

Cluster-Randomized Trials

NATHAN R. TODD AND PATRICK J. FOWLER

Community-based researchers often are interested in implementing and evaluating interventions at the level of the community. Whether a community-wide intervention to decrease youth violence (e.g., Hawkins et al., 2012) or a classroom-based universal prevention program to promote behavioral regulation (e.g., Kellam et al., 2014), community interventions frequently are conceptualized, designed, and implemented for entire groups. Testing the effectiveness of the intervention and the mechanisms responsible for change is of paramount importance not only to community-based researchers but also to funders, policymakers, and others interested in how best to promote health and wellness. In this chapter we present (a) cluster-randomized trials (CRTs) as one useful research design for evaluating community-level interventions and (b) multilevel modeling as an efficient way to analyze the results of such trials. We present a general introduction to CRTs with a focus on design basics and strategies to increase power and precision. We then connect these designs to the appropriate multilevel model for analysis. Finally, a case study showcases the process of CRT design and analysis, as well as the benefit of using CRTs to understand whether and how community-based interventions achieve their goals.

DESIGN AND ANALYSIS OF CLUSTER- RANDOMIZED TRIALS

Overview of Cluster-Randomized Trials

Cluster-randomized trials (CRTs), also known as group-randomized trials, community trials, or cluster-randomized studies, are characterized by randomly assigning intact social groups (e.g.,

schools, neighborhoods, entire cities) to intervention and control conditions (Murray, 1998). The group, or “cluster,” is the unit of randomization. For example, 20 schools may be randomly assigned to an intervention or control condition where all students in the same school receive the intervention (or not). CRTs enable researchers to study naturally occurring groups where the randomization of individuals is not possible due to ethical, logistical, political, or other reasons (Cook, 2005). For example, spillover (i.e., contamination) may be more of a concern if people within the same setting are randomly assigned to different interventions; however, randomizing groups minimizes spillover (because everyone in the same group receives the same intervention) and may be a more palatable option to communities.

Moreover, the purpose of community-based interventions often is to change something about the social environment, norms, community practices, or setting as the mechanism to shape individual behavior (Cook, 2005; Raudenbush, Martinez, & Spybrook, 2007). CRTs randomize at the same level of intervention deployment and offer a more ecologically valid approach to examining group-based interventions, such as randomizing entire classrooms to receive an intervention rather than dividing students within classrooms (Raudenbush, 1997). Also, the inclusion of randomization strengthens internal validity and begins to address bias of individuals self-selecting into preexisting groups, as well as other threats to internal validity (Cook, 2005; Murray, 1998). Clearly, there are many benefits to CRTs.

Although there are many benefits to CRTs, they also can be more costly and complex in scope, design, and analysis than other types of randomized experiments (Cook, 2005). CRTs require data

collection on units within settings. Thus, considerations must account for reliable measurement, as well as adequate numbers of both individuals and settings to provide a feasible test of the intervention. In addition to data collection, necessary resources must be committed to monitor implementation of the intervention in order to assess threats to the validity of the design (e.g., uptake, compliance, contamination, attrition). For example, many CRTs have at least 20 unique groups to provide an adequate test of intervention effects, with substantial cost in implementing and monitoring the intervention in such a large number of groups. Also, because individuals are nested within groups, analytic methods that account for this clustering need to be used (Murray, 1998; Raudenbush, 1997). Importantly, ethical considerations, especially informed consent, become more challenging given that individuals may not be able to fully agree or avoid exposure to a setting-level intervention decided upon by representatives from the larger group (Sim & Dawson, 2012). The complexities require active collaboration with community gatekeepers and an engaged institutional review board to help ensure ethical practices.

The challenges of CRTs must be weighed with their potential benefits, addressing important research questions concerned with group-level processes. This chapter reviews a number of advances in theory and application that mitigate some of these complications, especially pertaining to the number of settings required to power the study and other adaptive design issues that make CRTs more feasible in terms of implementation and budget. We start with a presentation of CRT design basics. Although not exhaustive, these designs show possibilities and illustrate initial decisions that must be made by community-based researchers.

Design of Cluster-Randomized Trials

Cluster-Randomized Trial

Design Basics

Plans for implementation of a CRT begin with a clear articulation of the research questions. Deliberation must determine whether a CRT represents the most useful and practical design to address questions. Considerations must account for ethical practice, time, available resources, and threats to internal and external validity (Murray, 1998; Shadish, Cook, & Campbell, 2002). Although designs may include more than two conditions (e.g.,

factorial CRTs with a control group and multiple intervention conditions; e.g., Peters et al., 2003), we focus on a traditional intervention and control group design and use Murray's (1998) terminology to describe the different components.

