

Alternatives to Randomized Experiments

Stephen G. West

Alternatives to Randomized Experiments

Stephen G. West

Arizona State University

Address correspondence to Stephen G. West, Psychology Department, Arizona State

University, Tempe, AZ 85287-1104; e-mail: sgwest@asu.edu.

ABSTRACT—Randomized experiments are preferred for making inferences about causality when they can be implemented and their assumptions are met. Yet assumptions can fail (e.g., attrition, treatment noncompliance) or randomization may be unethical or infeasible. I describe alternative design and statistical approaches that permit testing causal hypotheses and present current empirical evidence related to alternative designs. Alternative designs permit a wider range of research questions to be answered and permit more direct generalization of causal effects; however, when using such designs, estimates of the magnitude of the causal effect may be more uncertain.

KEYWORDS—experiment; quasi-experiment; causal inference

The randomized experiment (RE) enjoys a reputation as the gold standard of research designs. When this design can be properly implemented and its assumptions are met, it enables making strong, transparent inferences about causality that are unrivalled by those produced by other designs. These conditions can often be met in short-term laboratory experiments. However, as experiments become extended in time to weeks or years, attrition may occur—some people no longer participate. Or important active treatments (e.g., medical procedures, educational interventions) may incorporate risks or control treatments may fail to incorporate benefits, so that participants may seek out and receive other than their assigned treatment. In many settings, randomization may be infeasible or unethical: Consider proposing a RE on the health outcomes of hurricane Katrina or of secondhand tobacco smoke. And sometimes, only atypical participants may agree to be randomized or be willing to fully participate, for example, in a study of the effects of a religion-based program for alcoholism.

TWO COMPLEMENTARY PERSPECTIVES

Problems with REs or an inability to randomize do not mean that causal hypotheses cannot be entertained. Two complementary perspectives provide useful tools for strengthening causal inferences in both randomized and nonrandomized designs.

Campbell's Perspective

Campbell's (e.g., Shadish, Cook, & Campbell, 2002) perspective emphasizes the identification of plausible threats to the validity of causal inference. Threats are confounders that provide alternative explanations for the results. Campbell and colleagues have compiled an extensive list of common threats that may occur in the full range of research designs. The specific confounders that threaten a given study depend on

the research design and prior empirical findings in the research area. Cook and Campbell (1979) identified four general types of validity threats (statistical conclusion, internal, construct, and external). Internal validity focuses on threats relevant to causal inference. Whenever a plausible threat is identified, the researcher's task is to incorporate additional design elements that rule out that threat. Specific patterns of results in combination with design elements can permit the researcher to logically eliminate the threat (see examples below). Among design elements, randomization is given a privileged status because of its ability, when successfully implemented, to rule out the widest range of plausible threats. The mutual criticism fostered by peer review serves to expose plausible threats not identified by the researcher.

Rubin's perspective

Rubin's potential-outcomes perspective (e.g., Holland, 1986; Rubin, 2005), influential in medicine and public health, is a deductive mathematical approach based on making explicit, ideally verifiable assumptions (see Table 1 for examples). Rubin defines a causal effect as the difference between the outcomes for the same single unit (e.g., person, community) given two different treatments at the identical time and in the identical context. This definition is a Platonic ideal, unrealizable in practice since each unit can only receive one treatment at a given time. The ideal can be approximated through within-subject designs, identical units (e.g., matching), or randomization—the preferred method when feasible because REs involve the least stringent assumptions. When assumptions are violated, design or statistical-analysis approaches must be taken to minimize bias in estimation of the causal effect.

In Rubin's perspective, a key conceptual comparison is between the outcome for the treatment received for a single participant and the *potential outcome* that would have occurred if the same participant had received the alternative treatment. Based on the expectation that randomization will theoretically equate the treatment and control groups, on average, on all participant characteristics at baseline, the RE permits unbiased estimates of the magnitude of the *average* causal effect given that its assumptions are met (see Table 1). The magnitude of the causal effect will typically vary across individuals.

BROKEN RANDOMIZED EXPERIMENTS

Even originally well-designed REs can turn into "broken" REs. Attrition of participants and their nonadherence to the assigned treatment are common violations of the RE's underlying assumptions. Campbell's and Rubin's perspectives offer researchers strategies for strengthening causal inferences.

