

Designs for Testing Group-Based Interventions with Limited Numbers of Social Units: The Dynamic Wait-Listed and Regression Point Displacement Designs

Peter A. Wyman · David Henry · Shannon Knoblauch ·
C. Hendricks Brown

Published online: 7 December 2014
© Society for Prevention Research 2014

Abstract The dynamic wait-listed design (DWLD) and regression point displacement design (RPDD) address several challenges in evaluating group-based interventions when there is a limited number of groups. Both DWLD and RPDD utilize efficiencies that increase statistical power and can enhance balance between community needs and research priorities. The DWLD blocks on more time units than traditional wait-listed designs, thereby increasing the proportion of a study period during which intervention and control conditions can be compared, and can also improve logistics of implementing intervention across multiple sites and strengthen fidelity. We discuss DWLDs in the larger context of roll-out randomized designs and compare it with its cousin the Stepped Wedge design. The RPDD uses archival data on the population of settings from which intervention unit(s) are selected to create expected posttest scores for units receiving intervention, to which actual posttest scores are compared. High pretest–posttest correlations give the RPDD statistical power for assessing intervention impact even when one or a few settings receive intervention. RPDD works best when archival data are available over a number of years prior to and following intervention. If intervention units were not randomly selected, propensity scores can be used to control for non-random selection factors. Examples are provided of the DWLD and RPDD used to evaluate, respectively, suicide prevention training (QPR) in 32 schools and a violence

prevention program (CeaseFire) in two Chicago police districts over a 10-year period. How DWLD and RPDD address common threats to internal and external validity, as well as their limitations, are discussed.

Keywords Group-based designs · Roll-out designs · Small sample designs · Dynamic wait-listed design · Regression point displacement design

Group-based interventions, which have effects postulated to occur at both the individual and group level or through their interaction (Hutton 2001), introduce challenges to identifying for whom and in what contexts they are effective (Murray et al. 2004). When a group is assigned to receive an intervention based on some physical, social, or geographic connection, the units of observation are individuals nested within their groups and condition. Constraints in resources typically limit the number of groups that can implement or evaluate an intervention, making it difficult to distribute and account for potential confounding group differences (e.g., readiness) in group-based randomized trials (Murray, et al. 2004). The nested nature of a group-level intervention also reduces statistical power required for accurate testing of intervention effects, due to expected similarity or common variance accounted for by group membership (intraclass correlation). Due to the multi-level nature of group-based programs, statistical power is typically constrained by the number of groups, even if the number of individuals within groups is very large. Thus, when interventions are implemented in large social units such as entire communities or schools, it may be cost-prohibitive to include a sufficient number of units for adequate statistical power using traditional group-randomized trials.

Research designs that allow all community members to ultimately receive a group-based intervention are needed in many contexts. Decisions about implementing a group-based

P. A. Wyman (✉)
University of Rochester School of Medicine and Dentistry,
Rochester, NY, USA
e-mail: Peter_Wyman@URMC.Rochester.edu

D. Henry · S. Knoblauch
University of Illinois at Chicago, Chicago, IL, USA

C. H. Brown
Feinberg School of Medicine, Northwestern University
Chicago, IL, USA

intervention are often made by community representatives, and individual community members may have little or no opportunity to participate in decision-making or provide informed consent (Hutton 2001). Community decision-making may influence which research designs are feasible. If cultural norms forbid exclusion of some community members from activities open to others, such as is the case in some American Indian and Alaska Native communities, a group-based randomized trial involving a no-intervention group may be considered unethical or unacceptable. In other circumstances (e.g., multiple suicides), a community may decide to implement a program to everyone even if evidence regarding its efficacy is lacking (Brown et al. 2007). Another common scenario is when a community has already implemented, or began implementing, a group-based intervention and wishes to determine the extent to which prevention goals were met (Brown et al. 2014a).

Below we describe two designs that are well suited to evaluating group-based interventions when the number of social units is limited and all community members will receive a specified intervention. The choice of which design is appropriate illustrates the range of community contexts for evaluating group-based interventions. The dynamic wait-listed design (DWLD) is appropriate when prospective random assignment is possible. A variation of the traditional wait-listed design, the DWLD increases statistical power when the number of groups is limited and the rate of occurrence of the primary outcome is dependent on time. The regression point displacement design (RPDD) is an observational design suited to circumstances when prospective designs are not feasible, such as after a program has been implemented. Both designs make use of efficiencies that increase statistical power compared to alternative designs. For each design, we describe key features and a motivating example, including statistical analysis. We end with a discussion of the strengths and limitations of both designs and community contexts that influence decisions to include all participants in a selected intervention.

The Dynamic Wait-Listed Design

The dynamic wait-listed design (DWLD) is a randomized design that is useful for evaluating an intervention's efficacy or effectiveness as it is rolled out to a population or set of communities or groups such as all schools within a school district. If policy makers or community stakeholders have already decided that a novel intervention should be introduced to everyone in a population, a DWLD may well be an efficient way to conduct a rigorous randomized trial to test the effectiveness of the new intervention. Also, because it allocates equitably when individuals or groups—we use “units” as a generic term for this—are assigned to this intervention, and leaves no unit in a non-intervened control condition, policy

makers and community stakeholders may find this a more appropriate design than a traditional randomized trial or a traditional wait-listed design. The DWLD extends the traditional randomized wait-listed design by dividing the study period into multiple time periods during which subjects or groups are randomly assigned to receive the intervention (Brown et al. 2006). DWLDs first balance units into equivalent blocks and then randomly assign blocks of units to a time period for adopting the novel intervention. As we will see, DWLDs make use of all the period of adoption to evaluate effectiveness whereas wait-listed control designs only make use of half the time. This strategy increases efficiency and statistical power of the DWLD to assess intervention impact, especially when outcomes can be treated as count or time to event data.