An initial decision involves whether the design will include assessment only at the completion or endpoint of the trial (i.e., posttest-only control group design), will also collect pretest data (pretest-posttest control group design), or will include more than two assessment points. As with all experimental research designs (Shadish et al., 2002), the main weakness of the posttest-only design is the lack of information regarding selection bias and maturation (Murray, 1998). However, Murray noted that randomization of enough groups to conditions may begin to mitigate these concerns. The pretest design includes baseline or pretest data collection that allows conditions to be compared prior to the intervention and may also decrease other threats to internal validity. Importantly, baseline information also can be used to match groups prior to random assignment or as covariates in later analyses to increase power. As described later, availability of pretest assessment provides a number of benefits to testing intervention effects; however, time and resource limitations may influence feasibility. These tradeoffs should be carefully considered when planning a CRT.

Another early design decision pertains to sampling. Researchers must determine whether to collect data in a cross-sectional versus cohort design. As discussed by Murray (1998), a cohort design collects data on the same group of individuals followed over time and sampled at each measurement occasion, yielding longitudinal data. In a cross-sectional design, the group is sampled but different group members are assessed at each measurement occasion. For example, in a city-wide intervention to decrease smoking, at baseline a group may be randomly sampled from the city, whereas at the conclusion of the intervention a new group, not including any of the original members, would be sampled. The distinction between the cohort and cross-sectional designs is important, as it reflects different research questions. In a cross-sectional design the question is about change within the population, whereas a cohort design focuses on average individual change, requiring repeated observations of the same set (i.e., cohort) of individuals. Additionally, the cohort or cross-sectional nature

of the design necessitates slightly different analytic models that influence the ability to detect program effects, as described later. Interested readers should consult Murray (1998) for other considerations when selecting a design and should be guided by the primary research question of interest.

Power

Power refers to the ability to detect an effect if the effect actually exists. It depends on such aspects of the design as size of the effect, number of participants, and level of statistical significance. However, there are more factors that influence power in CRTs compared with designs at the individual level (Murray, 1998). Power in CRTs also depends on the number of clusters and the variance between clusters on the outcome variables (Raudenbush et al., 2007). It should be no surprise that the additional considerations for power are directly connected to the nested nature of the CRT design, as nesting often creates correlated, dependent observations. Raudenbush (1997) noted that dependence may occur within a group for multiple reasons, such as people self-selecting into a group based on similar characteristics or having common experiences or mutual interactions once they are in the group. The intraclass correlation (ICC) is used as an index of the degree of dependence in a group or the proportion of variance that is between groups (Raudenbush, 1997). Dependence violates the independence assumption of ordinary least squares regression and tends to produce downward biased standard errors, which, in turn, results in a more liberal test of significance (Murray, 1998). Failure to account for this dependence increases the risk of committing a Type I error. However, appropriate methods of analysis (i.e., multilevel mixed models) account for this between-group variability and produce accurate estimates of the standard error of the treatment effect.

Holding other factors constant, CRTs tend to have less power than traditional individual-level randomized trials because the variance of the condition mean will systematically be higher in nested designs (Moerbeek & Terenstra, 2011; Murray, 1998). In fact, the stronger the dependence (i.e., the larger the ICC), the greater the variance of the condition mean. Scholars call this the design effect, or variance inflation factor, quantified as $(1+(n_i-1)ICC)$, where n_i is the sample size per group (Moerbeek & Teerenstra, 2011; Murray,

1998). The intuitive implication is that increased dependence, as indexed by the ICC, increases the variance of the condition mean, which, in turn, decreases power. Thus, one strategy for increasing power is to lower the ICC (Murray & Blitstein, 2003). One way to do this is to select an outcome that tends to exhibit less variability between groups (Murray, 1998). Another is to include statistical controls that lower the ICC. Indeed, as noted by Cook (2005), it is the conditional ICC (i.e., the ICC conditioned on all variables in the model) that contributes to the design effect; thus, reducing the ICC through covariates may increase power. Both matching and covariates are ways to statistically reduce the ICC and noise in the study as well as to increase precision.

The general idea behind both matching and covariates is to include other variables in the model that are strongly related to the outcome of interest in order to reduce the unexplained variance, to decrease noise, and to decrease the ICC, all of which may increase power (Raudenbush, 1997; Raudenbush et al., 2007). Similar to other experimental designs (Shadish et al., 2002), matching involves selecting a variable that is correlated with the outcome of interest, ranking each group based on this variable, and then randomly assigning pairs of similar groups to the treatment or control condition. Raudenbush et al. (2007) showed that such matching may increase power when between-group variation is large and the variable is strongly related to the outcome. Matching in effect cuts down on the random noise among groups between conditions. Groups also may be matched on other characteristics to improve the face validity of comparing treatment and control conditions, such as balancing groups based on race/ethnicity or other demographic variables (Raudenbush et al., 2007). The drawback to matching is that this information also needs to be included in the analysis, which begins to cost degrees of freedom, which, in turn, decreases power (see Murray, 1998, for how to include matching as a random effect in the analytic model). Clearly, a balance exists between increasing power through matching with potential loss of power by losing degrees of freedom.

