Improved Designs for Cluster Randomized Trials

Catherine M. Crespi

Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, California 90095-1772; email: ccrespi@ucla.edu

Annu. Rev. Public Health 2016. 37:1–16

First published online as a Review in Advance on January 18, 2016

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

This article's doi: 10.1146/annurev-publhealth-032315-021702

Keywords

blocking, crossover, factorial, group randomized, sample size, power

Abstract

Studies in which clusters of individuals are randomized to conditions are increasingly common in public health research. However, the designs utilized for such studies are often suboptimal and inefficient. We review strategies to improve the design of cluster randomized trials. We discuss both older but effective design concepts that are underutilized, such as stratification and factorial designs, as well as emergent ideas including fractional factorial designs and cluster randomized crossover studies. We draw examples from the recent literature and provide resources for sample size and power planning. Given the inherent inefficiencies of cluster randomized trials, these design strategies merit wider consideration and can lead to studies that are more cost-effective and potentially more rigorous than traditional approaches.

INTRODUCTION

In an individually randomized trial (IRT), individuals are randomized to treatment conditions and the intervention is applied at the individual level. In a group or cluster randomized trial (CRT), naturally occurring groups or clusters of individuals are the experimental units randomized to conditions and the units to which the intervention is applied. The experimental units in CRTs have included families, schools, work sites, churches, health care centers, and entire communities.

Since public health investigators first conducted cluster randomized trials in the 1970s (6, 26), the CRT has become a widely used design. CRTs are well suited to public health research. Researchers have increasingly recognized that in order to change complex behaviors such as tobacco use, physical inactivity, and unhealthy diet, which are associated with chronic diseases, it is not sufficient to focus solely on interventions applied to the individual; rather, behavior change also requires systems-level interventions focused on the societal or environmental factors that influence these behaviors (47). CRTs are a natural design for the evaluation of health system and policy interventions and other interventions applied to organizations of individuals (28, 47). A relatively new intervention-delivery strategy is to target social networks, on the theory that doing so will accelerate the spread of information and behaviors within a community and will maximize behavior change at the community level (34). CRTs are a natural design for testing the effectiveness of such delivery strategies.

Other factors can motivate the choice of clusters as the units of experimentation (6, 23, 26, 31). These include the desire to avoid contamination, that is, exposure of individuals randomized to one condition to the comparison condition. Other considerations are the logistical convenience and cost containment that accrue when an intervention is applied at a group level, as well as a desire to avoid the acceptability and ethical issues that can arise when individuals in the same group are treated differently.

Literature reviews in the recent past have found widespread deficiencies in the design, analysis, and reporting of CRTs, such as analyzing them as if they were IRTs, which typically leads to underestimated standard errors and increased type I error rates (7, 18, 50). Such gross errors have diminished. However, when resources are limited, it is not enough simply to avoid overt errors; rather, we should strive to design studies to be as efficient as possible. Thus it behooves investigators to understand all strategies available for improving a CRT and to use design features that will achieve study objectives as cost-effectively and efficiently as possible.

To that end, this article reviews ideas for CRT design that may aid public health investigators who seek improved, cost-effective, and rigorous studies. We discuss new design options that merit serious consideration because of their economy and versatility but are not yet widely used. We also discuss older but effective design strategies for CRTs. Classic statistical design, discussed in texts such as that of Casella (8), offers many ideas for improving statistical efficiency. However, these principles tend to be overlooked when designing CRTs. Our main focus is on strategies that increase efficiency, that is, increase the precision of the intervention effect estimator. A more efficient design achieves the same power with fewer resources.

This review is organized as follows. First, we discuss the most basic cluster randomized trial design, the completed randomized parallel group CRT, which serves as a reference point. We then discuss some classic methods of improving efficiency. These include blocking and factorial designs. Factorial designs are especially useful for evaluating multicomponent interventions. We then discuss some emergent design ideas, including fractional factorial designs and cluster randomized crossover trials. We provide examples and sample size formulas and discuss resources for conducting sample size and power calculation.

For readers seeking general resources on the design and analysis of CRTs, a number of texts are available, including those of Murray (41), Donner & Klar (23), Hayes & Moulton (31), Eldridge & Kerry (26), Campbell & Walters (6), and Moerbeek & Teerenstra (39).

COMPLETELY RANDOMIZED PARALLEL GROUP DESIGN

The completely randomized parallel group design is the best understood CRT design. In this design, clusters are allocated to intervention and control conditions using simple randomization and remain in their allocated condition throughout the trial. This design is the analog of the completely randomized IRT, except that clusters rather than individuals are randomized.

A key feature of CRTs is that they yield estimates of intervention effects that have higher variance (higher standard error) than an IRT with the same number of individuals. The higher variance arises from the cluster-based sampling of observations. This was first recognized by Cornfield (17) and is explained in several texts (see, for example, 6, 23, 26). Suppose that we randomly and independently sample n observations, e.g., children, from a population and compute the sample mean of an outcome. The variance of this sample mean is σ^2/n, where σ^2 is the variance of a single observation, and its standard error is $\sigma/\sqrt{n}$. Now suppose that the observations are naturally aggregated into clusters, e.g., children aggregated into schools, and we sample clusters. In this case, we can apportion the total variance of an observation into two components, $\sigma^2 = \sigma_B^2 + \sigma_W^2$, where σ_W^2 represents the variance of observations within a cluster and σ_B^2 represents the variance of cluster-level means. Note that $\sigma_W^2 \leq \sigma^2$, reflecting that observations within the same cluster tend to vary less than randomly selected observations from the overall population (23). The extent of clustering will vary from setting to setting. We often quantify the magnitude of clustering using the intraclass correlation coefficient (ICC), defined as the proportion of the total variance in the outcome that is attributable to variance between clusters,

$$\rho = \frac{\sigma_B^2}{\sigma^2} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}.$$

If we sample c clusters with m observations in each cluster for a total of $n = mc$ observations, we can show that the sample mean will have variance $\sigma^2\left[1 + (m-1)\rho\right]/n = (\sigma_W^2 + m\sigma_B^2)/n$ (40). Because ρ is generally positive, $1 + (m-1)\rho$ will usually be greater than 1, and the variance will be inflated compared with independent sampling. The factor $1 + (m-1)\rho$ is the well-known design effect or variance inflation factor for a CRT. Thus, a cluster-based sample will yield an estimate of the intervention effect that has a larger standard error than yielded by independent sampling, all else being equal, and more subjects will be required to achieve the same power.

