Alternatives to the Randomized Controlled Trial

Stephen G. West, PhD, Naihua Duan, PhD, Willo Pequegnat, PhD, Paul Gaist, PhD, MPH, Don C. Des Jarlais, PhD, David Holtgrave, PhD, José Szapocznik, PhD, Martin Fishbein, PhD, Bruce Rapkin, PhD, Michael Clatts, PhD, and Patricia Dolan Mullen, DrPH

Public health researchers are addressing new research questions (e.g., effects of environmental tobacco smoke, Hurricane Katrina) for which the randomized controlled trial (RCT) may not be a feasible option.

Drawing on the potential outcomes framework (Rubin Causal Model) and Campbellian perspectives, we consider alternative research designs that permit relatively strong causal inferences. In randomized encouragement designs, participants are randomly invited to participate in one of the treatment conditions, but are allowed to decide whether to receive treatment.

In quantitative assignment designs, treatment is assigned on the basis of a quantitative measure (e.g., need, merit, risk). In observational studies, treatment assignment is unknown and presumed to be nonrandom. Major threats to the validity of each design and statistical strategies for mitigating those threats are presented. (*Am J Public Health.* 2008;98:1359–1366. doi:10.2105/AJPH.2007.124446)

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

—John Tukey[1]

THE RANDOMIZED CONTROLLED trial (RCT) has long been the gold standard for clinical research, representing the best way to determine efficacy and effectiveness for many intervention and prevention programs. However, public health researchers are increasingly addressing questions for which the RCT may not be a practical (or ethical) option or for which the RCT can be complemented by alternative designs that enhance generalization to participants and contexts of interest.

When structural or policy interventions are being examined, practical problems in conducting RCTs may arise[2]—for example, research participants may not want to be randomized, randomization may not be feasible or not accepted in the research context, or only atypical participants may be willing to be randomized. Such problems might be a concern in studies of the effects of environmental tobacco smoke on nonsmokers or of the effects of the severe disruption of Gulf Coast communities by Hurricane Katrina on HIV risk behaviors and medical care. Only atypical participants may agree to participate in the evaluation of a faith-based intervention. Highly religious participants may refuse to be assigned to a non–faith-based treatment group, whereas nonreligious participants may refuse or be unable to participate sincerely in a faith-based group. Randomization may be precluded if the religious organization implementing the intervention strongly believes that all people desiring the faith-based intervention should receive it. With one exception, our focus is on designs in which participants are assigned to treatment or control conditions. Parallel designs exist in which settings, time, or even dependent measures are the unit of assignment.[3]

THE RANDOMIZED CONTROLLED TRIAL

The RCT has its origins in the pioneering work of the English scientist and statistician Sir Ronald Fisher[4] in agriculture during the 1920s and 1930s. Fisher's key insight was that random assignment of units to treatment conditions leads to 2 related expectations: (1) the mean level for each of the treatment conditions is equal, on average, on any conceivable participant background variable prior to the beginning of the experiment; and (2) treatment assignment is, on average, unrelated to any conceivable participant background variable.

In the context of Fisher's agricultural studies, these expectations

guaranteed that the design would provide an unbiased estimate of the true causal effect. However, other features of the public health context require additional assumptions when traditional RCTs are utilized.[5] Unlike the corn plants in Fisher's agricultural studies, people can seek out alternative treatments or refuse treatments (nonadherence to treatment). People can refuse to be measured or migrate to another locale (attrition). Important advances addressing the challenges of nonadherence, attrition, and their combination have been made during the last half century. Advances in alternative designs and statistical analyses have also occurred.[6–11] Two perspectives have guided this work.

TWO PERSPECTIVES ON STRENGTHENING CAUSAL INFERENCE

Potential Outcomes Perspective

The potential outcomes perspective was originally introduced by Neyman[12] and developed by Rubin et al.[5,13,14] It takes as its starting point a comparison of an individual unit's outcome when the treatment is applied, $Y_t(u)$, versus the same unit's outcome when the alternative (or control) treatment is applied, $Y_c(u)$. The causal effect is defined as

$$(1) \qquad Y_t(u) - Y_c(u),$$

where $Y_t(u)$ represents the response of unit u to treatment t, and $Y_c(u)$ represents the response of the same unit u to the control treatment c at the identical time and in the identical setting. Theoretically, comparison of these 2 outcomes provides the ideal design for causal inference. Unfortunately, this ideal can never be achieved in practice. Additional assumptions are required depending on the choice of alternative to the ideal design. For the RCT, the additional assumptions required are (1) the units are independent, (2) participants actually received the treatment as intended (e.g., complete treatment adherence), (3) attrition from posttest measurement did not occur, and (4) the existence of other treatment conditions did not affect the participant's outcome.[5] If these assumptions of the RCT are met, strong inferences can be drawn about the average causal effect of treatment t relative to treatment c on the outcome. However, these assumptions are often not met. For example, in RCTs of mammography screening, one third of participants in the treatment group have refused screening and many participants in the control group have obtained screening outside the trial.[15]