Similar to matching, covariates are collected before the intervention and should be strongly related to the outcome of interest. Covariates can be at the level of the cluster or individual, and an emerging literature (e.g., Konstantopoulos,

2012) examines the benefit of including covariates at different levels of analysis. For example, a study of school-based intervention could include individual student characteristics (e.g., pretest scores) and aspects of the school, such as size, teacher-student ratio, or socioeconomic status. Interestingly, a covariate will be more effective at increasing power the more the covariate helps to explain between-group differences, thus decreasing the ICC. Thus, carefully chosen covariates can dramatically increase power or alternatively decrease the number of groups needed to achieve desired power.

Inclusion of covariates requires additional assumptions. For instance, variables must demonstrate similar associations in treatment and control conditions, and residuals must be normally distributed with constant variance (Raudenbush, 1997). Also, there is a tradeoff involved including covariates, as this also lowers the degrees of freedom in the model; usually, though, the benefit to power is in favor of including the covariate. In general, the use of covariates tends to increase power more than the use of matching, although matching may help to increase face validity by balancing groups on certain characteristics (Raudenbush et al., 2007). As noted later, resources exist to calculate the exact benefit of including covariates to increase power.

In the planning stages of a CRT, estimates of key factors (e.g., ICC, potential variance explained by covariates) are needed to calculate power. Similar to other designs, researchers also must specify an effect size; they need to forecast how much the treatment groups will differ on outcomes as a result of the intervention. Such an effect may be based on previous research or on what constitutes practical differences. Also, scholars have compiled common ICCs found in educational research (Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007) and other types of community samples (Murray, Varnell, & Blitstein, 2004) that can inform estimates used to calculate power. Researchers also may be guided by previous research and experience to estimate the potential impact of covariates at both individual and group levels of analysis.

Based on this information, researchers can use free, intuitive programs like Optimal Design to calculate power (Raudenbush et al., 2007). As noted in the Optimal Design Documentation (Spybrook et al., 2011), the user can plot power charts versus

the cluster size, the total number of clusters, the ICC, the effect size, and the amount of variance explained by covariates. One approach is to specify the minimum detectable effect size (MDES), or the smallest program effect in standard deviations of the outcome expected to be seen given other design considerations (i.e., number of clusters, conditional ICC, subjects, significance level, power). Optimal Design plots the MDES on the *y*-axis against the same information to provide a visual tool to aid researchers making design decisions. Outcomes can include continuous or binary distributions. This flexible program is a valuable resource for estimating power and gives community-based researchers the necessary resources to calculate power when designing a CRT.

Clearly, power costs resources. In order to maximize power for a given budget, the CRT literature (Moerbeek & Teerenstra, 2011; Raudenbush, 1997) has focused on determining the “optimal design” for a study. For example, there are differential costs and impacts on power for sampling more individuals within a setting versus recruiting an entire new setting. Calculating the optimal design usually focuses on minimizing the variance of the condition effect while considering the costs of modifying various aspects of the design (Raudenbush et al., 1997). Fortunately, the program Optimal Design includes a module to enter the total budget, cost per cluster, cost per cluster member, the significance level, the ICC, and the MDES. The program then calculates the optimal sample size per cluster and number of clusters as well as reports the power for such a design. This information may be a useful starting place to then use other modules in the program to examine the impact of including covariates or adjusting other design parameters. Clearly, resources such as the program Optimal Design provide the necessary tools for community-based researchers to design adequately powered CRTs.

Adaptive Designs

A number of modifications can be incorporated into basic CRT designs to address common questions in community research (Brown et al., 2009). Longitudinal designs that integrate multiple repeated measures allow investigation of change over time associated with group-based intervention. Beyond testing simple differences between groups, researchers can test hypotheses regarding

whether treatment effects grow or diminish over time, as well as the shape of change in outcomes (e.g., linear, quadratic, exponential). Repeated measures also enhance the validity of CRTs; modeling within-person variation in outcomes provides more precise and, thus, more powerful estimates of program effects. Having more measurements also may provide information to test threats to internal validity, such as maturation, regression, and instrumentation. Likewise, the design enables the testing of mechanisms that may occur at the group level (e.g., social norms, social processes) to better understand how changing the setting or social ecology shapes individual behavior and attitudes (Fowler & Todd, *in press*). The increasing availability of administrative records provides an inexpensive way to leverage longitudinal designs.