Attrition

Any participant whose outcome cannot be measured can potentially bias the estimate of the causal effect. Equipment malfunctions, participants decline further participation, or they fail to answer questions. Any participant characteristic related to both the outcome and the probability of attrition can bias the results. For example, in an educational study, low income may be related both to the likelihood that a child misses outcome measurement and to the degree of improvement from the treatment. On the design side, procedures to help retain participants have been developed (Shadish et al., 2002). On the statistical analysis side, modern missing-data methods (Schafer & Graham, 2002) permit appropriate adjustment of the results for any variable that is missing at random (MAR). When a variable is MAR, whether or not its value is observed is

accounted for by other variables in the data set (e.g., an expensive medical test is given only to patients with scores above a threshold on an inexpensive screening test).

Adjustment may also substantially reduce bias in the results associated with other variables whose source of missingness is unknown to the researcher. In planning research, measurement of possible predictors of participant attrition can improve the effectiveness of later adjustment.

Treatment Nonadherence

Particularly when important benefits, costs, or risks are associated with a treatment, not all participants may adhere to their assigned treatment condition. In an RE of screening mammography for early diagnosis of breast cancer, approximately one third of women assigned to the screening condition refused screening (Baker, 1998). Researchers have developed design strategies that increase treatment adherence (e.g., monetary incentives), but many of these potentially decrease the ability to generalize the results. Statistically, many researchers have followed the traditional intention-to-treat strategy (ITT)—“analyze them as you have randomized them”—the actual treatment received is ignored. This strategy may produce a substantial loss of statistical power; it can also lead to potential bias in the estimate of the magnitude and even the direction of the causal effect if both treatment nonadherence and attrition have occurred. Other simple statistical procedures (e.g., analysis by treatment received; see West & Sagarin, 2000) often produce highly misleading results. Rubin has advocated comparing the treatment and control outcomes only for adherers, individuals who would adhere to the treatment regardless of the treatment condition to which they would be assigned.. The challenge of this strategy is that researchers can only observe who adheres to the treatment in the

active-treatment condition. Those individuals in the control condition who would have adhered *if* assigned to the treatment condition are unknown. Experimental comparisons may be problematic if they include individuals in the control condition who would never have accepted the active treatment. Given randomization and a key additional assumption known as the exclusion restriction—the effect of treatment assignment operates fully through the actual treatment and not through other processes—the effect due to nonadherers can be statistically removed from *both* the treatment and control groups. The exclusion restriction is often plausible in practice; it is always met given successful masking of treatment assignment from participants and treatment providers. This approach can provide an unbiased, useful estimate: that of the causal effect of treatment for those who would actually receive it.

QUANTITATIVE ASSIGNMENT DESIGNS

Regression Discontinuity Design (RDD)

Ethical objections to randomization can sometimes be bypassed by taking advantage of known societal assignment rules. Intervention programs (e.g., school lunches) are given to those below a specified level (cutpoint) of measured baseline risk or need (e.g., family income). Gifted education programs are given to those who exceed a specified achievement test score. In the RDD, treatment is assigned on the basis of a measured quantitative baseline variable. The relationship between the measured baseline variable and the outcome is modeled separately for individuals below and above the cutpoint. The difference in the estimates of the levels of estimated outcome at the cutpoint represents the causal effect.

Mark and Mellor (1991) studied hindsight bias in union workers. The measured baseline variable was seniority. In an economic downturn, workers with 20 or more years of seniority retained their jobs, whereas those with less than 20 years were laid off. At the cutpoint of 20 years, retained workers reported that the layoffs were far more predictable, relative to laid-off workers. The RDD provides strong, unbiased estimates of the causal effect at the cutpoint. The key new additional assumption is that the form of the relationship between the baseline (assignment) variable and the outcome is properly represented in the statistical model (see Table 1).

Interrupted Time Series (ITS)

Other interventions occur at a specific point in time. A new policy (e.g., no-fault divorce) goes into effect on a specific date; or a single patient could have a treatment introduced or withdrawn at specific time points. If a large number of pre- and post-treatment observations can be taken at equally spaced time intervals, the ITS design permits careful modeling of the series of observations and a strong estimate of the causal effect, given that its assumptions are met (see Table 1).