In the classic wait-listed experiment, participation is divided into two phases. In the first phase, one half of participants are randomly assigned to receive the intervention and the other half to a control condition for a specified period of time. In a second phase, the control condition receives the intervention. A major limitation of the traditional wait-listed design is well known: only short-term intervention effects can be evaluated. Once intervention begins for the wait-listed group, the experiment is ended because no units remain in the control condition. Although the DWLD is also limited to evaluating short-term effects—because all units eventually enter the intervention condition—statistical power to detect differences in time to event data is increased by a longer total time for comparing units that have received the intervention against both their own response prior to adopting the novel intervention as well as against those units that still have not adopted the intervention yet. As we will see in an example, we can also use DWLDs to improve overall statistical modeling of short-term intervention impact compared to the traditional wait-listed design, such as determining if immediate intervention effects diminish or increase. Below we define formally the DWLD and compare it to other related designs.

As originally introduced, the DWLD can be used to evaluate the efficacy or effectiveness of a novel intervention to affect the rate of events that occur in a continuous process, such as the incidence of detecting and referring suicidal youth in secondary schools (Brown et al. 2006). The design begins with a complete set of units (e.g., all schools in a district), none with prior exposure to an intervention, and randomly assigning them to when they will receive this novel intervention. The assignment times could all be distinct, or units may be blocked first into comparable groupings, then all units within the block are randomized to receive the intervention at the same time. The individual units' rates on the outcome of interest (e.g., referrals of suicidal youth) are recorded throughout every interval of time, so by the end of the study nearly all units would have been measured prior to, during initial adoption, and after adoption. This initially proposed design could

thus be used to compare intervention impact across units and across time. If there are N units to be allocated and T time periods, with $B=N/T$ units in each block, then the first block is measured one time when initially adopting the intervention and $T-1$ times after adoption when the intervention is still continuing. The second block is measured one time without intervention and $T-1$ times when exposed to the intervention. The last block is measured $T-1$ times while unexposed to the intervention and once after adoption. If count data can be collected cheaply at each time point on each unit, say through a registry or through administrative records, this design uses repeated measures across units as well as between units to provide a very efficient means of assessing intervention impact (Brown et al. 2006).

Alternatively in a DWLD, we may start with a small number of units and “grow” a randomized trial over time by cumulating small numbers of randomized wait-listed studies. We call this a pairwise enrollment DWLD. Indeed, this “cumulative trial” idea where multiple smaller trials are combined over time (Brown et al. 2009) was, in fact, planned for a community-based HIV intervention study that provided the inspiration for DWLDs (Kegeles et al. 1996; Brown et al. 2006). Here is how this second version of a DWLD works. We would select two units at a time and randomly assign one to receive the intervention in the first time interval and the remaining one at the second interval. Data on a measurable outcome variable is then assessed or recorded during the first time interval when the units differ on intervention exposure. Because of random assignment to time, we can legitimately treat differences in response while one unit has received the intervention and the other has not started yet as causal to the short-term effects of the intervention condition. Typically, no further data would be recorded for these two units. We can then go on to the next pair of units, randomizing them again to whether they receive the intervention immediately or at a delayed time, and record differences in response in the duration where the pair has a discordant intervention condition. We continue to enroll pairs of units in this manner until sufficient units have been assessed to provide sufficient statistical power. We remark that it is possible to compress this cumulative design using what we call a single selection DWLD. In this design, we begin with a pair of units as before, randomizing one to receive the intervention at time 1 and the other at time 2. From a waiting list, we draw a third unit that serves as comparison for the second unit at time 2, and this unit then receives the intervention at time 3. Units continue to be selected from the waiting list one at a time. In this design, it is important to minimize drop out from the waiting list, otherwise there could be some systematic variation in units that occur over time. Because the pairwise enrollment and single selection DWLDs have much less longitudinal data on each unit, they often do not have the gains in statistical power that the DWLD has.

The first DWLD (Brown et al. 2006) is very similar to the stepped wedge design (Brown and Lilford 2006). They share the same motivation; i.e., evaluating a novel intervention when the community, organization, or policy maker has decided to adopt it throughout a system, even though it may not have been evaluated fully. They share the same general randomization schedule, but the stepped wedge designs are typically silent on the potential to enforce balancing units into comparable blocks over time. Also, descriptions of the stepped wedge design are silent about its potential use for individuals rather than groups, but it certainly could be used at the individual level. We take the view that these two designs can be treated as essentially comparable classes. Neither of these names, however, is well suited for use outside of academia for conveying to communities their important characteristics. We recommend that they be called roll-out designs (Brown et al. 2009) where (1) all units eventually receive the intervention, (2) the timing of units to receive the intervention is determined equitably (i.e., randomization), and (3) the design is used to evaluate effectiveness as an intervention is rolled out. Our experience is that when this design is described, communities are often comfortable with utilizing randomization to evaluate an intervention. There can be advantages for receiving a desired intervention first, and those who receive it later may be able to get a slightly improved intervention based on the experience to date in delivering the intervention (Brown et al. 2006). We also recommend beginning with a randomized roll-out design even when there are as few as two sites whose timing can be randomized. Even with a sample of size two, one can conduct analyses that look inside each site to examine how a novel program is being adopted, an existing program is being dismantled, or other changes are occurring. Secondly, the start of randomizing even two sites can lead to further pairs receiving random starts in complementary studies at a later time and therefore can legitimately be used in an analysis that synthesizes all findings.