Campbellian Perspective

Campbell et al. have developed a practical theory of causal inference that follows the logic and strategies of the working scientist.[16,17] Researchers need to identify plausible threats to the validity of the causal inference based on design considerations and prior empirical research. Then they need to rule out the possibility that any of those threats are responsible for the observed effect. If the initially proposed design does not rule out important plausible threats to causal inference, enhancements to the design are introduced that address the identified threats. Through a process of continued critical evaluation and additional research, plausible threats to validity can be identified and eliminated, yielding improved estimates of the causal effect.

Although Campbell et al. discussed 4 types of threats to validity, space limitations restrict our discussion to 2 types. Threats to internal validity are confounding factors that may potentially produce the observed results. These threats include factors that may lead to changes between baseline and posttest (e.g., differential history, maturation) and factors that may lead to differences between the treatment and control groups (e.g., differential selection, differential attrition) in the absence of a treatment effect. Threats to external validity limit the potential generalization of the results, an important consideration given the increasing emphasis on the translation of research results in public health into practice.

ALTERNATIVE DESIGNS FOR STRENGTHENING CAUSAL INFERENCE

Randomized Encouragement Designs

Trial participants are expected to adhere to their treatment assignments in classic RCTs. They may be given strong incentives that are outside usual practice to ensure adherence with the full protocol. Alternatively, participants may be randomly assigned to an opportunity or an encouragement to receive a specific treatment, but allowed to choose whether to receive the treatment. This variation from the classic RCT model is useful for interventions for which it is impractical or unethical to require adherence or in which the necessary incentives would be unrealistic, thus precluding generalization to practice.

For example, this design was used by Vinokur et al.[18] to study the impact of a job seeking skills program (JOBS) on depression in participants. This study recruited eligible participants (e.g., laid off and seeking a new job) at unemployment offices. All participants received a brief booklet describing job search methods. Participants were randomly assigned (stratified by baseline risk) to receive or not receive an invitation to participate in the JOBS program, a 20-hour group training program that emphasized learning and practicing job seeking skills, inoculation against setbacks, and social support. Of invited participants, 54% attended the program. Attempts were made to measure all participants on depression 6 months after baseline measurement (87% response rate).

Intention to treat analyses can be applied to randomized encouragement designs to assess the impact of treatment assignment (the offer of or encouragement to participate in the program) on participant outcome (depression). To the extent that missing data are negligible, the estimated effects are unbiased. Under the assumption of the exclusion restriction (the impact of treatment assignment is mediated entirely through the receipt of treatment), instrumental variables analysis[6] provides an unbiased estimate of the more informative complier average causal effect—the effect of the receipt of treatment (JOBS attendance) averaged across adherers who are expected to adopt the treatment if assigned to the treatment group. Little and Yau[8] compared the subgroup of participants who adhered to treatment in the JOBS program with the subgroup of individuals in the control group who would be expected to adhere to the treatment if invited to participate in the JOBS program. The JOBS program led to decreased depression

for high-risk participants who would adhere to treatment. The combination of randomization and the assumption of the exclusion restriction provided a strong basis for the unbiased estimate of the average effect of the JOBS program and proper standard errors for treatment adherers in a community population. More-complete discussions of randomized encouragement designs are available.[19,20]