Flexibility also exists around random assignment in CRTs. Studies may include multiple arms of an intervention, such as when testing dosage effects or multiple new interventions. For example, school-based prevention programs might stack intervention components such that schools within a district are randomized to receive a universal prevention program for all students, and intervention schools are further randomized to receive a selective or targeted intervention for at-risk students or both. Looking at school outcomes, the design simultaneously tests (a) the effects of the universal program versus treatment as usual; (b) the benefit of the universal program plus the selective component; and (c) effects of the combination of universal, selective, and targeted interventions. Interest in broad policy reforms common in community research makes CRT a useful tool; however, these designs require considerable forethought as to the degrees of freedom and number of groups needed to test research questions.

Researchers also may leverage CRT designs to study the rollout of new programs across units within a network of groups (Wyman, Henry, & Brown, 2015). The design works well when communities intend to make policy or program changes across groups but limited resources preclude making changes all at once. For instance, new procedures might require training across large numbers of geographic locations that are logistically impossible to do all at once. Relatedly, communities might be waiting to secure additional funding for new services before fully implementing the new intervention across the organization.

By randomly assigning when groups receive an intervention, researchers create a rigorous test of the short-term benefits of the program; that is, random assignment occurs for both groups and time. Groups waiting to receive the intervention serve as the randomly assigned control group until it is their “turn” to get the intervention, at which time they become part of the treatment group. Outcome data are collected at consistent time points across the entire rollout period, providing repeated measures before and after the intervention time point. By incorporating longitudinal outcomes into the CRT design, statistical power to detect small effects is greatly enhanced and makes the design very feasible with a relatively small number of groups (Brown, Wyman, Guoa, & Penab, 2006). The rollout design often appeals to community-based organizations because everyone receives the intervention and strong information is provided on the program. Importantly, the rollout CRT tests only the immediate effects of the intervention—the control group immediately transitions into the treatment. Thus, researchers must carefully consider the research questions of interest and the nature of the intervention. Delayed or compounding effects of the intervention will not be captured in rollout designs. Other adaptive features exist for CRTs, and interested readers are encouraged to review additional resources on the strengths and limitations of different components (Brown et al., 2009).

Analysis of Cluster-Randomized Trials Using Multilevel Modeling

This section provides an overview of multilevel modeling as an appropriate approach for analyzing data from CRTs, with a focus on connecting basic designs with analytic models. Multilevel models (MLMs; also known as multilevel linear models, mixed- or random-effects models, and hierarchical linear models) are an appropriate analytic strategy for analyzing CRTs because they use information about variance at multiple levels of analysis and produce accurate standard errors (Murray, 1998; Raudenbush, 1997). These models are now more accessible to understand and analyze within many statistical programs (Murray, 1998). Although a full treatment of MLM is beyond this chapter (see Fowler & Todd, *in press*, and Raudenbush & Bryk, 2002, for additional information), a presentation of

the multilevel model illustrates applications to analyzing CRTs.

Multilevel models are unique due to the inclusion of random effects and other variables at both the individual and group levels of analyses. Usually the lowest level is called "Level 1" (such as students), and the cluster they are nested in is the higher level, known as "Level 2" (such as classrooms). Blending the notation of Raudenbush (1997) and Murray (1998), in the simplest posttest-only CRT design, the multilevel model (Model A) may be written as:

$$Y_{ij} = \gamma_0 + \gamma_1 C_j + \mu_j + e_{ij}$$

where the *i*th individual is nested within the *j*th group, γ_0 is the grand mean, γ_1 is the treatment contrast for condition (usually effect coded $-.$.5 and $.$.5), μ_j is the Level-2 error term (also known as the random effect for group), and e_{ij} is the Level-1 error term. Random effects are bolded in the model. This model also assumes that $\mu_j \sim N(0, \tau^2)$ and independent, and $e_{ij} \sim N(0, \sigma^2)$ and independent, where τ^2 is the between-cluster variance and σ^2 is the within-cluster variance (Raudenbush, 1997). The intraclass correlation (ICC) is $\frac{\tau^2}{\tau^2 + \sigma^2}$.

What is important to note from this model are the separate random effects at Level 1 (i.e., e_{ij}) and Level 2 (i.e., μ_j) and that the ICC will increase the more variability there is between relative to within clusters. Also, the effect (i.e., γ_1) for C_j is of primary interest in determining the effect of the intervention.

As noted by Murray (1998), for a posttest-only design either a cross-sectional or cohort sample would be analyzed with this same Model A because data are collected only at the conclusion of the trial. The only difference is that in the cohort sample the only people included in the survey sample would be those who were present at the start of the intervention. In either case, covariates assessed at the beginning of the intervention (or that would not change due to the intervention, such as gender) also could be added to the model to increase precision and power.