A related type of design is called multiple baseline, which has a long history of use in behavioral sciences (Baer et al. 1968) especially with individual-level intervention. A non-concurrent multiple baseline design looks similar to a DWLD, and interrupted time series designs have been used in prevention (Biglan et al. 2000), but these designs do not use random assignment to determine the timing of the intervention (Carr 2005).

We also note that the class of roll-out designs is actually larger than the DWLD/stepped wedge design described above, both of which rely on one active intervention being newly introduced in communities. Another type of roll-out design can be used to compare two alternative interventions head-to-head. In such a design, all units begin in a wait-listed state, then at a random time they are assigned, again randomly, to one of two alternative interventions. This type of head-to-head randomized roll-out design was used to test two

implementation strategies for the evidence-based multidimensional treatment foster care (MTFC) in 51 counties in two states (Brown et al. 2014b; Chamberlain et al. 2010).

Statistical Modeling for DWLDs

We now return to the statistical modeling that has been proposed for the DWLD; this is somewhat more general than that proposed for stepped wedge designs (Hussey and Hughes 2007). The count of events for each unit during time interval t can be represented by Y_{gt} . At a randomly determined time τ_g , the g th group begins to receive the new intervention. The basic model for the counts Y_{gt} is based on a Poisson model, with the mean depending on a random effect of which unit they are in, a_g , a random effect for the time period b_t , plus time-independent covariates X_g , that may include stable population factors, as well as time-independent covariates Z_{gt} that include indicators of intervention condition, time trends, and changing population factors. In particular, the intervention covariate is coded as zero until the time that the intervention is introduced, and from then on is set to one. The overall model for Y_{gt} given X_g and Z_{gt} is often presented as a Poisson distribution with a mean μ_{gt} given by $e^{a_g+b_t+\alpha X_g+\beta Z_{gt}}$. The random effects a_g and b_t allow for group-level variations and variations in time of the events that are not accounted for by the fixed covariates. A standard approach for group-assigned trials is to use a fixed “offset” of the logarithm of the number in the population at risk times the interval of time, with coefficient fixed to 1. This allows us to interpret the intervention coefficient as the change in the per-person rate of reporting per unit time. In stepped wedge designs, the traditional models do not include random effects involving the time interval (Hussey and Hughes 2007). We have found it useful to include both fixed effects involving time, e.g., linear changes in the rate over time and seasonality, and residual random effects that may occur because of unpredictable, exogenous factors (e.g., increased attempts after a celebrity’s suicide). It is also easy to include time-dependent covariates involving exposure to the intervention condition, such as the proportion of the population trained as we discuss below.

Statistical Power

By balancing the time at which groups receive the intervention based on relevant baseline characteristics and including more than two time points for randomization, this dynamic wait-listed design can have substantial increased efficiency and statistical power over that of the standard wait-listed design. Thus, dynamic wait-listed designs are particularly useful when sample sizes are limited in group-based trials. Detailed power calculations may be found in Brown et al. (2006) and are only summarized here. First, there is a large gain in statistical power by increasing from 2 time points in a standard

wait-listed design to even a few more. Secondly, there is typically much more statistical power improvement in doubling the number of intervals from, say 2 to 4, than from say 16 to 32, if the overall time for an entire design to be completed is held fixed. The implication of this is that we often do not need a large number of time points to achieve sufficient gain in statistical power.

Example: DWLD to Enhance Efficiency in Suicide Prevention Research

Co-authors Brown and Wyman were invited by a large school district (Cobb County) in Georgia to partner in evaluating a suicide prevention training (Question, Persuade, Refer; Quinnett 1995) that the District leadership decided to provide all secondary school staff. QPR is a widely used gatekeeper training program designed to increase the adults’ ability to identify signs that students may be contemplating suicide, initiate a conversation and question them about suicidal thoughts, and, as needed, refer them for services. Gatekeeper training had been shown to increase adults’ knowledge and attitudes, but no prior study had evaluated impact on detecting of suicidal youth.

The Cobb County School District experienced a 6 % prevalence of self-reported suicide attempts in the past year reported by 8th and 10th graders, implying that approximately 3600 of the 60,000 middle and high school students could be expected to be harboring significant thoughts and/or plans about suicide; however, only 127 students were referred by the District that year for crisis evaluation, indicating that up to 95 % were not identified (Brown et al. 2006). In light of this small proportion of students referred for crisis evaluation, even a small increase in each staff member’s ability to identify and refer a potentially suicidal student could have a large effect in increasing detection of the population of suicidal students in the District.

Selection of Treatment Settings

Secondary school was selected as the unit for randomization and analysis, since each school had a counselor trained as a QPR Instructor and training was to be provided at the school level. Through a collaborative design process, the District agreed first to implement and evaluate QPR using a traditional randomized wait-listed design. Of the District’s 35 middle and high schools, three had already begun training in QPR and were excluded; the remaining 32 schools serving a total of 60,000 students were enrolled in 2003. After stratifying schools by middle/high school and previous year rates of crisis referrals, 16 schools were assigned to an early intervention condition to receive training in 2004 and the remaining 16 to a waiting condition for training in 2005. The primary outcome

was the rate of detection of suicidal students identified by their schools and referred for evaluation.