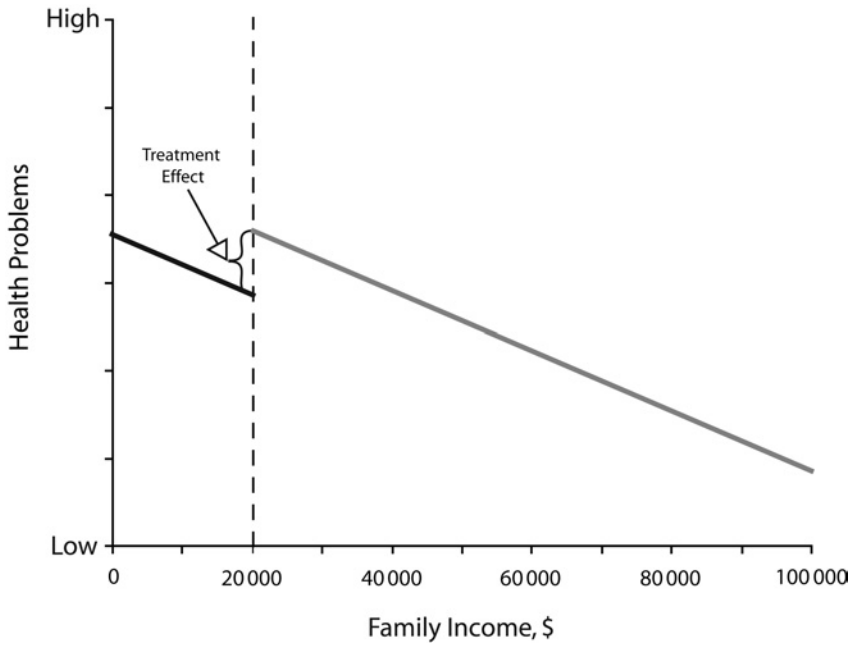
Nonrandom Quantitative Assignment of Treatment

In quantitative assignment designs, participants are assigned to treatment groups on the basis of a quantitative measure, often a measure of need, merit, or risk.[17,21–24] For example, school lunch programs in the United States are assigned to children whose household income falls below a prespecified threshold related to need (e.g., poverty line). Causal inference is based on modeling the functional relationship between the known quantitative assignment variable (household income) and the outcome variable (e.g., health, school achievement), estimated separately for the treated group that falls below the threshold and the control group that falls above the threshold. Because the assignment variable fully determines treatment assignment, proper adjustment for the assignment variable permits the inference of a treatment effect for the school lunch program if there is a discontinuity at the threshold where the treatment is introduced (Figure 1).

As part of the launch of the Head Start program in 1965, US counties with a poverty rate above 59.2% (the 300 poorest in the country) received technical assistance in writing Head Start proposals. A very high proportion (80%) of the poorest counties received Head Start funding, approximately double the funding rate of counties that were slightly better off economically (49.2%–59.2% poverty rates) that did not receive technical assistance. The original Head Start program provided basic health services (e.g., nutrition, immunization, screening) to children in addition to its educational component. Using a regression discontinuity design, Ludwig and Miller[25] found results that demonstrated the introduction of Head Start had led to substantially lower mortality rates in children aged 5 to 9 years from diseases addressed by the program (e.g., measles, anemia, diabetes).

Quantitative assignment designs can be applied to units at various levels such as individuals, community health clinics, neighborhoods, or counties. These designs offer an important alternative to classic RCTs in situations in which randomization is impractical or unethical. Many important public health communities might be resistant to RCTs, in which case quantitative assignment designs might be more acceptable. In addition, the RCT might be unethical when there are doubts about equipoise. Quantitative assignment designs that utilize a clinically meaningful assignment variable (e.g., risk level) might provide a stronger ethical basis for such studies. Quantitative assignment designs can also be implemented based on time (interrupted time series) or settings (e.g., measured risk of neighborhoods).[3]

For example, Khuder et al.[26] analyzed 6 years of monthly data on hospital admissions in 2 cities for coronary heart disease and for other diseases unrelated to smoking. In one city, a public ban on indoor smoking was implemented after the third year of data collection. Khuder et al. showed that hospital admissions for coronary heart disease (but not other diseases) declined following the introduction of the smoking ban. By contrast, no change in hospital admissions for either coronary heart disease or other diseases was observed in the comparison city, which did not institute a smoking ban. Any alternative explanation of these results must clearly account for why the change occurred at the point at which the smoking ban was introduced only in the treatment city and only on the outcome related to smoking.[17]



Note. All children whose family income was below the threshold, here $20 000 (dotted line), received the treatment program (school lunch program); all children whose family income was above the threshold did not receive the program. The difference in level between the regression lines for the program and no program groups at the threshold represents the treatment effect.

FIGURE 1—Illustration of regression discontinuity design.