If the design collected both pretest and posttest measures, for a nested cross-sectional design the model (Model B) would be as follows (Murray 1998):

$$Y_{ijk} = \gamma_0 + \gamma_1 C_j + \gamma_2 T_k + \gamma_3 CT_{jk} + \mu_j + T\mu_{jk} + e_{ijk}$$

where T_k indicates if the person was in the first or second wave of data collection and CT_{jk} indicates the interaction between wave and condition. $T\mu_{jk}$ is the random effect for the interaction. In this design, the primary interest is in the CT_{jk} interaction as a way to determine the effectiveness of the intervention, with follow-up tests focusing on decomposing the interaction to understand how mean values on the outcome are similar or different for the control and intervention condition across time points. One would hope for differences in means from Time 1 to Time 2 for the intervention but not control condition, with similar means between conditions at Time 1. Covariates also could be added to this model to increase power.

A cohort design with pretest and posttest data would be analyzed with a very similar model to Model B, but individuals in this sample would be assessed at both time points and the model would add additional random effects for the person (see Murray, 1998). The addition of these random effects, and the ability of individuals to serve as their own control across time, serves to further increase power. Covariates also could be added to this model to increase power. Alternatively, as a special case, in a pretest-posttest cohort design, the outcome measured at Time 1 (e.g., reading score at Time 1) can serve as a covariate in the model predicting the outcome at Time 2 (e.g., Time 2 reading score). Thus, time is incorporated into the model (Model C) in a different way by including this covariate, such as:

$$(Time2Y)_{ij} = \gamma_0 + \gamma_1 C_j + \gamma_2 (Time1Y)_{ij} + \mu_j + e_{ij}$$

Model C is exactly the same as Model A, but information for Time 1 is introduced as a covariate in the model, while other covariates also could be added. The larger point is that the CRT design (posttest only, pretest-posttest, cross-sectional, or cohort) has direct implications for how to specify the analytic model. Power is likely increased when repeated observations and person- and group-level covariates are included. Covariates, such as Time 1 scores, may be especially potent in increasing power (Cook, 2005). Readers interested in further elaboration of these models, how to incorporate matching in the design and analysis, and SAS syntax for implementing such models are directed to Murray (1998).

A final advantage of multilevel modeling is the ability to generalize to other types of outcomes beyond continuous variables. Multilevel models fall under the broad umbrella of the generalized linear mixed model, which allows for outcomes that are discrete, binary, count, rates, and continuous. Scholars have discussed how MLM can incorporate such outcomes in general (Raudenbush & Bryk, 2002) and in particular for CRTs (Eldridge & Kerry, 2012; Murray, 1998; Murray et al., 2004). Such resources should be consulted to determine how power may be impacted by the type of outcome when planning a CRT.

Summary of General Steps in Designing a Cluster-Randomized Trial

As is clear from this description, careful planning of a CRT can help community-based researchers achieve adequate power while minimizing cost. However, there are many steps to consider beyond power, cost, and analysis (Murray, 1998). Among the most crucial are to clearly articulate the guiding research question, the theory underlying the intervention, and the mechanisms that are proposed to result in change (Cook, 2005). Such clarity informs whether the focus is on population change or individual change, which may help guide the researcher in selecting a cross-sectional or cohort design. In particular, clarity is needed in specifying theory and mechanism at multiple levels of analysis with respect to how processes may operate differently at individual versus group levels. If the social ecology or setting is the intervention target, the mechanisms of change expected to result in the desired outcomes should clearly be explicated. Statistical analysis cannot redeem an intervention that does not have a clear theoretical focus that produces testable hypotheses.

Given the high cost of a CRT, scholars also recommend conducting a pilot study to provide a general proof of principle that the intervention tends to work in the way that it is proposed (Murray, 1998). Such a pilot study may be conducted with only a few groups but also will provide an opportunity to refine the intervention and to anticipate further challenges with implementation. Even small effects in a pilot study may warrant a larger trial. Although beyond the scope of this chapter, plans also should be made to monitor the implementation of the intervention and to collect ongoing process data (Cook, 2005; Murray, 1998). Especially in the case of null

findings, such information may be incorporated into the analysis (such as dose effects) or may further serve to contextualize why effects were present or not. A pilot study provides the opportunity to work out these details before investing in a larger CRT.