Time Unit Assignments

The pattern of staff training provided to the first group of 16 schools reflected the demands of a typical wait-listed design in which all training begins at the same time for the early intervention group. The first training occurred 40 days after the start of the trial. After the first 125 days of the trial, the District trained 1387 of 2498 staff (56 %) in the first 16 schools, and within each school nearly all training occurred within a few days. Moreover, schools that trained earlier in this first phase trained a higher proportion of their staff (75–90 %) compared to schools that trained later (<60 %). Delays in training and variable start-dates due to differential readiness have the potential to attenuate intervention effects, since typically the date of randomization or of first training would be the start of the trial and no differences between intervention and control groups would be expected until training is actually provided.

These delays in training that resulted from having to start QPR training simultaneously in 16 schools motivated switching to a DWLD for the second year (Brown et al. 2006), a change approved by the funders and data safety monitoring board. Rather than starting training simultaneously, the 16 schools initially assigned to the waiting condition were randomly assigned to four blocks, each to begin training at different times. In contrast to a standard wait-listed design in which the experiment would have ended as training began in the second group, shifting to a dynamic wait-listed design extended comparisons between intervention and control schools on rates of crisis evaluation referrals of suicidal students until the final group of schools began training.

Calculating Increased Efficiency

Brown et al. (2006) provided detailed analyses showing the increased power for estimating QPR impact using a DWLD versus a traditional wait-listed design. Whereas the standard design had 80 % power to detect a 32 % increase in student referral rates in the intervention condition versus controls, shifting to a dynamic wait-listed design with four time blocks had 80 % power to detect a 23 % increase in referral rates, an increase equivalent to adding six schools to the design. The DWLD nearly doubled the total time during which legitimate comparisons could be made between intervention and control groups, from approximately 1/2 to 7/8 of the 2-year study period.

Analysis

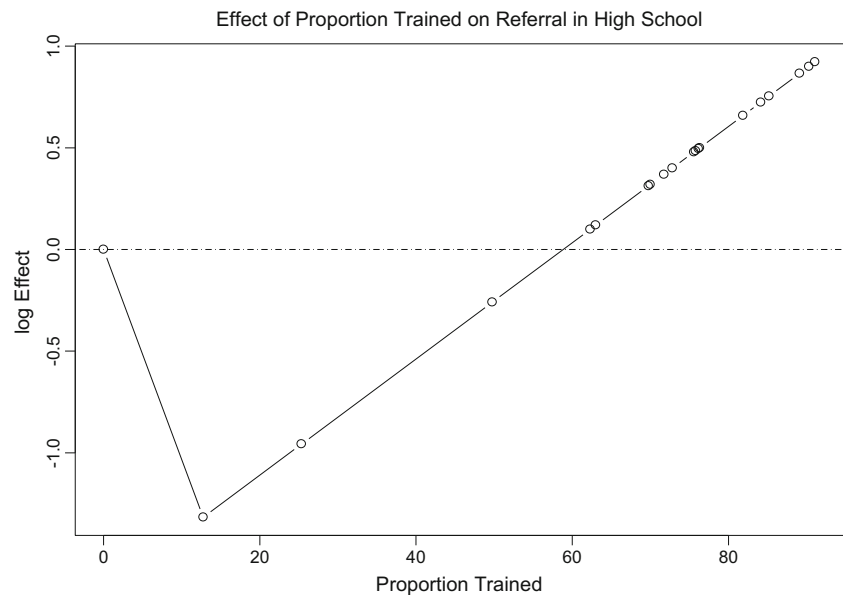
As an illustration of the type of complex modeling that is possible using the DWLD, we have examined the immediate

consequences of beginning QPR training in these schools. One question of interest was whether providing the first QPR training to staff in a school was associated with a large increase in referrals, followed by a much smaller rate of referral beyond this point. If such a pattern were found, it would suggest that staff were already aware of students who were suicidal but had little intention to refer or skills to talk to these youth. In these analyses, we included as time-dependent covariates an indicator of the interval in which initial QPR training began in that school, as well as one reflecting the proportion of the staff in that school that was trained. We also divided the time intervals more finely than the intervals that we had for training each block, starting a new time period whenever a new group-level training occurred in any of the schools, as the proportion trained in that school would change as of that time. Figure 1 shows the result of this analysis in the high schools. At the beginning of this plot, the effect of training was fixed to zero, standardizing the rates of referral across each school prior to training. As training began, the rate of referral actually decreases considerably, which was unexpected from what we had earlier predicted based on an expected surplus of youth known to be suicidal but not referred. Only after approximately 20 % of a school's staff received training did the rate of referral begin to climb with the proportion trained, and only after 60 % or more are trained do referral rates increase above the prior period. These findings suggest that training a few suicide prevention gatekeepers in a school will not overcome barriers that many staff experience in engaging distressed students (e.g., discomfort; fear of angering students) and may, in fact, heighten those barriers. However, staff referrals of suicidal students did increase after a majority of staff was trained, perhaps due to staff members perceiving a school-wide norm of collective responsibility has been established to engage distressed youth in conversations about possible suicidal intentions.

Other findings from staff surveys across the 32 schools showed that while QPR training consistently enhanced knowledge and attitudes, increased suicide identification behaviors were limited to staff already actively involved in conversations about distress with students before they were trained (Wyman et al. 2008). Combined with evidence that students at high risk for suicide (i.e., prior attempt) endorsed less favorable help-seeking attitudes, these Georgia Gatekeeper Project findings suggest that increasing youth–adult communication may be essential for gatekeeper trainings to identify more students at high risk for suicide. Final analyses of QPR training impact, using total crisis referrals and results from evaluations of students referred by community mental health providers, are in preparation (Brown, Wyman et al., in preparation).

