparisons in different substantive domains. *Multivariate Behavioral Research*, 44, 828-847.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945-960.
- Hong, G., and Raudenbush, S. W. (2006). Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, 101(475), 901-910.
- Hong, G., and Raudenbush, S. W. (2013). Heterogeneous agents, social interactions, and causal inference. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 331-352). Springer Netherlands.
- Imbens, G. W., and Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press, New York.
- Imbens, G. W., and Wooldridge, J. W. (2009). Recent developments in the economics of program evaluation. *Journal of Economic Literature*, 47, 5-86.
- Little, R. J. A., and Rubin, D. B. (2002). *Statistical inference with missing data (2nd Ed.)*. Wiley, Hoboken, NJ.
- Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., and Bootsmiller, B. J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning*, 19(1), 1-25.
- Robins J. M., and Hernán M. A. (2009). Estimation of the causal effects of time-varying exposures. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds), *Advances in Longitudinal Data Analysis*. Chapman and Hall, New York.
- Rosenbaum, P. R. (2010). Design of observational studies. New York: Springer.
- Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 34-58.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100, 322-331.
- Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press, New York.
- Rubin, D. B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, 15(1), 38-46.
- Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., and Hansen, E. J. (2014). Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological Methods*, 19(3), 317-333.
- Shadish, W. R. (2014). Statistical analyses of single-case designs: The shape of things to come. *Current Directions in Psychological Science*, 23, 139-146.
- Shadish, W. R., M. H. Clark and Peter M. Steiner (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment (with commentary). *Journal of the American Statistical Association*, 103, 1334-1356.

- Shadish, W. R., and Cook, T. D. (1999). Design rules: More steps towards a complete theory of quasi-experimentation. *Statistical Science*, 14, 294–300.
- Shadish, W. R., Cook, T. D., and Campbell, T. D. (2002). *Experimental and quasi-experimental designs for general causal inference*. Wadsworth, Boston.
- Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., and Cook, T. D. (2011). A randomized experiment comparing random to cutoff-based assignment. *Psychological Methods*, 16(2), 179–191.
- Shadish, W.R., Hedges, L.V., Pustejovsky, J., Rindskopf, D.M., Boyajian, J.G. and Sullivan, K.J. (2014). Analyzing single-case designs: *d*, *G*, multilevel models, Bayesian estimators, generalized additive models, and the hopes and fears of researchers about analyses. In T. R. Kratochwill and J. R. Levin (Eds.), *Single-Case Intervention Research: Methodological and Statistical Advances* (pp. 247–281). American Psychological Association, Washington, D.C.
- Shadish, W.R. and Sullivan K. (2012). Theories of causation in psychological science. In H. Cooper (Ed.), *APA Handbook of Research Methods in Psychology* (Vol. 1, pp. 23–52). Washington, D.C.: American Psychological Association.
- St. Clair, T., Cook, T. D. and Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted times series design by comparison with a randomized experiment. *American Journal of Evaluation*, 35(3), 311–327.
- Thistlethwaite, D. L., and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51, 309–317.
- Thoemmes, F. J., and West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research*, 46(3), 514–543.
- West, S. G., and Koch, T. (2014). Restoring Causal Analysis to Structural Equation Modeling. Review of Judea Pearl, *Causality: Models, Reasoning, and Inference* (2nd. Ed). *Structural Equation Modeling*, 21, 161–166.
- West, S. G., and Thoemmes, F. (2010). Campbell’s and Rubin’s perspectives on causal inference. *Psychological Methods*, 15, 18–37.
- West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., and Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of consulting and clinical psychology*, 82(5), 906–919.
- Yang, M., and Maxwell, S. E. (2014). Treatment effects in randomized longitudinal trials with different types of non-ignorable dropout. *Psychological Methods*, 19, 188–210.