Early in this process a power analysis should be conducted to ensure that enough resources are available for an adequately powered CRT. As discussed earlier, estimates of the ICC from previous research can be used, along with thoughtful selection of variables for matching or covariates. As a part of this process, the optimal design should be calculated to ensure that there are enough resources to sample an adequate number of individuals and clusters. Also, before launching the CRT, the analytic method should be selected to ensure that all appropriate information is gathered during the CRT for use in analysis. Likely this will all be an iterative process (selection of design, power analysis, pilot study, determination of feasibility) in the planning of an adequately powered CRT.

Applications of Cluster-Randomized Trials

CRTs are increasingly being used in community-based research to test important questions regarding the influence of setting characteristics. In particular, prevention and intervention trials have examined the effects of programs, as well as strategies for implementing evidence-based practices to scale (i.e., with a larger number of settings or communities). Classroom-based interventions targeting low-income students demonstrate long-term effects on healthy child development (Kellam et al., 2014). Universal prevention programs that build coalitions to support the use of evidence-based practices show decreases in youth substance abuse and delinquency into high school (Hawkins et al., 2012). Moreover, studies compare implementation strategies to promote the use of evidence-based practices, including delivering parent training for youth in foster care across child welfare agencies in multiple states (Chamberlain et al., 2013) and addressing culture within community mental health agencies (Glisson, Hemmelgarn, Green, & Williams, 2013). The studies randomize intact groups—including classrooms, schools, counties, and states—in order to investigate theories of setting-level processes. To demonstrate the use and flexibility of CRT designs, we next examine a set of studies targeting adolescent suicide prevention.

CASE STUDY

Background

Adolescent suicide represents a key concern of communities in both the United States and internationally. Suicide represents the third leading cause of death among children and adolescents, with 9% completing suicide each year—a trend that has increased during recent decades (Brown, Wyman, Brinales, & Gibbons, 2007). Communities, and especially school leaders, seek information on practices that promote protective factors and reduce the risk of youth suicide. Evidence suggests the importance of active surveillance for warning signs, as well as immediate action to connect at-risk youth with appropriate mental health resources. Although schools engage with students and their social networks, teachers and staff struggle to provide systematic monitoring, and mental health resources often fail to meet demands. School-based efforts too often provide services haphazardly or revert to traditional one-on-one counseling models that inherently cannot generate reductions in suicide rates.

Challenges also exist in implementing and evaluating evidence-based programs. Individual-level random assignment is not a feasible option; effective surveillance requires participation by school staff interacting across student groups, and students interact within peer networks that extend beyond classrooms. Statistically, suicide represents a relatively rare event that requires considerable power to detect the true effects of prevention efforts; this means large samples followed over time. Brown et al. (2007) have quantified the scope of the challenge by calculating the number of person years (i.e., number of people in a study multiplied by the number of years the people are followed) needed to detect a 50% reduction in the incidence of youth suicide through a universal prevention program; 1 million person years would be needed to detect a 50% drop in the rate of adolescent suicide through a universal prevention program!

Methodology

To address these challenges, a coalition of researchers, school officials, program developers, and other community partners was formed to design a series of adaptive CRTs that would maximize statistical power and provide an adequate test of evidence-based prevention implementation within

schools (Brown et al., 2007). An initial study tested a gatekeeper training model used within schools that educated school staff on (a) recognizing suicide warning signs and (b) communicating with students at risk (Wyman et al., 2008). All teachers and staff received 30-minute group trainings and a brief refresher. A pretest-posttest CRT design asked whether school staff increased their awareness of suicide risk indicators and if they were more likely to act on warning signs. To further test program theory, it was hypothesized that follow-up communications would occur more frequently among the staff who interacted with students around emotionally laden topics before the intervention. This tested whether education was enough to motivate behavior or if additional channels were needed to facilitate communication.

All secondary schools in a Georgia school district ($N = 35$) were stratified by middle versus high school, and by the number of crisis referrals made by schools in the previous academic year. Matched schools were randomized to gatekeeper training or a waitlisted control group. To assess staff awareness and communication, a random sample of staff in different roles (e.g., teacher, nurse) from each school were invited to complete surveys at baseline and 1 year later ($N = 249$). Because the same group was sampled at both time points, this was a cohort sample. Students ($N = 2,059$) also completed anonymous online assessments of suicidal ideation, help-seeking attitudes, and risk behaviors in order to further evaluate the effects of gatekeeper training. Balance of school and teacher characteristics existed across treatment conditions.