Results from this trial of QPR training using a DWLD prompted the co-authors and Cobb County School District to select, in a next phase, a universal intervention (Sources of

Fig. 1 Effect of proportion trained in QPR on referrals in high school



Strength) that trains student peer opinion leaders to modify help-seeking norms throughout their friendship networks. Six high schools previously trained in QPR were selected, grouped into matched pairs, and randomly assigned to begin Sources of Strength training in either the fall or spring semester of the next year. Using a cumulative roll-out trial approach (Brown et al. 2009), another 12 high schools were added in the next 2 years, resulting in a trial with 18 schools serving 2600 students who had sufficient power to evaluate short-term program impact. Results showed that trained peer leaders modified school-wide risk and protective factors associated with lower risk for suicidal behavior (Wyman et al. 2010). In a next phase, we are extending this cumulative roll-out trial design to test Sources of Strength impact over 2 years on student suicide attempts in 40 high schools. Using this roll-out trial design that enrolls a new cohort of schools each year for four school years, it was possible to distribute training efficiently and to conduct social network assessments to test if school network changes (e.g., fewer isolated students; more ties to adults) serve to mediate the impact of this intervention on suicidal behavior.

Regression Point Displacement Design: Evaluation with as Little as One Intervention Unit

Preventive interventions that are specifically adapted to the needs of single communities present significant evaluation challenges including difficulties finding appropriate controls (Catalano et al. 1998; Henry et al. 2012). The regression point displacement design (RPDD; Linden, et al. 2006; Trochim and Campbell 1996) is a little-used variation on the regression discontinuity design (RDD; Campbell et al. 1963). The RPDD

shows promise for quasi-experimental evaluation of prevention programs conducted with a single intervention unit or a very small number of intervention units, when archival data for multiple units from the same population prior to and following implementation of the intervention are available. Furthermore, the RPDD can help increase the strength of causal inference for settings that may have already been planned, and potentially implemented, at a set point in time without the researcher's involvement as well as those settings that were non-randomly selected to receive an intervention due to practical limitations, ethics, or cultural norms.

The basic implementation of the RPDD is as follows. Archival data from units (such as communities or schools) that have not received intervention are used to create an expected posttest score for each unit receiving intervention. The actual posttest score of each unit receiving intervention is then compared to the expected posttest score. High correlations between pretest and posttest scores give the RPDD statistical power for assessing intervention impact when only a single or a few units have received intervention.

Design and Use Requirements of the Regression Point Displacement Design (RPDD)

Selection of Treatment Settings

The most important decision in applying the RPDD is the process used to choose the intervention group or groups. Control groups are chosen from the remainder of the population that have not selected or received intervention and on whom archival data are available. The more the selection process approximates random selection, the more valid will be the causal inference.

Unfortunately, in designs using the RPDD, the intervention unit(s) often will have been selected prior to the analysis, based on willingness or need. For example, the Communities that Care (CTC) program involves community stakeholders in a process of assessment and planning prior to implementation (Quinby et al. 2008), so well prior to the delivery of an intervention a CTC community is likely to differ from others that could be considered for comparison. Community-based participatory research studies such as those conducted by the Center for Alaska Native Health Research are initiated by community request due to concerns about levels of alcohol use and suicide, which distinguish them in important ways from communities that do not make such requests. Regardless of the reason for non-randomness in selection, it may be possible to combine propensity score methods (Rosenbaum and Rubin 1983) with RPDD, i.e., to first conduct analyses to predict selection of the treatment settings from available data.

The Unit of Analysis

Although the RPDD can theoretically be used with any unit of analysis, and its proponents (Trochim and Campbell 1996; Linden, et al. 2006) stress its flexibility, the unit of analysis chosen should have two characteristics. First, it should be sized to maximize the pretest–posttest correlation. Second, measures must be available before and after the initiation of an intervention, allowing estimation of the extent to which intervention units deviate from their expected values based on all similar units in the population.

Selection of Pretest and Posttest Measures

Community-level pretest and posttest measures should be selected to maximize the correlation between them, and thus, the accuracy with which an expected value for the intervention unit(s) can be estimated. For example, a tribal health corporation collects health record data on multiple small communities, only one of which has implemented a preventive intervention. If records, aggregated at the community level, can be obtained for periods of time prior to and following implementation of the intervention, the RPDD can be used.

It is not necessary for the pretest and posttest measures to be the same instruments, making the RPDD applicable in situations where there are historic changes in the indicators assessed, as, for example, might occur when new street drugs appear, changing the composition of measures of drug abuse. The statistical power of the design depends on the correlation between the pretest and posttest measures; the larger the correlation, the smaller will be the standard error for comparing actual to expected posttest measures. The pretest variable should also be selected to predate the intervention sufficiently so that displacement of one or more units from their expected

values will be possible. The RPDD does not model the effects of time or the interaction between time and intervention. Instead, values gathered from a time period after intervention are regressed on values from a time prior to intervention to provide expected values for the intervention unit(s).

Selection of Covariates

As in any regression model, the addition of covariates can sometimes improve the accuracy of estimation in RPDD models. As Linden et al. (2006) points out, covariates in an RPDD will be more effective if there are multiple treatment groups. Also, because the RPDD tends to be used with aggregates such as neighborhoods, police beats, cities, or states, the effect of adding covariates on degrees of freedom for error, and thus sample size requirements, should not be overlooked.