Table 1 provides the total number of subjects required to achieve the desired power for a comparison of means in a two-arm trial for various trial designs. The sample size formulas are expressed in two different but equivalent forms, a variance components version and a correlation coefficient version. Comparing the correlation coefficient versions for completely randomized IRT and CRT designs, we see that the required sample size is inflated by the design effect, $1 + (m-1)\rho$, for the CRT. Similarly, comparing the variance components versions, we note that $\sigma_W^2 + m\sigma_B^2 > \sigma^2 = \sigma_W^2 + \sigma_B^2$, which also reflects variance inflation due to clustering.

The reader should note that in the formulas in **Table 1**, in order to focus on important concepts, we have made simplifying assumptions such as equal variances, equal allocation, and constant cluster size, which may not hold for all studies. Researchers should consult references such as the texts cited in this article when designing a real trial. We illustrate the impact of the design effect with an example.

Table 1 Total number of subjects required to achieve desired power for comparing two means for various trial designs[a,b]

Design	Variance components form	Correlation coefficient form
Completely randomized IRT	$\frac{4(z_{\alpha/2}+z_\beta)^2\sigma^2}{(\mu_0-\mu_1)^2}$	$\frac{4(z_{\alpha/2}+z_\beta)^2\sigma^2}{(\mu_0-\mu_1)^2}$
Completely randomized CRT	$\frac{4(z_{\alpha/2}+z_\beta)^2(\sigma_W^2+m\sigma_B^2)}{(\mu_0-\mu_1)^2}$	$\frac{4(z_{\alpha/2}+z_\beta)^2\sigma^2[1+(m-1)\rho]}{(\mu_0-\mu_1)^2}$
Matched or stratified CRT	$\frac{4(z_{\alpha/2}+z_\beta)^2(\sigma_W^2+m\sigma_{BM}^2)}{(\mu_0-\mu_1)^2}$	$\frac{4(z_{\alpha/2}+z_\beta)^2\sigma^2[1+(m-1)\rho-m\rho_M\rho]}{(\mu_0-\mu_1)^2}$
Crossover IRT	$\frac{2(z_{\alpha/2}+z_\beta)^2\sigma_e^2}{(\mu_0-\mu_1)^2}$	$\frac{2(z_{\alpha/2}+z_\beta)^2\sigma^2(1-\rho_e)}{(\mu_0-\mu_1)^2}$
Crossover CRT, crossover at cluster level	$\frac{4(z_{\alpha/2}+z_\beta)^2[\sigma_W^2(1-m\rho_C)+m\sigma_B^2(1-\rho_C)]}{(\mu_0-\mu_1)^2}$	$\frac{4(z_{\alpha/2}+z_\beta)^2\sigma^2[1+(m-1)\rho-m\rho_C]}{(\mu_0-\mu_1)^2}$
Crossover CRT, crossover at individual level	$\frac{2(z_{\alpha/2}+z_\beta)^2\sigma_e^2}{(\mu_0-\mu_1)^2}$	$\frac{2(z_{\alpha/2}+z_\beta)^2\sigma^2(1-\rho_e)}{(\mu_0-\mu_1)^2}$
Factorial CRT	For factors with two levels, use formula for completely randomized CRT, possibly with adjustment of α for multiple comparisons.	

Abbreviations: IRT, individually randomized trial; CRT, cluster randomized trial.

[a]Formulas assume equal variances and correlations in the two study arms, equal numbers of clusters in each arm, constant cluster size, a two-sided test, and a sufficient number of clusters to allow a normal approximation for the test statistic. Several authors recommend adding 1–2 clusters per arm to offset loss of degrees of freedom from use of the *t* distribution when the number of clusters is small; see, e.g., Hayes & Moulton (31), Donner & Klar (23), or Eldridge & Kerry (26).

[b]Notation: z_γ, γ th quantile of the standard normal distribution; α, desired type I error rate; β, desired type II error rate; power, $1-\beta$; m, cluster size; μ_0, μ_1, true means in the control and intervention arms, respectively; σ^2, marginal (total, unconditional) variance of an observation in the population; σ_W^2, variance of observations within the same cluster; σ_B^2, variance of cluster-level means; σ_{BM}^2, variance of cluster-level means between clusters in same pair or stratum; σ_e^2, variance of measurements on the same individual; ρ, intraclass correlation, correlation between two observations in same cluster, equal to $\sigma_B^2/(\sigma_W^2+\sigma_B^2)$; ρ_M, correlation between cluster-level means of matched pairs of clusters, equal to $\sigma_{BM}^2/(\sigma_W^2+\sigma_{BM}^2)$; ρ_C, correlation between two observations in same cluster in different periods; ρ_e, correlation between two measurements on the same individual.

Example of Completely Randomized Parallel Group CRT: Tai Chi for Stroke Patients

The protocol for a CRT in China evaluating Tai Chi Yunshou exercises for community-based stroke patients calls for randomly allocating community health centers to a Tai Chi exercise group or a balance rehabilitation training group (54). A cluster-randomized design was selected over individual randomization to avoid contamination and for logistical reasons. The study will recruit 25 participants from each center, and the ICC is projected to be 0.05. Thus the variance inflation factor is $1+(25-1)0.05=2.2$, meaning that owing to the cluster randomized design, the study requires 2.2 times as many participants as would be required for an IRT, all else being equal.

To improve the situation, we might consider increasing the sample size of each cluster, m. However, even though increasing m increases the overall sample size, it can also increase the variance inflation factor, so this strategy is typically of limited help. Rather, the key to increasing the power and precision of a CRT lies in reducing the impact of between-cluster variability.

DESIGN STRATEGIES TO IMPROVE EFFICIENCY

A completely randomized parallel group CRT design does nothing to reduce the inefficiency attributable to between-cluster variation. We now discuss several strategies to improve efficiency for CRTs.

Blocked Designs

Blocking, in the form of matching or stratification, is often used in CRTs. However, its advantages are often overlooked, and it is not always utilized when it could prove helpful.