Results

Because the outcomes of interest focused on staff learning, and not on the relatively rare event of suicide, 1 million person years were not needed for this phase of the study. Analyses suggested adequate power to detect the anticipated moderate program effects (Cohen's $d = .60$) at a 95% significance level given a modest ICC (ICC = .06) among outcome variables within schools. Multilevel models regressed outcomes on treatment condition, baseline measures as covariates to increase power, school means for outcome variables, and treatment \times baseline interactions, while also including school as a random effect. Intent-to-treat analyses suggested gatekeeper training improved awareness of risk behaviors among all types of school

staff 1 year later, with staff who had reported lower baseline awareness showing the largest improvements. Overall rates of communication with distressed students increased, but the effects were driven by a small subset (14%) of school staff who regularly interacted with at-risk students. In addition, students who reported a history of suicidal behavior were less likely than other students to talk with adults about their distress. The findings supported the use of surveillance and identified primary mechanisms for prevention efforts to achieve reductions in suicide incidence.

Follow-Up

This initially successful CRT spurred follow-up studies that investigated modifications to the gatekeeper model based on initial findings, as well as tested short-term effects on adolescent suicide behaviors (Wyman et al., 2010). An adapted gatekeeper model leveraged adolescent leaders within high schools to deliver key prevention messages across peer networks. School staff nominated adolescent peer leaders, who received 4 hours of training on protective factors and on engaging with trusted adults. Schools also broadly disseminated messages of identifying and talking to a trusted adult through presentations, videos, and texts. Key research questions asked whether short-term changes occurred among peer leader attitudes and behaviors and in school norms around suicide protective factors.

In particular, one follow-up study used a CRT design to balance the needs of researchers and school officials. School officials wanted to implement prevention programming based on positive outcomes of the initial CRT and other piloting being done, while researchers wanted a rigorous test of implementation and outcomes of the program. Through a collaborative process, the team designed a multisite CRT that randomized schools to either the intervention or a 5-month waitlist. In particular, 18 high schools in three states (Georgia, New York, and North Dakota) were matched by state, region, and number of students; schools were randomized to treatment conditions on a one-to-one ratio within each state. Pretest-posttest assessments occurred with 453 peer leaders—half of whom received training during the intervention period—as well as other students at each school ($N = 2,675$). The MDES indicated an ability to detect the expected moderate effects on attitude

changes among peers and across schools. Multilevel models included Level-1 covariates (gender, grade, ethnicity, baseline outcomes) and Level-2 fixed effects of intervention condition and school. The results of the CRT showed significant improvements in both peer and school-wide norms concerning suicide and engaging with adults.

The next phase of research required more innovative methodologies. Based on the accumulation of positive findings of fidelity and norm changes, school officials wanted to apply the intervention across all schools; however, program effects on the specific behaviors that reduce suicide rates had yet to be tested. To test impact on behaviors, more power would be needed than for the prior trials that focused on more easily detected attitude changes. This meant programming would need to be implemented across many more schools, with data collection occurring across hundreds of thousands of students. Given the unfeasibility of such a design, the team decided to focus on a more proximal outcome that would be easier to detect. The theory of change hypothesized reductions in suicide incidence when at-risk youth were identified and connected to needed resources. Gatekeeper training emphasized channels to refer at-risk youth for school-based mental health assessment. The team decided to examine gatekeeper training impact on crisis referrals for school mental health services. Ongoing record keeping of referrals provided readily accessible longitudinal data on all students within schools, which reduced the burden on survey data collection and provided information on the population of students.

A dynamic waitlist CRT optimized the number of schools needed to detect moderate program effects (Brown et al., 2007). The CRT randomized both schools and time to the intervention. In particular, 16 schools were blocked on school characteristics and then randomly assigned to start training at one of four designated time points over a 2-year academic period. Thus, the study started with four schools receiving the intervention, which increased incrementally until all schools had been trained. Data were collected for all students within all schools before and after receipt of the intervention. Across time the design balanced school characteristics that would systematically influence outcomes, and repeated measures of referrals increased efficiency to detect program effects. In particular, the MDES in a traditional waitlist design was powered

to detect a 32% increase in referrals, while the dynamic waitlist was powered to identify a 23% increase. Change in referral rates was modeled over time with a time-varying indicator of whether the school received training. The results demonstrated a short-term effect of gatekeeper training after full implementation of the program (Wyman et al., 2015). The design addressed central research and practice questions; however, other methods would be needed to evaluate other relevant questions, such as maintenance of longer-term effects.

CONCLUSION

The CRT design provides a powerful and flexible approach for testing research questions that address setting characteristics through ecologically valid methods. However, the costs, complexities, and ethical considerations must be weighed when planning a study. Considerations must balance feasibility and accuracy, and a specific theory of change provides a necessary framework for guiding design and analyses choices. Using multiple methods provides greater opportunities to address important research questions. Moreover, community partnerships are key in all stages of designing and implementing CRTs. Deliberations among a wide range of stakeholders must consider issues of informed consent, prioritize questions of interest, ensure fidelity of interventions and their evaluations, and plan utilization of findings. Despite these challenges, CRTs offer much potential for addressing questions at the core of social interventions, as well as for developing truly community-engaged research.