Khuder et al. (2007) analyzed monthly hospital admissions over a 72-month period in an Ohio city. A ban on indoor smoking in the city was implemented after 36 months. Admissions for coronary heart disease decreased after the ban. However, causal inferences in the ITS may be threatened if other events occur (e.g., new heart medicine introduced), record-keeping practices change, or the patient mix (selection) changes simultaneously with the intervention. Additional design elements are needed to address these threats. Khuder et al. also analyzed coronary heart disease admissions in another similar Ohio city that did not institute a smoking ban, finding no change at the

comparable time point. They also found no change in either city for hospital admissions for non-smoking-related diseases. This clear pattern of results produced by ITS designs with added design elements nicely rules out plausible validity threats.

OBSERVATIONAL STUDIES

Even in weaker designs, the two perspectives provide strategies that allow researchers to maximize the strength of causal inference. In observational studies (also called nonequivalent-control-group design) participants are measured at baseline and then again after the intervention. Participants may receive a control or active treatment, but the rule for assignment to (selection into) treatment conditions is unknown to the researchers. From Campbell's perspective, participants in the two groups may be maturing at different rates, have differential regression to the mean, have different histories, or the baseline and outcome measures may have different psychometric properties. Design elements must be added to rule out plausible threats.

Reynolds and West (1987) compared the effectiveness of a sales campaign to no campaign in increasing sales of state lottery tickets in convenience stores. Store managers refused randomization. For each treatment store, a similar control store was chosen—stores from the same chain were matched on prior lottery ticket sales and neighborhood. Sales in the game in which the campaign was implemented were measured in each store as the outcome. Treatment stores showed (a) an increase in sales of lottery tickets relative to control stores, (b) an increase in lottery tickets relative to other sales categories, and (c) a similar level of sales to the control store in the weeks prior to the beginning of the campaign but higher sales thereafter (Fig. 1). Additional design elements, (b) and (c)—described more fully in the caption to Figure 1—greatly strengthened causal inference.

Rubin's perspective emphasizes close matching of treatment and control group participants at baseline on *all* covariates believed to be potentially related to both treatment-group selection and the outcome. If all such covariates have been identified and the relationship of the covariates to treatment-group assignment has been properly modeled, then the baseline distributions in the two groups will theoretically be balanced on all measured covariates, closely paralleling the RE (Rosenbaum & Rubin, 1983).

West and Thoemmes (in press) reanalyzed data from Wu, West, and Hughes' (2008) study of the effects of being held back in first grade (retention) versus promotion to second grade. All 72 variables that subject-matter experts suggested might possibly be related to retention and/or math achievement were measured at baseline. Propensity scores—each child's probability of being subsequently retained based on the 72 baseline variables—were estimated using logistic regression. From 784 low-achieving children, a subsample of 97 pairs, in which one child was subsequently retained and the other promoted, could be closely matched on their propensity scores. This procedure resulted in groups of retained and promoted children who had similar means (and frequency distributions) on nearly all of the 72 individual baseline variables. Retained students showed less gain in raw math achievement during the retention year than did promoted students, and they showed no increase in their rate of gain following the retention year. The propensity-score procedure rules out possible selection effects on measured baseline variables. However, this procedure does not rule out other unmeasured baseline variables on which the children might differ (and have been overlooked by the subject matter experts), if such variables exist. The robustness of the causal effect can be explored with sensitivity analyses describing how the magnitude of the causal effect would change

across a range of plausible scenarios involving unmeasured baseline variables. In addition, causal effects are only defined for the precise range of data over which treatment and control participants can be successfully matched. Other procedures permit inadvertent but unwarranted extrapolation of the causal effect (here, retention) to inappropriate subpopulations (e.g., high-achieving children who would never be retained).

HOW WELL DO ALTERNATIVE DESIGNS WORK?