In a matched or stratified CRT design, clusters that are similar on one or more prognostic characteristics are first grouped together, then clusters within groupings are randomized to conditions (23, 31, 41). In a matched-pair design, clusters are formed into pairs and one cluster within each pair is assigned to each condition. In stratified designs, clusters are grouped into larger strata; the number in each stratum exceeds the number of conditions, and multiple clusters within each stratum are assigned to each condition (31). Matching or stratification on one or more characteristics prior to randomization ensures balance between arms on these variables (23, 31). Thus these are useful strategies for achieving balance on baseline prognostic indicators when the number of clusters is relatively small and simple randomization cannot be relied on to achieve balance.

The advantages of matching and stratification in terms of improved power and precision arise when these designs are coupled with a matched or stratified analysis. In such an analysis, comparisons between conditions are made within matched pairs or within strata. If the matching is very close, these comparisons will be similar to comparing the same experimental units under two different conditions. This reduces between-unit (between-cluster) variability in the estimation of the intervention effect, reducing the standard error and increasing power and precision.

The sample size formulas in **Table 1** illuminate the increase in power. Compared with a completely randomized design, in a blocked design, the variance of cluster means over all clusters, σ_B^2, is replaced by the variance of cluster means within pairs or strata, σ_{BM}^2 (31). If the matching is good, we expect to have $\sigma_{BM}^2 < \sigma_B^2$, and the sample size requirement is reduced. Similarly, the correlation coefficient version of the formula shows that the design effect is decreased by an amount $m\rho_M\rho$, where ρ_M is the matching correlation, defined as the Pearson correlation for the outcome computed over blocks of clusters (22, 23) and equal to $1 - \sigma_{BM}^2/\sigma_B^2$ (26).

Variants of this formula exist. Some authors use the approximation that the reduction in the design effect for a matched-pair design compared with a completely randomized design is $1 - \rho_M$ (22, 23). Thus, if the matching correlation is 0.25, the design effect will be reduced by 25%, all else being equal.

There is an important caveat with matched designs, however. The degrees of freedom (df) of the test statistics for unmatched and matched analyses are different. For unmatched analysis, with c clusters in each condition, a t-test statistic comparing the two conditions will have $2(c - 1)$ df. For a matched analysis, with c matched cluster pairs, the t statistic will have $c - 1$ df, which is half the value. A lower df for a t-distribution corresponds with a larger critical value and hence lower power. The difference can be important when the number of clusters is small (23, 31).

Thus matching will yield an efficiency gain only if ρ_M is sufficiently high to offset the reduction in df. For this reason, it is worthwhile to match or stratify only on cluster-level factors known to be strongly associated with the outcome (23). Break-even values of ρ_M are provided in Hayes & Moulton (31).

A further disadvantage of the matched-pair design is that loss to follow-up of a cluster implies that both clusters in the pair must be dropped from a paired outcome analysis. Alternatively, an unpaired analysis could be conducted, but doing so would lose the efficiency gains of matching. This problem is less likely to occur in a stratified design, in which each stratum has multiple clusters in each condition, better ensuring that within-stratum comparisons can be made.

Stratified designs and matched-pair designs share similar advantages in that the intervention effect is estimated by comparing intervention and control clusters within strata, in which cluster characteristics are similar, thus reducing the amount of between-cluster variability that enters into

the comparison and ultimately the standard error. In practice, stratification into a few strata often achieves almost as much reduction in between-cluster variance as does pair-matching, with fewer df lost (31).

Example of matched-pair CRT design: the ACTIVITAL study. ACTIVITAL, a school-based intervention conducted in urban Ecuador, aimed to improve diet and physical activity among adolescents (1). A lifestyle intervention program tailored to the local context was developed and implemented at each school. Randomization at the school level was utilized because intervention development and implementation were school-based.

A matched-pair design was selected to ensure balance on four baseline characteristics: total number of students; monthly school fee, a proxy for socioeconomic level; gender composition; and time schedule of classes. In general, pair-matching on multiple criteria can be challenging. However, the pool of eligible schools was large. Forty-nine schools were available, but sample size calculations indicated that only 10 pairs (20 schools) were needed for adequate power. A total of 28 schools were paired, and 10 pairs were randomly selected.

Schools within pairs were randomly allocated to intervention and wait-list control conditions. Within each school, four classes participated; measurements were collected at baseline and 28 months. The outcome analysis used linear mixed models with a random effect for each pair.

In general, pair-matching will be more efficient than stratification only if pairs can be very closely matched on prognostic factors. In the ACTIVITAL study, there was a relatively large pool of schools and detailed information was available on each school, which facilitated the formation of well-matched pairs. In other studies, there is typically minimal information about the units available in advance, which makes pair-matching less viable and stratification a better option.

Example of stratified CRT design: the Korean health study. The Korean Health Study was a CRT to evaluate an intervention to promote hepatitis B virus testing among Koreans in Los Angeles County (3). The units of randomization were Korean churches. Prior to recruitment, a sampling frame of Korean churches in the county was developed, which included basic information such as geographic location and membership size. These two factors were used as strata. Fifty-two churches, stratified by size (small, medium, large) and location (Koreatown versus other), were randomized to intervention or control conditions. Intervention church participants attended a small-group session on liver cancer and hepatitis B testing, and control church participants attended a similar session on physical activity and nutrition. The outcome was self-reported receipt of a hepatitis B serological test.

Recruitment of churches into the trial spanned several years, during which the investigators cycled through the six strata, sampling two churches within each stratum then randomly allocating them to intervention and control. Although these pairs were not matched on additional church characteristics, they were matched on timing of recruitment, which helped to control for secular trends. The pairing was also a logistical convenience.

Several benefits accrued from the stratified design. The investigators ensured balance between conditions on the stratification factors. The replicates of intervention and control units within each stratum made it possible to obtain stratum-specific estimates of the intervention effect, and these stratified analyses had internal validity because they were essentially randomized trials nested within the larger study.

The pairing within stratum also proved beneficial. During the study, several churches in the large Koreatown stratum were contaminated by free hepatitis B screening events conducted by an outside organization. Sensitivity analyses were conducted by repeating the intervention effect estimation dropping all members of any affected church pair, a strategy that preserved the balance created by randomization. The results confirmed an overall intervention effect as well as

intervention effects within all strata. Thus pairing within a stratum can be considered a strategy to mitigate contamination. Cluster dropout could be handled in the same manner.

Factorial Designs

Many interventions involve multiple components, and their effectiveness is tested together as a package. However, investigators are often interested in studying the effectiveness of the individual intervention components. Factorial designs can be an efficient method to accomplish this objective.