Analyzing the RPDD

Analysis of the RPDD requires a simple linear model. Each unit assigned to intervention receives a code of 1 and other units are coded 0 on an intervention indicator. If there are multiple units and little expectation of homogeneous effects among them, dummy codes may be created for each intervention unit, leaving the non-treatment units as the comparison level. Post-intervention scores are regressed on pre-intervention scores and the intervention dummy variable(s). Additional covariates may be added. Prior to beginning the analysis, the data should be examined to determine whether a linear regression equation is appropriate for modeling the relation between the pretest and posttest variables.

Type I Error Protection of the RPDD

To ascertain the extent to which the RPDD provides adequate protection from Type I error, we conducted a Monte Carlo study. Selecting one intervention unit out of a “population” of 25, we created random variables with pretest–posttest correlations varying between 0.75 and 0.99. We used the cumulative probability density of the pretest score as the probability of selection in order to approximate non-random selection of intervention units. We conducted 10,000 iterations of each correlation level, fitting the RPDD regression model following Linden et al. (2006). The proportion of significant results, even with intentionally biased selection, averaged 0.049 and ranged from 0.047 (with $r=0.99$) to 0.051 (with $r=0.75$). This suggests that the RPDD may afford adequate protection against Type I error, even when selection is biased in the direction of the pretest value. As long as biased selection is due to a measured baseline covariate in the model, the Type I error is quite stable. However, there may be other non-measured biasing factors in selection, and these may impact Type I error protection.

Power of the RPDD

Power for the RPDD is strongly related to the pretest–posttest correlation:

$$SE_{\text{pred}} = \sigma_Y \sqrt{1-r^2}$$

When the correlation is zero, the standard error of prediction is equal to the standard deviation of the outcome variable, and the smallest detectable effect may be quite large. However, as the pretest–posttest correlation increases, the detectable effect size grows smaller. At a correlation of 0.9, with a “population” of 25 and a single intervention unit, the RPDD provides power of 0.8 to detect a true standardized mean difference effect size of $d=0.5$. With a “population” of 100 units, one of which receives intervention, the detectable effect size would be $d=0.25$.

Example: Applying the RPDD in Violence Prevention Research

CeaseFire is a violence prevention intervention that involves placing workers whose mission is to stop violent altercations through outreach work with high-risk youth and families and violence interruption, which involves taking a direct role in mediating potentially violent conflicts (Dymnicki et al. 2013; Skogan et al. 2009). Based on a public health model, CeaseFire views the spread of violence as having similarities to the spread of infectious diseases. Implementation of CeaseFire intervention in neighborhoods depends on the vicissitudes of state funding as well as the status of political alliances and rivalries. In this particular case, a spike in homicide rates in 2010 prompted the Chicago Police Department to ask for CeaseFire assistance in reigning in the violence in two Chicago Police Districts.

The police districts selected for CeaseFire intervention had high homicide rates in 2010–2011, but other districts had homicide rates that were as high or higher. This is important because the presence of other districts with equal or higher rates in the population to be tapped for expected values protects the design against regression to the mean as a threat to validity. Regression to the mean is expected and modeled in the design. As long as the range of the non-intervention population covers the pretest values of the intervention units, regression to the mean would not be expected to a greater extent in intervention as compared to non-intervention units.

Selection of Units

The City of Chicago funded CeaseFire operations for four police beats, two each in two of the 25 police districts in

Chicago. As mentioned above, they chose two districts to focus on and then selected two beats from each district. We considered conducting analysis at the beat level, but rejected this for two reasons. First, we regarded it as unlikely that the operations of CeaseFire personnel, though concentrated in a single beat, would not affect surrounding beats. Including the surrounding beats in the population used to estimate expected values would result in contamination and an inability to fairly evaluate the effects of CeaseFire. The second reason had to do with the pretest–posttest correlation of homicide at the beat level, which was $r=0.38$, insufficient to provide statistical power for a test. The correlation of homicide at the district level over 3 years was $r=0.91$, much higher than the beat because the larger size of districts meant the rates were more stable. Our power simulations estimated that such a correlation would provide power of approximately 0.73 to detect an intervention effect of $d=0.4$ with 25 police districts, two of which received intervention.

Selection of Pretest and Posttest Measures

Because preventing homicides was the City’s overall aim, we chose counts of homicides as the outcome measure.

Selection of Covariates

We used logistic regression to calculate propensity scores representing the probability that a particular police beat would be selected for CeaseFire services. These analyses used crime data, poverty data, data on building violations, and an indicator of whether influential politicians requested CeaseFire services as predictors. The model explained approximately 37 % of the variability of selection for CeaseFire. The predicted probabilities of each beat being selected from the logistic regression were used as the propensity scores. Because analyses would be conducted at the district level, we aggregated propensity scores from the beat to the district level by taking the mean of all beats in each district. The second covariate was the total number of police responses in each district, entered in order to control for changes in police presence in the targeted districts.

Analysis

Using a generalized linear model with a Poisson distribution, we regressed homicide counts from the year of CeaseFire intervention in the 25 police districts on homicide counts from the year prior to negotiation of the city contract. The model used a Poisson distribution and a logarithmic link function:

$$\ln(E(Y/t)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

where Y/t is the posttest year homicide rate for each district, $E(Y/t)$ is the expected value under a Poisson model, X_1 is the

pretest value of the outcome, X_2 is a dummy variable indicating intervention District A, X_3 is a dummy code for intervention District B, both of which are CeaseFire districts, X_4 is the propensity score, and X_5 is the total number of police responses in the intervention year for each district. β_0 is an intercept, β_1 is the pretest–posttest slope where β_2 and $\beta_3=0$, and β_2 and β_3 are the effects of the intervention on Districts A and B, respectively.