One primary weakness of quantitative assignment designs is that the functional form of the relationship between the assignment variable and response variable is usually unknown. With unknown functional forms, bias might be introduced if the functional relationship is misspecified (e.g., assuming a linear functional form when the true functional form is curvilinear). In smaller samples, separate nonparametric smooths (e.g., lowess) of the data for the participants who receive the treatment and control conditions provide some information about functional form. In large samples, this bias can be minimized by using nonparametric regression methods to estimate the functional relationship. In the regression discontinuity design, the assignment threshold can sometimes be modified in subsequent studies (e.g., some states have different incomes necessary for treatment receipt) to further strengthen causal inference. In the time series design, cities that introduce the intervention at different time points can be compared.[17,27]

Observational Studies

In observational studies, participants in preexisting or constructed groups receive various treatment conditions, often through voluntary selection.[28,29] The selection of participants into each treatment condition may be associated with confounding factors, resulting in bias that might occur in naive statistical analyses. However, advances in methodology have provided a much stronger toolkit for observational studies. We discuss 2 general approaches below.

First, within the potential outcomes perspective, an important focus has been on the development of matched sampling strategies and analyses.[30,31] Among the most developed strategies are causal inference methods based on propensity scores.[29,32] Propensity scores represent the predicted probability that a participant will receive the treatment given his or her baseline measurements, estimated using either logistic regression to predict treatment status, or more-complex functional forms such as regression tree models.[33] If the researcher can accurately construct propensity scores that balance the treatment and control participants on all potentially relevant baseline variables, the difference between the response in the treatment condition and the control condition (conditioned on the propensity scores) will represent the causal effect. In essence, conditioning on the basis of the propensity scores attempts to create homogeneous units whose responses in the treatment and control groups can be directly compared.

The propensity scores method can only mitigate *overt* selection bias attributed to those baseline characteristics that have been accurately measured.[29] The adequacy of the comparison depends strongly on baseline assessment of the full set of variables believed to be potentially related to treatment selection and outcome. Assessment of a few convenient baseline variables (e.g., demographics) is unlikely to substantially mitigate selection bias.

Haviland et al.[34] studied the effect of gang membership on violent delinquency, an important question for which an RCT was not feasible. They conducted a longitudinal study of boys living in lower socioeconomic areas of Montreal, Quebec, and identified boys who were not members of any gang prior to age 14 years. Based on the boys' behaviors between the ages of 11 and 13 years, they identified groups with a history of low violence, declining violence, and chronic high violence. Within each of these groups, they measured a large number of baseline covariates known to be related to gang membership and violence. Propensity to join a gang at age 14 years was estimated separately within each violence history group from the baseline covariates, with the result that boys who did and did not join gangs at age 14 years could be closely matched within both the low and declining violence groups, but not the chronic high violence group. This finding illustrates that the propensity scores method often *appropriately* limits generalization of the causal effect by restricting comparisons to only the range of propensity scores within which adequate comparisons can be constructed. In the low and declining groups, joining a gang at age 14 years increased violent delinquent acts.

Haviland et al. also performed a sensitivity analysis that investigated how large hidden bias would need to be before the treatment effect was no longer statistically significant. They found that even if hidden variables existed that led to a 50% increase in the odds of joining a gang, a significant treatment effect would still exist. Such causal sensitivity analysis against hidden bias can be used to bracket the plausible range of the magnitude of the causal effect.[29,35] Alternatively, hidden bias caused by unobserved confounding factors can sometimes be mitigated using instrumental variables analysis.[36,37]

Second, within the Campbellian framework, design elements are added that address likely threats to internal validity.[17,38] These design elements include strategies such as matching and stratifying, use of pretests on multiple occasions to estimate preexisting trends, use of multiple control groups with different strengths and weaknesses to bracket the effect, and the use of nonequivalent dependent measures that are expected to be affected by the threat but not by the treatment (see also Rosenbaum[29,39]). Reynolds and West[40] provide an illustration of the use of several of these strategies in an observational study designed to evaluate the effectiveness of a program to increase the sales of state lottery tickets in convenience stores. The store managers refused randomization. Those stores that agreed to implement the program were matched with other stores in the same chain on baseline sales volume and geographical location. Increases in sales were observed in (1) the treatment but not the control group; (2) within the treatment group, for lottery ticket sales, but not other sales categories; and (3) in the weeks following the introduction of the intervention, but not before (Figure 2). Taken together, inclusion of these additional design elements made it extremely difficult to identify any potential confounding factors that might be responsible for the observed pattern of results. In the Campbellian framework, strong priority is given to design enhancements over statistical corrections with their associated assumptions.[38]