Factorial experimental designs have a history that reaches back to the nineteenth century (61) and discussion of factorial designs can be found in many texts (4, 33, 35, 49). However, factorial designs have not been widely used in public health research or as part of a CRT (41). Articles explaining and promoting the use of factorial designs for intervention research (9, 13) and within CRTs (31, 41) have recently begun to appear in the literature.

Factorial designs involve two or more independent variables or "factors" varied within a single study. Each factor has two or more levels, e.g., present/absent. In a full factorial design, all factors are crossed with each other, such that all possible combinations of factor levels are represented; for example, if one factor has two levels and a second factor has three levels, all $2 \times 3 = 6$ possible combinations are evaluated. In a fractional factorial design, discussed later, only a fraction of the possible combinations are utilized.

The completely randomized CRT can be viewed as a one-factor factorial design. Formulas for one-factor designs can be used to predict power for the main effects of dichotomous factors when there is more than one factor (24). In some studies, adjustments to sample size requirements may be needed owing to df, alpha control for multiple hypothesis testing, or other considerations.

One key advantage of factorial experiments is that they can require much smaller sample sizes than comparable single-factor experiments (13, 49). For example, consider a CRT setting involving clusters of size 20 with an ICC of 0.04. Trial A will test Intervention A versus control. To detect a standardized effect size (difference in means divided by pooled standard deviation) of $d = 0.4$ with 80% power and significance level 0.05, a total of 20 clusters will be required. Trial B will test Intervention B versus control. To detect $d = 0.3$ with 80% power at level 0.05, 34 clusters will be required. Conducted separately, the two trials require 54 clusters. In contrast, a 2×2 factorial design, even with a Bonferroni-adjusted significance level of 0.025 for each comparison, would require a total of only 40 clusters. The smaller sample size requirement occurs because all the clusters are included in the estimation of each main effect.

Factorial designs are also "rich with information" (33, p. 195); they provide information about not only the main effects of each factor, but also their combined or interaction effects. In our example, the factorial design would provide information about the combined effect of Interventions A and B. Interactions can, however, be more difficult to detect than main effects, so larger sample sizes may be required if tests for interaction are important (31).

Factorial designs also have advantages from an equity standpoint, in that a greater proportion of participants receive intervention. Whereas in a two-arm trial with equal allocation, 50% of participants receive intervention, in a 2×2 factorial design, 75% of participants receive at least some intervention. This higher probability of receiving treatment can enhance recruitment (5). Factorial designs can also facilitate mediation analyses examining specific mediators for each intervention component (12). Factorial designs have been promoted as useful for screening studies, in which many candidate intervention components are simultaneously assessed to identify and screen out inactive components (9, 24).

Factorial designs do have some limitations. Because the individual components are likely to have smaller effect sizes than the full multicomponent intervention, a factorial experiment may require a larger sample size than would a two-arm trial comparing a full intervention package with a control

condition (58). However, the two-arm trial would not provide information on the effectiveness of individual components. Factorial designs require close control over the combination of treatments given to each unit, which may be difficult to achieve in community settings (49). The complexity of dealing with interactions is off-putting for some investigators. However, main effects can still be estimated in the presence of interactions in balanced designs (14), and interactions are often a key object of interest for multicomponent interventions and should be a motivator for conducting factorial studies rather than a reason to avoid them (13). Indeed, factorial designs are the only experimental designs that can systematically investigate interactions (9). None of these limitations is specific to CRTs. There are a limited but increasing number of examples of 2×2 factorial CRTs in health research (37). Here we discuss a CRT with a three-factor design.

Example of a full factional CRT design: the HWStudy2. Caldwell et al. (5) report on the design of a translational study implementing a substance use and sexual risk prevention program in schools in South Africa. HealthWise (HW) South Africa is a curriculum of 18 lessons for students in grades 8 and 9 that was previously shown in a CRT with nine schools in the Cape Town area to be efficacious at reducing risky behaviors (51). Given its efficacy, administrators wished to disseminate the program to the remaining 56 schools in the district. Researchers used this opportunity to design a translational study, called HWStudy2, whose objective is to examine factors affecting the quality and fidelity of program implementation and, ultimately, student outcomes.

Because the study outcomes have not been reported yet, we focus solely on the trial design. The investigators were interested in the causal effects of each of three factors: enhanced teacher training; enhanced teacher support, structure, and supervision; and enhanced school environment. Because they were interested in the main effects of each enhancement, a factorial design was selected over an intervention package approach. They hypothesized that each factor would have a main effect on the outcomes, which were measures of implementation fidelity and student behaviors. Each factor will have two levels, present and absent, for a total of $2 \times 2 \times 2 = 8$ conditions. Seven schools will be randomly allocated to each condition, for a total of 56 schools.

This design is not an eight-arm trial in which each condition is compared with the control condition; such a design would have very low power. Rather, in this factorial experiment, the effect of each of the three factors will be estimated using the entire sample of 56 schools. The authors report that in order to achieve the same power as the factorial experiment, conducting individual experiments on each factor would have required three times as many schools (5). Furthermore, in this design, only one out of eight schools received none of the enhancement elements (and even these schools will receive the basic HW program). Other advantages of the factorial design cited by the investigators include the ability to estimate interaction effects and the effects of each of the factors on specific mediators.

Fractional Factorial Designs

Sometimes a full factorial design is not an option because resources are too limited to implement all possible combinations of factors or because some combinations cannot or should not be implemented (14). In these cases, a fractional factorial design could be considered. A fractional factorial design is derived from a full factorial by dropping some conditions (combinations of factors). Although unfamiliar to many public health investigators, fractional factorial designs deserve consideration because of their economy. These designs are discussed in texts on experimental design (8, 35, 59) and in articles aimed at behavioral researchers (14).

The obvious advantage of a fractional factorial design is the reduction in the number of conditions and experimental units utilized. One drawback is that when conditions are removed, certain effects become completely confounded with each other and cannot be estimated separately. This

is termed aliasing. However, researchers can strategically drop conditions so that effects of interest are confounded only with effects that are expected to be minimal, e.g., high-order interactions (59). Hence, it is quite feasible to achieve study objectives with a fractional factorial design. Because fractional factorial designs are widely used in engineering and other fields, software to assist with designs is widely available and includes the FACTEX procedure in SAS, the orthoplan procedure in SPSS, and the R package RfR2.