REFERENCES

- Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, 29, 30–59.
- Brown, C. H., Ten Have, T. R., Jo, B., Dagne, G., Wyman, P. A., Muthén, B., & Gibbons, R. D. (2009). Adaptive designs for randomized trials in public health. *Annual Review of Public Health*, 30, 1–25.
- Brown, C. H., Wyman, P. A., Brinales, J. M., & Gibbons, R. D. (2007). The role of randomized trials in testing interventions for the prevention of youth suicide. *International Review of Psychiatry*, 19, 617–631.
- Brown, C. H., Wyman, P. A., Guoa, J., & Pena, J. (2006). Dynamic wait-listed designs for randomized trials: New designs for prevention of youth suicide. *Clinical Trials*, 3, 259–271.
- Chamberlain, P., Roberts, R., Jones H., Marsenich, L., Sosna, T., & Price, J. M. (2013). Three collaborative models for scaling up evidence-based practices. *Administration and Policy in Mental Health and Mental Health Services Research*, 39, 278–290.
- Cook, T. D. (2005). Emergent principles for the design, implementation, and analysis of cluster-based experiments in social science. *Annals of the American Academy of Political and Social Science*, 599, 176–198.
- Eldridge, S., & Kerry, S. (2012). *A practical guide to cluster randomized trials in health services research*. West Sussex, England: Wiley.
- Fowler, P. J., & Todd, N. R. (in press). Methods for multiple levels of analysis: Capturing context, change, and changing contexts. In M. A. Bond, C. Keys, & I. Serrano-García (Eds.), *APA handbook of community psychology*. Washington, DC: American Psychological Association.
- Glisson, C., Hemmelgarn, A., Green, P., & Williams, N. (2013). Randomized trial of the availability, responsiveness and continuity (ARC) organizational intervention for improving youth outcomes in community mental health programs. *Journal of the American Academy of Child and Adolescent Psychiatry*, 52, 493–500.
- Hawkins, J. D., Oesterle, S., Brown, E. C., Monahan, K. C., Abbott, R. D., Arthur, M. W., & Catalano, R. F. (2012). Sustained decreases in risk exposure and youth problem behaviors after installation of the Communities That Care prevention system in a randomized trial. *Archives of Pediatric Adolescent Medicine*, 166, 141–148.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60–87.
- Kellam, S. G., Wang, W., Mackenzie, A. C., Brown, C. H., Ompad, D. C., Or, F., ... Windham, A. (2014). The impact of the Good Behavior Game, a universal classroom based preventive intervention in first and second grades, on high risk sexual behaviors and drug abuse and dependence disorders in young adulthood. *Prevention Science*, 15, S6–S18.
- Konstantopoulos, S. (2012). The impact of covariates on statistical power in cluster randomized designs: Which level matters more? *Multivariate Behavioral Research*, 47, 392–420.
- Moerbeek, M., & Teerenstra, S. (2011). Optimal design in multilevel experiments. In J. Hox & J. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 257–281). New York, NY: Routledge.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

- Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review*, 27, 79–103.
- Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, 94, 423–432.
- Peters, T. J., Richards, S. H., Bankhead, C. R., Ades, A. E., & Sterne, J. A. C. (2003). Comparison of methods for analyzing cluster randomized trials: An example involving a factorial design. *International Journal of Epidemiology*, 32, 840–846.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2, 173–185.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, 29, 5–29.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Houghton Mifflin.
- Sim, J., & Dawson, A. (2012). Informed consent and cluster-randomized trials. *American Journal of Public Health*, 102, 480–485.
- Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal design plus empirical evidence: Documentation for the "Optimal Design" software*. Retrieved June 2015, from <http://hlmsoft.net/od/od-manual-20111016-v300.pdf>
- Wyman, P. A., Brown, C. H., Inman, J., Cross, W., Schmeelk-Cone, K., Guo, J., & Pena, J. B. (2008). Randomized trial of a gatekeeper program for suicide prevention: 1-year impact on secondary school staff. *Journal of Consulting and Clinical Psychology*, 76, 104–115.
- Wyman, P. A., Brown, C. H., LoMurray, M., Schmeelk-Cone, K., Petrova, M., Yu, Q., . . . Wang, W. (2010). An outcome evaluation of the Sources of Strength suicide prevention program delivered by adolescent peer leaders in high schools. *American Journal of Public Health*, 100, 1653–1661.
- Wyman, P. A., Henry, D. B., & Brown, C. H. (2015). Designs for testing group-based interventions with limited numbers of social units: The dynamic wait-listed and regression point displacement designs. *Prevention Science*. Epub ahead of print.