CONCLUSION

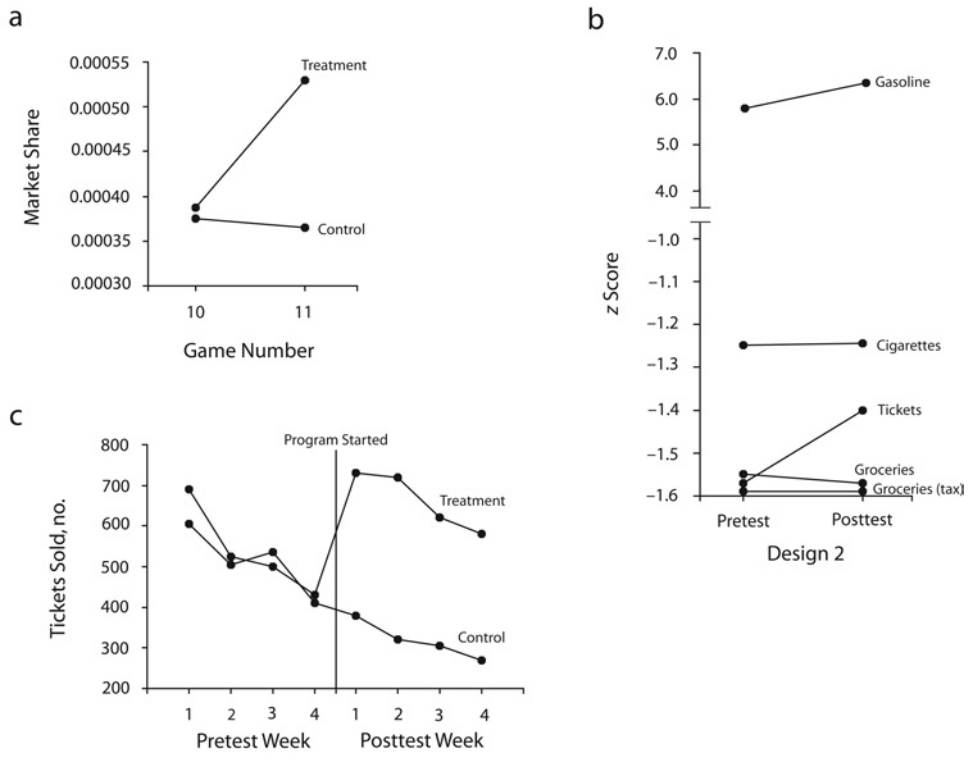
The RCT is the gold standard among research designs. It has

the highest internal validity because it requires the fewest assumptions to attain unbiased estimates of treatment effects. Given identical sample sizes, the RCT also typically surpasses all other designs in terms of its statistical power to detect the predicted effect. Nonetheless, even with the best planning, the RCT is not immune to problems common in community trials. These threats potentially weaken the causal inferences.

When RCTs cannot be implemented in settings or with participants of interest, it is far better to use a strong alternative design than to change the treatment (e.g., using an analog rather than an actual faith-based treatment) or study population (e.g., using only participants indifferent to the treatment choice) so that an RCT may be implemented. Such changes may severely limit the external validity of the findings, potentially distorting the inference about the causal effect for the specific population, treatment, and setting of interest. Even when RCTs can be implemented, alternative designs can be valuable complements that broaden the generalizations of RCTs in multistudy programs of research.

The alternative design and statistical approaches permit relatively strong causal inference in the RCT when common problems such as treatment nonadherence and participant attrition occur and in alternative designs when randomization is not possible. Researchers need to give careful attention to the additional assumptions required by these approaches. Table 1 lists each of the designs considered in this article. The first section lists the basic assumptions and internal validity threats of the RCT, together with design and statistical approaches for addressing these issues. Each subsequent section lists key assumptions and threats to internal validity in addition to those of the RCT, together with design and statistical approaches for addressing these issues.

To illustrate, the key additional threat in the regression discontinuity design is misspecification of the functional form of the relationship between the assignment and outcome variables (typically assumed to be linear). Statistically, nonparametric regression in large samples and sensitivity analyses in small samples that probe the extent of misspecification necessary to undermine the observed treatment effect can help bracket the possible range of the effect size. Adding the design feature of a nonequivalent dependent variable that is expected to be affected by important confounders, but not by the treatment, can help rule out many of the threats to internal validity. In general, the causal effect estimated from the alternative designs and analyses is likely to be associated with more uncertainty than those from the ideal RCT in which no attrition or treatment nonadherence has occurred. Confidence intervals that provide a range of plausible effect sizes caused by sampling fluctuations should be supplemented with estimated brackets on effect sizes that indicate how large or small the effect might plausibly be if key assumptions are not met.[3,17] Remaining