Investigators may be concerned that the statistical analysis for a cluster-randomized fractional factorial design would be overly complex. However, it is relatively straightforward to analyze data from such studies using a multilevel modeling framework (60). A few examples of the use of fractional factorial designs for health behavior research can be found in Nair et al. (42) and in Strecher et al. (53). These studies used individual randomization. We discuss a rare example of a CRT using a fractional factorial design.

Example of fractional factorial CRT design: the myPlaybook study. Wyrick et al. (60) describe the design of a study to develop and test an online program, myPlaybook, for the prevention of substance use among college student–athletes in the United States. A pilot study of myPlaybook used the classic treatment package approach, in which the entire program—six online lesson modules targeting risk and protective factors for substance use—was delivered as a single package. The results did not provide clear guidance on how the effectiveness could be improved. Thus the researchers decided to use a fractional factorial design to help optimize the multicomponent intervention.

The study by Wyrick et al. used the multiphase optimization strategy, a three-phase framework for the development and evaluation of multicomponent interventions proposed by Collins et al. (12, 15, 16). During the preparation phase, researchers identified intervention components to be studied and performance criteria to be optimized. In the optimization phase, incremental experiments were used to evaluate the components, with revision or deletion of underperforming components. The final evaluation phase involved a randomized trial to evaluate the optimized intervention.

It was the evaluation phase of the myPlaybook study that used a cluster-randomized fractional factorial design. Although the intervention was delivered to individuals in an online format, student–athletes are clustered within schools. Thus, the investigators decided to randomize at the school level to reduce the risk of contamination.

The investigators wanted to evaluate five intervention components corresponding to five specific lesson modules. A full factorial design would have required $2^5 = 32$ conditions. Additionally, the investigators wanted to stratify schools on division (level of college athletics). To have at least 1 school from each of the 3 divisions in each condition, at least 96 schools would have been required. To reduce the sample size, the investigators used a fractional factorial design. Reasoning that high-order interactions would be negligible, they selected a design that combines main effects with four-way interactions and two-way interactions with three-way interactions. Doing so cut the number of conditions in half, to 16. Implementing 16 different conditions was feasible because the program was delivered online, enabling close control over content delivery. The power analysis for the CRT fractional factorial design was carried out using a SAS macro developed by Dziak et al. (24, 38), and the investigators determined that 56 schools with 100 students per school would provide 90% power to detect an effect size of $d = 0.3$.

Cluster-Randomized Crossover Design

The individually randomized crossover design is widely used in clinical research (43, 48). In a 2×2 individually randomized crossover trial, each participant receives two treatments during two

successive time periods, in random order, with a washout period between them. Operationally, individuals are randomized to receive treatments A and B in the order AB or BA.

The main advantage of a crossover IRT is increased statistical efficiency compared with a parallel group trial. Comparisons are made within individuals rather than between individuals, which removes interindividual variability from the treatment effect estimator and thus reduces the standard error (10). As seen in **Table 1**, whereas the key variance component for a completely randomized IRT is the variance between measurements on different individuals, σ^2, the key for a crossover IRT is the variance between measurements on the same individual, σ_e^2, where $\sigma_e^2 = \sigma^2(1 - \rho_e)$ and ρ_e is the correlation between measurements on the same individual (see, for example, 32). Because ρ_e is expected to be positive, $\sigma_e^2 < \sigma^2$. Note that in a 2×2 crossover IRT, the outcome of each individual is measured twice rather than once; hence the formula for total individuals involves a factor of 2 rather than 4. The formula assumes no period or carryover effects.

CRT crossover designs are rare but becoming more popular (2). In a 2×2 CRT crossover design, clusters are randomized to AB or BA conditions. Like the crossover IRT, the crossover CRT has increased statistical efficiency compared with a parallel group CRT. Comparisons are made within clusters rather than between clusters; each cluster serves as its own control, which removes between-cluster variability from the variance of the intervention effect estimator and thus increases precision and power (31). Because the inefficiency of CRTs compared with IRTs is due principally to between-cluster variation, this can represent a large efficiency gain.

A further advantage is that only half the total number of clusters is needed to achieve the same number of person-years of observation compared with a parallel group trial (31). This can be an important consideration when the available number of clusters is limited. In addition, all participants receive all treatments.

In a crossover CRT, each cluster experiences each experimental condition. However, individuals within clusters may not necessarily experience both conditions. Two design subtypes can be distinguished (46). In a CRT with crossover at the cluster level, subjects are included in only one period. This can occur naturally in short-term or acute care settings such as emergency rooms or intensive care units (ICUs), in which individuals accrue and attain their outcome within a single period, or can be implemented by design, by sampling different individuals within a cluster in each period. In a CRT with crossover at the individual level, the same subjects are included in both periods. Such designs may be used in longer-term settings such as schools or residential facilities, where the same individuals are available to be observed during both time periods.

Table 1 shows that compared with a completely randomized CRT, the design effect of a CRT with cluster-level crossover is reduced by $m\rho_C$, where ρ_C is the correlation between two observations in the same cluster in different periods [see Giraudeau et al. (29)]. From a variance components perspective, the within- and between-cluster variances are reduced by factors of $1 - m\rho_C$ and $1 - \rho_C$, respectively. These formulas assume there are no period or carryover effects. Variants of these formulas appear in the literature (46).

CRTs with individual-level crossover have the advantage of eliminating both between-cluster and between-subject variation from the variance of the treatment effect estimator because the comparison is between the same individual in the same cluster in different conditions. This strategy reduces the standard error to an even greater extent, making CRTs with individual-level crossover potentially extremely efficient compared with standard parallel group CRTs (46). In fact, as shown in **Table 1**, if there are no period or carryover effects, a CRT with crossover at the individual level is as statistically efficient as a crossover IRT without clustering (46).

Crossover CRTs have important limitations, however. When crossover is at the individual level, the trials have the same limitations as IRT crossovers. They are appropriate only when treatment

effects are reversible (not curative) and short-lived (25), and the outcome must be repeatable, which rules out end points such as primary infection or mortality (48). Period effects may arise when subjects do better or worse in a subsequent period because their physical or psychological status changes, independent of treatment (48). Another concern is carryover effects, in which the first treatment changes the subject's status and that change persists into the second period (43, 46, 48).