Results were mixed, but informative. Both districts had fewer homicides than in the year prior to the contract, but differed from their expected numbers of homicides in different directions. District A had a higher than its expected number of homicides in the intervention year, and District B had a lower than its expected number of homicides in the intervention year. The difference in effects may be due to the tactics used. In District A, local gang leaders agreed that gang members would stay in their territory and not start anything and would alert CeaseFire workers to brewing violence. District B used more typical CeaseFire tactics, which relied on the “street intelligence” of outreach workers who were a consistent neighborhood presence.

Discussion

Small numbers of social units implementing group-based interventions, and the need to balance community values and research priorities, present challenges to prevention researchers. The dynamic wait-listed design (DWLD) and regression point displacement design (RPDD) address some of these challenges. Both designs utilize efficiencies that can increase statistical power and enable communities and researchers to provide the intervention to all members of units selected for intervention. In addition, both address many common threats to internal and external validity. By providing communities with options to evaluate and obtain useful information about interventions that are chosen to address a range of health problems, these designs can contribute to determining which interventions are effective and in which contexts. The DWLD and RPDD have important strengths as well as some limitations.

Strengths of the Designs

A key strength shared by both designs is the capacity to respond to community needs and sustain meaningful community involvement in research partnerships. In some contexts, such as American Indian and Alaska Native communities, the concept of using randomization to assign some individuals, but not others, for participation may be incompatible with cultural values. In other contexts, a community may have selected an intervention to address a pressing problem (e.g., youth suicide)

and perceive that intervention as essential to its public health goals. For example, in rural communities where suicide rates are substantially higher than the national average (Brown et al. 2007) and accessible, acceptable mental health and substance abuse services are less common, communities may select population-level interventions that have minimal empirical evidence, yet may be justified in deciding that the alternative of no intervention does not suit the best interests of its members. In such contexts where communities decide that all participants will receive an intervention, using the DWLD or RPDD can yield valuable information about intervention impact.

A strength of the DWLD is to increase the length of time intervention that can be compared with control by blocking more groups randomized to begin intervention at different times. Ultimately, all participating individuals or settings receive intervention. By increasing the number of subdivisions of the sample groups, the logistics of intervention implementation also become more manageable. Particularly in large, multi-site projects, the DWLD turns the logistical problem of beginning intervention at the same time in all units into a strength.

Because it relies on archival data, the RPDD compares social units that did and did not receive intervention. Handling control in this manner makes the RPDD well suited to situations where a single community receives intervention and archival data are available over a number of years for that community and others like it. In regards to RPDD, the statistical power depends primarily on the strength of correlation between pre- and posttest measures. This research design does not require that two measures be the same, but the stronger the correlation between measures, the greater the statistical power. Strong pretest–posttest correlations allow the RPDD to be effectively used for evaluating prevention with smaller samples (Brown and Liao 1999). Additionally, the RPDD’s ability to provide inferences to compare one unique site against all others may make it useful in policy evaluation. This type of design was used by Gibbons et al. (2007) with mixed effects modeling to assess racial disparities in foster placements in two counties implementing a change in child welfare policy.

Furthermore, these designs have a high degree of flexibility. The RPDD is analyzed using relatively simple linear models. However, if the data has a nonlinear pattern, polynomial terms can be added. This flexibility opens up options for the types of research designs that can accurately use this system of analysis. The DWLD has flexibility in a different way. Unlike typical wait-listed designs that spread finite training resources at a single time to large number of participants, DWLD has fewer participants in each group and fewer sites beginning at each time.

Internal and External Validity

The DWLD has features that improve internal validity over traditional wait-listed designs. Although there is more

opportunity for cross-site contamination in the DWLD than in traditional wait-listed designs (Wyman et al. 2010), shorter waiting periods make it less likely that participants will themselves seek out intervention content. Mortality or differential attrition may be less of a problem in the DWLD because of shorter waiting periods and the DWLD is more resilient to the internal validity threat of history. Instead of having two major groups that can experience different historical events, DWLD creates multiple groups making it unlikely that a major historical event will greatly influence all of the data or dampen the possibility of making effective comparisons. Finally, because it increases the length of time that a control group exists and uses randomization to determine when groups will receive treatment, the DWLD may be less susceptible to readiness to participate and other biases.

The internal validity of the RPDD depends on how the intervention settings and measures are selected. The closer the selection of settings is to random selection, the greater will be the internal validity. In the absence of random selection, using propensity scores can improve internal validity. The threat of regression is also addressed by selection of settings. Also critical is the timing of the pretest and posttest assessments, the former of which should be gathered at a time when contamination by intervention or the expectation of intervention is not possible.

With regard to external validity, the DWLD tends to be generalizable because there is a greater number of groups for which the intervention effect is being tested and replicated. Furthermore, by increasing the number of time intervals for which comparisons can be made, DWLD decreases the variance of the intervention effect, making estimations of intervention effects more efficient and accurate. On the other hand, the RPDD's external validity derives from its nonreactive assessments and random selection of intervention units from a population. Admittedly, it will often be employed in situations where treatment settings are not randomly selected. The solution to this problem is the use of propensity scores that control for non-randomness of selection using a rich set of covariates. Although propensity scores may not fully account for non-randomization, by making full use of the available information on important variables associated with bias within samples when random selection is not feasible, they allow meaningful and generalizable evaluation of intervention effects.