Early reviews considered all available randomized and nonrandomized studies involving diverse sets of treatments and participants and came to pessimistic conclusions about the comparability of their results. A meta-analysis (Thoemmes, West, & Hill, 2009) showed that other features of methodological quality (e.g., was attrition addressed?) also differ and account for much of the obtained difference in effect sizes between randomized and nonrandomized intervention studies. Cook, Shadish, and Wong (2008) provided a more focused comparison, considering only available studies that included a RE and a nonrandomized alternative design that shared the identical treatment condition. ITS and RDD led to estimates of the causal effect that did not differ from those produced by the RE. Observational studies also produced similar causal-effect estimates given that a control group of similar participants was used or the selection mechanism was known. Hernán et al. (2008) reanalyzed data from an RE (Women's Health Initiative) and an observational study (Nurses Health Study) of postmenopausal hormone therapy on coronary heart disease that produced highly disparate results. When the observational study was analyzed using propensity-score methods and the same causal effect (ITT) for the same treatment for similar participants who met the same eligibility criteria was

estimated, discrepancies were minimal. Shadish, Clark, and Steiner (2008) randomly assigned participants to a RE or observational study of the effects of math or vocabulary training, finding little difference in the estimates of the causal effect after adjusting for an extensive set of baseline covariates in the observational study. The available limited set of focused comparisons provide the encouraging finding that similar estimates are obtained from REs and non-randomized alternative designs if the same causal effect for identical treatments and similar populations is estimated. Additional focused comparisons are needed to determine the generality of this finding.

CONCLUSION

When properly implemented, REs require the fewest assumptions and lead to the most transparent causal inferences. Violations of assumptions, such as attrition and treatment nonadherence, muddy this transparency. Each nonrandomized alternative design involves the four assumptions listed in Table 1 for the RE plus its own new assumptions. With each additional assumption comes increased uncertainty about the magnitude of the causal effect. These assumptions must be addressed through design or statistical-analysis strategies to reduce uncertainty and place proper bounds on the estimate of the causal effect. Given a choice, addition of design elements is preferred over statistical-analysis strategies (Shadish et al., 2002).

Alternative nonrandomized designs have the advantage of permitting scientists to address a wider range of important research questions rather than abandoning or altering them because they cannot be placed in the sometimes Procrustean bed of the RE. Alternative designs do not depend on having a sufficient number of units to randomize, which often privileges interventions focused on individuals or small groups rather than

larger scale (e.g., city-wide or state-wide) interventions that might have a larger societal impact. Finally, alternative designs also often permit more transparent generalization of causal effects to the specific population, setting, treatment, and outcome of interest, which is important in many applied and policy settings.

Recommended Reading

Imbens, G., & Rubin, D. (in press). (See References). The most complete and advanced exposition of Rubin's potential outcomes perspective and its application to various designs.

Shadish, W.R., & Cook, T.D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, 60, 607–629. An up-to-date review of recent developments in experimental and quasi-experimental designs.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). (See References). A good general reference with the most complete presentation of Campbell's perspective and its application to various designs.

West, S.G., Biesanz, J.C., & Pitts, S.C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H.T. Reis & C.M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84). New York: Cambridge. A chapter applying Campbell's and Rubin's perspectives to common experimental and quasi-experimental designs.

West, S.G., & Thoemmes, F. (in press). (See References). A thorough comparison of Campbell's and Rubin's perspectives on causal inference.

REFERENCES

- Baker, S.G. (1998). Analysis of survival data from a randomized trial with all-or-none compliance: Estimating the cost effectiveness of a cancer screening program. *Journal of the American Statistical Association*, *93*, 929–934.
- Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- Cook, T.D., Shadish, W.R., & Wong, V.C. (2008). Three conditions under which experiments and observational studies produce comparable causal effects: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750.
- Hernán, M.A., Alonso, A., Logan, R., Grodstein, F., Michels, K.B., Willett, W.C., Manson, J.E., & Robins, J.M. (2008). Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease (with discussion). *Epidemiology*, *19*, 766–792.
- Holland, P.W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, *81*, 945–970.
- Imbens, G., & Rubin, D. (in press). *Causal inference: Statistical methods for estimating causal effects in biomedical, social, and behavioral sciences*. New York: Cambridge.
- Khuder, S.A., Milz, S., Jordan, T., Price, J., Silvestri, K., & Butler, P. (2007). The impact of a smoking ban on hospital admissions for coronary heart disease. *Preventive Medicine*, *45*, 33–38.