Because many interventions involve education or training that is expected to have an ongoing influence over an individual's behaviors, carryover effects limit the usefulness of individual-level crossover CRTs for intervention studies. For designs with cluster-level crossover, however, these effects may be less of a concern. Carryover effects may be reduced or absent because individuals do not carry over to the next period. Period effects may be reduced for the same reason, although they may still occur because of secular trends or other changes in the cluster environment (46). Of note, there are ways to account for carryover or period effects in the analysis model (43). Another limitation of crossover CRTs is that the overall duration of a crossover study may be greater than that for a parallel group study because time is needed for each cluster to undergo two conditions rather than only one (31). Despite their limitations, the cluster randomized crossover design has been gaining popularity and is increasingly used in nonclinical settings. We provide two examples.

Example of CRT with crossover at the individual level: the OPUS school meal study. Damsgaard et al. (20, 21) report on a cluster randomized crossover trial that investigated the effects on schoolchildren in Denmark of providing school meals that are rich in fish, vegetables, and fiber, consistent with the New Nordic Diet (NND). Many outcomes were assessed, including dietary intake, nutrient status, physical activity, fitness, sleep, growth, body composition, cardiometabolic markers, illness, school absences, well-being, and cognitive function. Nine schools participated. Year group within school was used as the unit of randomization because school activities are typically coordinated by year group in Denmark; thus this design was practical and helped to avoid dietary contamination among peers.

For two consecutive three-month periods, children received school meals based on the NND or their usual packed lunch from home (control condition) in AB or BA order within their year group. Crossover was at the individual level; the same children were exposed to both conditions. Children were assessed at three time points: baseline, after the first period, and after the second period. Owing to the crossover design and large sample, the study had high power, allowing investigators to detect small effect sizes ($d = 0.11$) with 80% power based on a conservative calculation (21).

In this crossover study, there is potential for period effects due to seasonality and child growth and for carryover effects due to long-term metabolic change induced by diet. However, many outcomes were examined, and these issues are less of a concern for some outcomes than for others. On balance, the merits of crossover outweighed the drawbacks.

Example of CRT with crossover at the cluster level: antibiotic resistance in ICUs. An example of a CRT with crossover at the cluster level is a planned trial involving eight ICUs in five European countries, which aims to evaluate the effectiveness of two antibiotic rotation strategies on reducing antimicrobial resistance in Gram-negative bacteria (56). In the mixing strategy, the antibiotic regimen is changed with every new antibiotic course given to a patient. This strategy aims to maximize antibiotic heterogeneity. In the cycling strategy, the regimen changes per time block (weeks or months), and a single regimen is used within each time block. This approach aims to maximize antibiotic homogeneity within each time block. The primary outcome is the prevalence of antibiotic-resistant Gram-negative bacteria among patients in the ICU, determined through monthly point-prevalence screening.

In the crossover design, the ICUs will be randomized to one of the two interventions, implemented for 9 months, followed by a 1-month washout, and then followed by the other strategy for 9 months. Owing to patient turnover, the patients assessed in the same ICU under different conditions are expected to be different.

Due to transmissibility, the bacterial colonization rates of patients within the same ICU are highly dependent. A cluster-randomized design was selected over an IRT to physically separate the two intervention populations (56). The crossover design enables comparison within ICUs, thereby controlling for differences between ICUs in factors such as patient case mix. However, the crossover design also increases the trial duration, making the trial susceptible to case mix fluctuation over time and possibly magnifying the effect of baseline resistance, which can carry forward over time. Thus this trial illustrates some of the trade-offs that are inherent in a crossover design.

SAMPLE SIZE AND POWER PLANNING RESOURCES

We have provided some formulas for sample size calculation in **Table 1**. These formulas apply only to the comparison of two group means and entail assumptions such as constant cluster size, equal variances and correlations in the two arms, and a sufficiently large number of clusters to use the normal rather than the t distribution. Thus we discuss additional resources for power and sample size planning.

Sample size and power calculation for completed randomized parallel group CRTs have been widely discussed (6, 23, 26, 31). Many authors have provided formulas that are adaptations of IRT formulas and are straightforward to apply for posttest comparisons of rates, proportions, or means. Other authors use a multilevel modeling perspective (36, 39, 41). R, SAS, and SPSS code for two- and three-arm CRTs as well as three-level and multisite CRTs for continuous outcomes are provided by Liu (36). Optimal Design Plus software (44, 52) supports these designs and also covers binary outcomes. R code for many different CRT designs is provided in Campbell & Walters (6). For matched-pair and stratified CRT designs, sample size and power formulas applicable to posttest comparisons of rates, proportions, and means are discussed in several texts (6, 23, 26, 31).

Several authors discuss sample size and power for 2×2 factorial designs for CRTs (26, 31, 55). Fewer resources are available for larger factorial or fractional factorial CRT designs. Dziak et al. (24) provide a detailed discussion, showing how calculations for unclustered factorial designs and single-factor CRT designs can be adapted, and they provide power formulas and a SAS macro (38) for both full and fractional factorial designs.

For CRTs with crossover at the cluster level, Harrison & Brady (30) and Giraudeau et al. (29) discuss sample size and power for continuous outcomes. Forbes et al. (27) present methods for binary outcomes and unequal cluster sizes. Reich et al. (45) provide a framework for estimating power for crossover and parallel group CRTs via simulation using the R package clusterPower.

Sample size and power methods for CRTs with crossover at the individual level have received less attention. Rietbergen & Moerbeek (46) discuss sample size and power for both types of crossover CRTs (cluster-level and individual-level crossover). They also provide formulas for optimal allocation of clusters and subjects for these designs.

DISCUSSION

We have reviewed several strategies that can help investigators design more efficient cluster randomized trials. Several of these strategies are well known in other fields but are underutilized in

public health, most likely owing to a lack of awareness. For example, public health researchers seem to have largely overlooked the merits of factorial experiments. We hope it will become more widely known that a factorial experiment analyzed using a classic factorial ANOVA can be a highly efficient design.

Some strategies may be underutilized owing to perceived complexity. Some of the approaches do require more statistical expertise. For example, in the OPUS School Meal Study, the data had a nested and longitudinal structure, which necessitated the use of a linear mixed model with multiple random effects for the outcome analysis (20). However, these designs can yield substantial benefits in terms of efficiency and cost savings, and a well-trained statistician—who should be part of the study team anyway—should be able to implement them.