Note. In panel a, treatment and control stores were selected from the same chain, were in the same geographical location, and were comparable in sales during baseline (lottery game 10). Introduction of the treatment at the beginning of lottery game 11 yielded an increase in sales only in the treatment stores. In panel b, within the treatment stores, sales of lottery tickets increased substantially following the introduction of treatment. Sales of other major categories (gasoline, cigarettes, nontaxable groceries, and taxable groceries that would be expected to be affected by confounding factors, but not treatment) did not show appreciable change. In panel c, treatment and control stores' sales show comparable trends in sales during the 4 weeks prior to and 4 weeks following the introduction of the treatment. The level of sales in the treatment and control stores is similar prior to the introduction of treatment but differ substantially beginning immediately after treatment is introduced.

Source. Adapted from Reynolds and West.[40]

FIGURE 2—Design elements that strengthen causal inferences in observational studies: matching (a), nonequivalent dependent variables (b), and repeated pre- and posttest measurement (c).

TABLE 1—Key Assumptions or Threats to Internal Validity and Example Remedies for Randomized Control Trials and Alternatives

Assumption or Threat to Internal Validity	Approaches to Mitigating the Threat to Interval Validity: Design Approach	Approaches to Mitigating the Threat to Interval Validity: Statistical Approach
Randomized controlled experiment		
Independent units	Temporal or geographical isolation of units	Multilevel analysis (other statistical adjustment for clustering)
Full treatment adherence	Incentives for adherence	Instrumental variable analysis (assume exclusion restriction)
No attrition	Sample retention procedures	Missing data analysis (assume data missing at random)
Other treatment conditions do not affect participant's outcome (SUTVA)	Temporal or geographical isolation of treatment groups	Statistical adjustment for measured exposure to other treatments
Randomized encouragement design		
Exclusion restriction	No design approach yet available	Sensitivity analysis
Regression discontinuity design		
Functional form of relationship between assignment variable and outcome is properly specified	Replication with different threshold; nonequivalent dependent variable	Nonparametric regression; sensitivity analysis
Interrupted time series analysis		
Functional form of the relationship for the time series is properly specified; 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
Observational study		
Measured baseline variables equated; unmeasured baseline variables equated; differential maturation; baseline variables reliably measured	Multiple control groups; nonequivalent dependent measures; additional pre- and postintervention measurements	Propensity score analysis; sensitivity analysis; subgroup analysis; correction for measurement error

Note. SUTVA = stable unit treatment value assumption. The list of assumptions and 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 controlled trial 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, the potential outcomes and Campbellian frameworks study differ so that the statistical and design approaches do not map 1-to-1 onto the assumptions or threats to internal validity that are listed. More in-depth descriptions can be found in Shadish et al.[17] and West et al.[22]

uncertainty about the causal effect can often be reduced by adding design features that help rule out the possibility that other unobserved confounders are producing the observed effect.

We have touched only briefly on the matter of external validity. Generalization of findings should not be assumed; features to enhance generalization need to be built into the design.[17] Some RCTs have features that decrease the generalizability of their results to the actual treatments, settings, and populations of interest.[22] This may limit the ability of public health research to provide information about the actual effectiveness of interventions to alleviate health problems. People can have preferences and capacities that interact with treatment effects. Important contextual variables can influence intervention effects as well as participant self-selection and attrition. Regardless of the design chosen, features that maximize external validity should be incorporated into the design. Shadish et al.[17] present procedures for doing this in both single and multiple studies.

Our opening quotation from John Tukey reminds us that the public health significance of the research question should be paramount in the design of research. Important questions should not be ignored if they cannot be fitted into the framework of an RCT. Rather, the strongest possible design that can feasibly be implemented should be chosen, whether an RCT or an alternative design. Whatever design is chosen, careful attention must be given to the viability of the assumptions of the design, adding design and analysis features to address plausible threats to internal and external validity.