- Mark, M.M., & Mellor, S. (1991). Effect of self-relevance of an event on hindsight bias: The foreseeability of a layoff. *Journal of Applied Psychology, 76*, 569–577.
- Reynolds, K.D., & West, S.G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review, 11*, 691–714.
- Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.
- Rubin, D.B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*, 322–331.
- Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.
- Shadish, W.R., Clark, M.H., & Streiner, P.M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments (with commentary). *Journal of the American Statistical Association, 103*, 1334–1356.
- Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Thoemmes, F., West, S.G., & Hill, E. (2009, May). *Propensity score matching in a meta-analysis comparing randomized and non-randomized studies*. Paper presented at the Atlantic Causal Modeling Conference, Philadelphia, PA.
- West, S.G., & Sagarin, B.J. (2000). Subject selection and loss in randomized experiments. In L. Bickman (Ed.), *Contributions to research design: Donald Campbell's legacy* (Vol. 2, pp. 117–154). Thousand Oaks, CA: Sage.

West, S.G., & Thoemmes, F. (in press). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*.

Wu, W., West, S.G., & Hughes, J.N. (2008). Effect of retention in first grade on children's achievement trajectories over four years: A piecewise growth analysis using propensity score matching. *Journal of Educational Psychology, 100*, 727–740.

Fig. 1. Some approaches using design elements to strengthen causal inferences in observational studies. (a) Matching: Treatment and control stores are selected from the same chain, are in the same geographical location, and are comparable in sales during baseline (lottery game 10); introduction of the treatment during lottery game 11 yields an increase in sales only in the treatment stores. (b) Nonequivalent dependent variables: Within the treatment stores; sales of lottery tickets increase substantially following the introduction of treatment; sales of other major categories (gasoline, cigarettes, groceries (nontaxable), and groceries (taxable) that would be expected to be affected by confounding factors, but not treatment, do not show appreciable change. (c) Repeated pre- and post-test measurements: Treatment- and control-stores sales show comparable trends in sales during the 4 weeks prior to and following the introduction of the treatment; the levels of sales in the treatment and control scores are similar prior to the introduction of treatment but differ substantially beginning immediately after treatment is introduced. Adapted from “A Multiplist Strategy for Strengthening Nonequivalent Control Group Designs,” by K.D. Reynolds & S.G. West, (1987), *Evaluation Review, 11*, (pp. 698, 701, 794). Copyright 1987, SAGE. Adapted with permission.

/tc/**TABLE 1**

Threats to Internal Validity/Key Assumptions and Example Remedies for Randomized Experiments and Alternatives

Assumption or threat to internal validity

Approaches to mitigating the threat

	<u>Design approach</u>	<u>Statistical analysis strategy</u>
Randomized controlled experiment		
Independent units assumption	Geographical or temporal 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 variations in treatments	Geographical or temporal isolation of units	Statistical adjustment for measured exposure to other treatments
Full treatment adherence assumption	Incentives for adherence	Instrumental variable analysis (exclusion restriction is assumed)
No attrition assumption (measurement of all randomized participants on outcome measure)	Sample retention procedures	Modern missing data analysis (outcome measure assumed to be missing at random)

Regression discontinuity design		
Functional form of the relationship between assignment variable and outcome is properly modeled	Replication with different cutpoint; nonequivalent dependent variable	Sensitivity analysis; nonparametric regression
Interrupted time series design		
Another historical event, a change in population (selection), or change in measures coincides with the introduction of the intervention; functional form of the relationship for the time series is properly specified	Nonequivalent dependent measure; nonequivalent control series in which intervention is not introduced; switching replication in which intervention is introduced at another time point in another group	Sensitivity analysis; diagnostic plots (autocorrelogram; spectral density)
Observational Study		
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

/tfn/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 randomized experiment 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 of the commonly used approaches and are not exhaustive. For the observational study design, Rubin's (2005; Imbens & Rubin, in press) and Campbell's (Shadish, Cook, & Campbell, 2002) perspectives differ so that the statistical and design approaches do not map 1:1 onto the assumptions/threats to internal validity that are listed.