There is increasing interest in the use of digital technologies to deliver behavioral and mental health interventions, owing to their low delivery costs and tailorability (11, 19, 57). Strategies such as fractional factorial designs should be more widely considered for such studies because the technology can easily facilitate control over the combination of intervention components delivered.

Crossover CRTs, and in particular CRTs with cluster-level crossover, can also be highly efficient. Although CRTs with individual-level crossover suffer from the same limitations as do individually randomized crossover trials, when crossover is at the cluster level, the individuals in each period within a cluster are different. This approach broadens the range of interventions and outcomes that are suitable for a crossover trial because of the reduced risk of carryover effects and the potential to accommodate nonrepeatable end points.

This review has not exhausted the possible strategies for improving the power and precision of CRTs. Other strategies include covariate adjustment, which can increase precision if the covariate is associated with variation in the outcome (6, 31, 39). This is the well-known analysis of covariance (ANCOVA) approach (8). Also on the horizon are optimal design methods for CRTs, which seek to find designs that minimize variance within cost constraints (see 55 for a recent review).

More power and sample size calculation tools are needed to facilitate adoption of these trial designs. Unfortunately, sample size and power procedures in standard statistical software packages typically do not accommodate CRT designs. Because clustered data are common, we hope that software developers will respond to this need. In sum, the field of public health would benefit greatly by increasing the adoption of both older and newer design ideas that can increase the efficiency of cluster randomized trials.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Andrade S, Lachat C, Ochoa-Aviles A, Verstraeten R, Huybregts L, et al. 2014. A school-based intervention improves physical fitness in Ecuadorian adolescents: a cluster-randomized controlled trial. *Int. J. Behav. Nutr. Phys. Act.* 11:153
2. Arnup SJ, Forbes AB, Kahan BC, Morgan KE, McDonald S, McKenzie JE. 2014. The use of the cluster randomized crossover design in clinical trials: protocol for a systematic review. *Syst. Rev.* 3:86
3. Bastani R, Glenn B, Maxwell AE, Jo AM, Herrmann AK, et al. 2015. Cluster-randomized trial to increase hepatitis B testing among Koreans in Los Angeles. *Cancer Epidemiol. Biomark. Prev.* 24:1341–49
4. Box GEP, Hunter JS, Hunter WG. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery*. Hoboken, NJ: Wiley. 2nd ed.

5. Caldwell LL, Smith EA, Collins LM, Graham JW, Lai M, et al. 2012. Translational research in South Africa: evaluating implementation quality using a factorial design. *Child Youth Care Forum* 41:119–36
6. Campbell MJ, Walters SJ. 2014. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research*. Chichester, UK: Wiley
7. Campbell MK, Piaggio G, Elbourne DR, Altman DG, CONSORT Group. 2012. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 345:e5661
8. Casella G. 2008. *Statistical Design*. New York: Springer
9. Chakraborty B, Collins LM, Strecher VJ, Murphy SA. 2009. Developing multicomponent interventions using fractional factorial designs. *Stat. Med.* 28:2687–708
10. Chow S-C, Shao J, Wang H. 2008. *Sample Size Calculations in Clinical Research*. Boca Raton, FL: Chapman & Hall/CRC
11. Christensen H, Batterham PJ, O'Dea B. 2014. E-Health interventions for suicide prevention. *Int. J. Environ. Res. Public Health* 11:8193–212
12. Collins LM, Baker TB, Mermelstein RJ, Piper ME, Jorenby DE, et al. 2011. The Multiphase Optimization Strategy for engineering effective tobacco use interventions. *Ann. Behav. Med.* 41:208–26
13. Collins LM, Dziak JJ, Kugler KC, Trail JB. 2014. Factorial experiments: efficient tools for evaluation of intervention components. *Am. J. Prev. Med.* 47:498–504
14. Collins LM, Dziak JJ, Li RZ. 2009. Design of experiments with multiple independent variables: a resource management perspective on complete and reduced factorial designs. *Psychol. Methods* 14:202–24
15. Collins LM, Murphy SA, Nair VN, Strecher VJ. 2005. A strategy for optimizing and evaluating behavioral interventions. *Ann. Behav. Med.* 30:65–73
16. Collins LM, Murphy SA, Strecher V. 2007. The Multiphase Optimization Strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART)—new methods for more potent eHealth interventions. *Am. J. Prev. Med.* 32:S112–18
17. Cornfield J. 1978. Randomization by group: a formal analysis. *Am. J. Epidemiol.* 108:100–2
18. Crespi CM, Maxwell AE, Wu S. 2011. Cluster randomized trials of cancer screening interventions: Are appropriate statistical methods being used? *Contemp. Clin. Trials* 32:477–84
19. Cunningham JA, Gulliver A, Farrer L, Bennett K, Carron-Arthur B. 2014. Internet interventions for mental health and addictions: current findings and future directions. *Curr. Psychiatry Rep.* 16:5
20. Damsgaard CT, Dalskov SM, Laursen RP, Ritz C, Hjorth MF, et al. 2014. Provision of healthy school meals does not affect the metabolic syndrome score in 8–11-year-old children, but reduces cardiometabolic risk markers despite increasing waist circumference. *Br. J. Nutr.* 112:1826–36
21. Damsgaard CT, Dalskov SM, Petersen RA, Sorensen LB, Molgaard C, et al. 2012. Design of the OPUS School Meal Study: a randomised controlled trial assessing the impact of serving school meals based on the New Nordic Diet. *Scand. J. Public Health* 40:693–703
22. Donner A. 1998. Some aspects of the design and analysis of cluster randomization trials. *J. R. Stat. Soc. Ser. C-Appl. Stat.* 47:95–113
23. Donner A, Klar N. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research*. New York: Oxford Univ. Press
24. Dziak JJ, Nahum-Shani I, Collins LM. 2012. Multilevel factorial experiments for developing behavioral interventions: power, sample size, and resource considerations. *Psychol. Methods* 17:153–75
25. Elbourne DR, Altman DG, Higgins JPT, Curtin F, Worthington HV, Vail A. 2002. Meta-analyses involving cross-over trials: methodological issues. *Int. J. Epidemiol.* 31:140–49
26. Eldridge S, Kerry S. 2012. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Chichester, UK: Wiley
27. Forbes AB, Akram M, Pilcher D, Cooper J, Bellomo R. 2015. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: application to studies of near-universal interventions in intensive care. *Clin. Trials* 12:34–44
28. Fretheim A, Zhang F, Ross-Degnan D, Oxman AD, Cheyne H, et al. 2015. A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation. *J. Clin. Epidemiol.* 68:324–33
29. Giraudeau B, Ravaud P, Donner A. 2008. Sample size calculation for cluster randomized cross-over trials. *Stat. Med.* 27:5578–85