In addition, the evaluation of important interventions is rarely limited to single studies but rather is based on the accumulated body of research. The use of systematic reporting frameworks, such as CONSORT[41] for RCTs and TREND[42] for nonrandomized studies, may encourage more in-depth appraisal of research designs both during the planning of the study and the evaluation of its results. Scientific progress in public health will be facilitated by asking the right questions, choosing the strongest feasible design that can answer those questions for the population of interest, and probing the assumptions underlying the design and analysis choices through the addition of carefully chosen design features and supplemental statistical analyses. ■

About the Authors

Stephen G. West is with Arizona State University, Tempe. Naihua Duan is with Columbia University, New York, NY, and New York State Psychiatric Institute, New York. Willo Pequegnat is with the National Institute of Mental Health, Bethesda, MD. Paul Gaist is with the National Institutes of Health, Bethesda. Don C. Des Jarlais is with Beth Israel Medical Center, New York. David Holtgrave is with the Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD. José Szapocznik is with the University of Miami School of Medicine, Miami, FL. Martin Fishbein is with the Annenberg School for Communication, University of Pennsylvania, Philadelphia. Bruce Rapkin is with Memorial Sloan-Kettering Cancer Center, New York. Michael Clatts is with the National Development and Research Institutes, Inc, New York. Patricia Mullen is with the University of Texas School of Public Health, Houston.

Requests for reprints should be sent to Stephen G. West, Psychology Department, Arizona State University, Tempe, AZ 85287-1104 (e-mail: sgwest@asu.edu).

Note. *The views in this article are those of the authors. No official endorsement by the US Department of Health and Human Services or the US National Institutes of Health is intended or should be inferred.*

Contributors

S.G. West participated in the initial workshop and helped develop the outline, wrote the initial draft and subsequent drafts of the article incorporating additions and edits, and wrote the final article. N. Duan participated in the initial workshop, participated in the development of the paper outline, drafted part of the article and reviewed and edited the entire article. W. Pequegnat conceptualized the initial workshop on which the article is based, co-chaired the workshop and guided development of original outline, wrote the introduction for the first draft, provided feedback on multiple drafts, and coordinated continued development of the article. P. Gaist participated in the original workshop, guided development of the original outline, provided significant input and contributions throughout the planning, writing, review, and revision stages of this article. He has served as 1 of the 2 primary coordinators responsible for overseeing each phase that has been required in the development and writing of this article. D.C. Des Jarlais chaired the initial workshop that led to the writing of the article, contributed text to various drafts, edited and approved the final draft. D. Holtgrave, J. Szapocznik, M. Fishbein, B. Rapkin, M.C. Clatts, and P.D. Mullen attended the workshop, helped conceptualize ideas, contributed text, and reviewed and edited drafts.

Acknowledgments

S.G. West was supported by a study visit grant at the Free University of Berlin by the German Academic Exchange Service.

An earlier version of this article was presented to the meeting of the Committee on the Prevention of Mental Disorders and Substance Abuse among Children, Youth, and Young Adults, Institute of Medicine, Washington, DC, October, 2007.

We thank Wei Wu for her help in the preparation of the figures.

Note. On November 14–15, 2005, the US National Institute of Mental Health and the Office of AIDS Research, US National Institutes of Health, convened a group of experts to consider the critical questions associated with the efficacy and effectiveness of interventions for preventing HIV and other chronic diseases that do not lend themselves to randomized controlled trials. This discussion led to the development of this article.

Human Participant Protection

No protocol approval was needed for this study.

References

1. Tukey JW. The future of data analysis. *Ann Math Stat.* 1962;33:13–14.

2. Bonnell C, Hargreaves J, Strange V, Pronyk P, Porter J. Should structural interventions be evaluated using RCTs? The case of HIV prevention. *Soc Sci Med.* 2006;63:1135–1142.

3. Reichardt CS. The principle of parallelism in the design of studies to estimate treatment effects. *Psychol Methods.* 2006;11:1–18.

4. Fisher RA. *The Design of Experiments.* Edinburgh, Scotland: Oliver & Boyd; 1935.

5. Holland PW. Statistics and causal inference (with discussion). *J Am Stat Assoc.* 1986;81:945–970.

6. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). *J Am Stat Assoc.* 1996;91:444–472.

7. Jo B. Statistical power in randomized intervention studies with noncompliance. *Psychol Methods.* 2002;7:178–193.

8. Little RJ, Yau L. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychol Methods.* 1998;3:147–159.

9. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* 2nd ed. New York, NY: John Wiley and Sons; 2002.

10. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods.* 2002;7:147–177.

11. West SG, Sagarin BJ. Participant selection and loss in randomized experiments. In: Bickman L, ed. *Research Design: Donald Campbell's Legacy.* Vol. 2. Thousand Oaks, CA: Sage Publications, 2000;117–154.

12. Neyman J. On the application of probability theory to agriculture experiments. Essay on principles. Section 9. *Statistical Science.* 1990;5:465–472. Originally published in *Roczniki Nauk Rolniczych* [Annals of Agricultural Science] 1923, Tom X, 1–51. Translated and edited by DM Dabrowska and TP Speed.

13. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol.* 1974;66:688–701.

14. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc.* 2005;100:322–331.

15. Baker SG. Analysis of survival data from a randomized trial with all-or-none compliance: estimating the cost-effectiveness of a cancer screening program. *J Am Stat Assoc.* 1998;93:929–934.

16. Campbell, DT. Factors relevant to the validity of experiments in social settings. *Psychol Bull.* 1957;54: 297–312.

17. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton-Mifflin; 2002.

18. Vinokur AD, Price RH, Schul Y. Impact of the JOBS intervention on unemployed workers varying in risk for depression. *Am J Community Psychol.* 1995;23:39–74.

19. Holland PW. Causal inference, path analysis, and recursive structural equation models (with discussion). In: Clogg C, ed. *Sociological Methodology 1988.* Washington, DC: American Sociological Association; 1988:449–493.

20. Barnard J, Frangakis CE, Hill JL, Rubin DB. A principal stratification approach to broken randomized experiments: a case study of school choice vouchers in New York City (with discussion). *J Am Stat Assoc.* 2003;98:299–323.

21. Mark MM, Reichardt CS. Quasi-experimental and correlational designs: methods for the real world when random assignment isn't feasible. In: Sansone C, Morf CC, Panter AT, eds. *Sage Handbook of Methods in Social Psychology.* Thousand Oaks, CA: Sage Publications, 2003:265–286.

22. West SG, Biesanz JC, Pitts SC. Causal inference and generalization in field settings: experimental and quasi-experimental designs. In: Reis HT, Judd CM, eds. *Handbook of Research Methods in Social and Personality Psychology.* New York, NY: Cambridge University Press; 2000:40–84.

23. Finkelstein MO, Levin B, Robbins H. Clinical and prophylactic trials with assured new treatment for those at greater risk: I. A design proposal. *Am J Public Health.* 1996;86:691–695.

24. Finkelstein MO, Levin B, Robbins H. Clinical and prophylactic trials with assured new treatment for those at greater risk: II. Examples. *Am J Public Health.* 1996;86:696–705.

25. Ludwig J, Miller DL. Does Head Start improve children's life chances? evidence from a regression discontinuity design. *Q J Econ.* 2007;122:159–208.

26. Khuder SA, Milz S, Jordan T, Price J, Silvestri K, Butler P. The impact of a smoking ban on hospital admissions for coronary heart disease. *Prev Med.* 2007;45:33–8.

27. Hawkins NG, Sanson-Fisher RW, Shakeshaft A, D'Este C, Green LW. The multiple baseline design for evaluating population-based research. *Am J Prev Med.* 2007;33:162–168.

28. Cochrane WG. The planning of observational studies of human populations (with discussion). *J R Stat Soc.* Series A (General). 1965;128:236–265.

29. Rosenbaum PR. *Observational Studies.* 2nd ed. New York, NY: Springer; 2002.

30. Rubin DB. *Matched Sampling for Causal Effects.* New York, NY: Cambridge University Press; 2006.

31. West SG, Thoemmes F. Equating groups. In: Alasuutari P, Brannen J, Bickman L, eds. *The SAGE Handbook of Social Research Methods.* London, England: Sage Publications; 2008:414–430.

32. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70:41–55.

33. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods.* 2004;9:403–425.

34. Haviland A, Nagin DS, Rosenbaum PR. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychol Methods.* 2007;12:247–267.

35. Sommer A, Zeger SL. On estimating

efficacy from clinical trials. *Stat Med.* 1991;10:45–52.

36. Winship C, Morgan SL. The estimation of causal effects from observational data. *Annu Rev Sociol.* 1999;25:659–706.

37. Morgan SL, Winship C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* New York, NY: Cambridge University Press; 2007.

38. Shadish WR, Cook TD. Design rules: more steps towards a complete theory of quasi-experimentation. *Stat Sci.* 1999;14:294–300.

39. Rosenbaum PR. Replicating effects and biases. *Am Stat.* 2001;55:223–227.

40. Reynolds KD, West SG. A multiplist strategy for strengthening nonequivalent control group designs. *Eval Rev.* 1987;11:691–714.

41. Moher M, Schulz KF, Altman D, the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA.* 2001;285:1987–1991.

42. Des Jarlais DC, Lyles C, Crepaz N, the TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health evaluations: the TREND statement. *Am J Public Health.* 2004;94:361–366.