30. Harrison D, Brady AR. 2004. Sample size and power calculation using the noncentral t-distribution. *Stata J.* 4:142–53
31. Hayes RJ, Moulton LH. 2009. *Cluster Randomised Trials*. Boca Raton, FL: CRC Press
32. Julious SA. 2010. *Sample Sizes for Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC
33. Keppel G, Wickens TD. 2004. *Design and Analysis: A Researcher's Handbook*. Upper Saddle River, NJ: Pearson Prentice Hall. 4th ed.
34. Kim DA, Hwong AR, Stafford D, Hughes DA, O'Malley AJ, et al. 2015. Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *Lancet* 386:145–53
35. Kuehl R. 1999. *Design of Experiments: Statistical Principles of Research Design and Analysis*. Pacific Grove, CA: Duxbury/Thomson
36. Liu X. 2014. *Statistical Power Analysis for the Social and Behavioral Sciences*. New York: Routledge
37. Mdege ND, Brabyn S, Hewitt C, Richardson R, Torgerson DJ. 2014. The 2×2 cluster randomized controlled factorial trial design is mainly used for efficiency and to explore intervention interactions: a systematic review. *J. Clin. Epidemiol.* 67:1083–92
38. Methodol. Cent. 2013. *FactorialPowerPlan*, version 1.0 [software]. Pa. State Univ., Univ. Park. **https://methodology.psu.edu/downloads/factorialpowerplan**
39. Moerbeek M, Teerenstra S. 2016. *Power Analysis of Trials with Multilevel Data*. Boca Raton, FL: CRC Press/Taylor & Francis Group
40. Moser CA, Kalton G. 1972. *Social Methods in Social Investigation*. New York: Basic Books
41. Murray DM. 1998. *Design and Analysis of Group-Randomized Trials*. New York: Oxford Univ. Press
42. Nair V, Strecher V, Fagerlin A, Ubel P, Resnicow K, et al. 2008. Screening experiments and the use of fractional factorial designs in behavioral intervention research. *Am. J. Public Health* 98:1354–59
43. Raghavarao D, Padgett L. 2014. *Repeated Measurements and Cross-Over Designs*. Hoboken, NJ: Wiley
44. Raudenbush S. 2011. *Optimal design software for multilevel and longitudinal research*, version 3.01 [software]. William T. Grant Found., New York. **http://www.wtgrantfoundation.org**
45. Reich NG, Myers JA, Obeng D, Milstone AM, Perl TM. 2012. Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. *PLOS ONE* 7:e35564
46. Rietbergen C, Moerbeek M. 2011. The design of cluster randomized crossover trials. *J. Educ. Behav. Stat.* 36:472–90
47. Sanson-Fisher RW, D'Este CA, Carey ML, Noble N, Paul CL. 2014. Evaluation of systems-oriented public health interventions: alternative research designs. *Annu. Rev. Public Health* 35:9–27
48. Senn S. 2002. *Cross-Over Trials in Clinical Research*. Chichester, UK: Wiley. 2nd ed.
49. Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton-Mifflin
50. Simpson JM, Klar N, Donnor A. 1995. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *Am. J. Public Health* 85:1378–83
51. Smith E, Palen LA, Caldwell L, Flisher A, Graham J, et al. 2008. Substance use and sexual risk prevention in Cape Town, South Africa: an evaluation of the HealthWise program. *Prev. Sci.* 9:311–21
52. Spybrook J, Bloom H, Congdon R, Hill C, Martinez A, Raudenbush S. 2011. *Optimal design plus empirical evidence: documentation for the "optimal design" software*, version 3.0 [software]. William T. Grant Found., New York. **http://www.wtgrantfoundation.org**
53. Strecher VJ, McClure JB, Alexander GL, Chakraborty B, Nair VN, et al. 2008. Web-based smoking-cessation programs - results of a randomized trial. *Am. J. Prev. Med.* 34:373–81
54. Tao J, Rao T, Lin L, Liu W, Wu Z, et al. 2015. Evaluation of Tai Chi Yunshou exercises on community-based stroke patients with balance dysfunction: a study protocol of a cluster randomized controlled trial. *BMC Complement. Altern. Med.* 15:555
55. van Breukelen GJP. 2013. Optimal experimental design with nesting of persons in organizations. *J. Psychol./Z. Psychol.* 221:145–59
56. van Duijn PJ, Bonten MJ. 2014. Antibiotic rotation strategies to reduce antimicrobial resistance in Gram-negative bacteria in European intensive care units: study protocol for a cluster-randomized crossover controlled trial. *Trials* 15:277
57. Wallace P, Bendtsen P. 2014. Internet applications for screening and brief interventions for alcohol in primary care settings - implementation and sustainability. *Front. Psychiatry* 5:151

58. West SG, Aiken LS. 1997. Toward understanding individual effects in multicomponent prevention programs: design and analysis strategies. In *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*, ed. KJ Bryant, M Windle, SG West, pp. 167–209. Washington, DC: Am. Psychol. Assoc.
59. Wu C, Hamada MS. 2000. *Experiments: Planning, Analysis and Parameter Design Optimization*. New York: Wiley
60. Wyrick D, Rulison K, Fearnow-Kenney M, Milroy J, Collins L. 2014. Moving beyond the treatment package approach to developing behavioral interventions: addressing questions that arose during an application of the Multiphase Optimization Strategy (MOST). *Transl. Behav. Med.* 4:252–59
61. Yates F, Mather K. 1963. Ronald Aylmer Fisher. 1890–1962. *Biogr. Mem. Fellows R. Soc.* 9:91–120

Annual Review of
Public Health
Volume 37, 2016

Contents

Epidemiology and Biostatistics

Environmental and Occupational Health

Public Health Practice

Social Environment and Behavior

Health Services

Indexes

Errata

An online log of corrections to *Annual Review of Public Health* articles may be found at http://www.annualreviews.org/errata/publhealth