Limitations

Like the simple wait-listed design, DWLD can only evaluate time-limited effects. However, negotiated wait periods could be fairly long, as in our current roll-out design for testing peer leader suicide prevention training (Wyman et al. 2010), which uses a 2-year waiting period to accommodate the minimum period posited to change peer group norms. There are also

situations in which the DWLD should not be used, such as when implementation schedules cannot be varied (e.g., teacher training to implement an intervention must occur at the beginning of a school year). As in most quasi-experimental analyses, causal inference from the RPDD is weakened by the extent that non-random factors play a role in the selection of communities for participation or nonparticipation. To address this limitation, propensity scores, which are increasingly being recognized as a means to obtain stronger causal inference, are added into the RPDD as covariates. However, even with propensity scores, when used on a post hoc basis, there may be unknown factors that complicate inferences from the RPDD. As was noted above, power in the RPDD depends on the existence of archival data collected prior to and following the intervention period. In the absence of such data, or if highly correlated pretest and posttest measures cannot be obtained, the RPDD will have low power.

Conclusion

The DWLD and RPDD offer advantages for testing group-based interventions in a variety of community contexts when the number of units is limited. The DWLD also may be valuable in evaluating interventions for low base rate conditions, which introduce a different type of small sample problem. For example, no single randomized trial is likely to be sufficiently powered to evaluate impact on reducing mortality rates due to suicide (~10 per 100,000). However, if a state, or states, decides to roll out population-level interventions over time, a DWLD could be used to randomly assign large population units (e.g., counties across a state) to begin intervention at different time points to achieve the large person years of follow-up required to detect intervention effects on rare problems such as suicide. The RPDD is valuable in establishing evidence for interventions that require substantial site-specific modifications, which a single RCT cannot address, or when implementation has already occurred. In such instances, applying an intervention and using historical and ongoing surveillance data to evaluate can contribute to establishing effectiveness in new settings.

These two designs cannot fully overcome the limitations of small samples, but they do use available resources efficiently. DWLD is a flexible design that increases the total time of legitimate comparisons between intervention and control groups while ensuring that every participant receives the intervention in a manner that is usually easier to implement than a traditional wait-listed design. Likewise, the RPDD can be modified to fit the needs of the research that cannot utilize randomization, through the use of propensity scores, as well as the research model that is not a linear design, through the addition of terms to the statistical equation. Both of these designs add statistical power to the field of small sample

research, and researchers would be wise to add them to their analytical tool belts.

Acknowledgments We thank the National Institute of Mental Health for support under grants R34MH071189 (P. Wyman, PI) and RO1MH091452 (P. Wyman, PI) and the National Institute on Drug Abuse (NIDA) under grants P30 DA027828 (C. H. Brown, PI) and R13040610 (C. T. Fok, PI).

Conflict of Interest The authors declare that they have no conflict of interest.

References

- Baer, D. M., Wolf, M. M., & Risely, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91–97.
- Biglan, A., Ary, D., & Wagenaar, A. C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science, 1*, 31–49.
- Brown, C. H., & Liao, J. (1999). Principles for designing randomized preventive trials in mental health: an emerging developmental epidemiology paradigm. *American Journal of Community Psychology, 27*, 673–710.
- Brown, C. A., & Lilford, R. J. (2006). The stepped wedge design: A systematic review. *BMC Medical Research Methodology, 6*, 54. doi:10.1186/1471-2288-6-54.
- Brown, C. H., Wyman, P. A., Guo, J., & Pena, J. (2006). Dynamic wait-listed designs for randomized trials: New designs for prevention of youth suicide. [Research Support, N.I.H., Extramural Research Support, U.S. Gov't, P.H.S.]. *Clinical Trials, 3*, 259–271.
- 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. doi:10.1080/0954026070179777.
- Brown, C. H., Ten Have, T. R., Jo, B., Dagne, G., Wyman, P. A., Muthén, B. O., & Gibbons, R. (2009). Adaptive designs in public health. *Annual Review Public Health, 30*, 1–25.
- Brown, C.H., Mason, W.A., Brown, E.C. (2014). Translating the Intervention Approach into an Appropriate Research Design—The Next Generation Designs for Effectiveness and Implementation Research. In: Z Sloboda and H Petras (Eds.), *Advances in Prevention Science: Defining Prevention Science*. Springer Publishing.
- Brown, C.H., Chamberlain, P., Saldana, L., Padgett, C., Wang W., Cruden G. (2014). Evaluation of two implementation strategies in fifty-one child county public service systems in two states: Results of a cluster randomized head-to-head implementation trial.
- Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton, Mifflin and Company.
- Carr, J. E. (2005). Recommendations for reporting multiple-baseline designs across participants. *Behavioral Interventions, 20*, 219–224.
- Catalano, R. F., Arthur, M. W., Hawkins, J. D., Berglund, L., & Olson, J. J. (1998). Comprehensive community- and school-based interventions to prevent antisocial behavior. In R. Loeber & D. Farrington (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions*. Thousand Oaks: Sage.
- Chamberlain, P., Saldana, L., Brown, C. H., & Leve, L. (2010). Implementation of multidimensional treatment foster care in California: A randomized control trial of an evidence-based practice. In M. Roberts-DeGennaro & S. Fogel (Eds.), *Using evidence to inform practice for community and organizational change* (pp. 218–234). Chicago: Lyceum Books.
- Dymnicki, A., Henry, D., Quintana, E., Wisnieski, E., & Kane, C. (2013). Outreach worker perceptions of positive and negative critical incidents: Characteristics associated with successful and unsuccessful violence interruption. *Journal of Community Psychology, 41*, 200–217. doi:10.1002/jcop.21523.
- Gibbons, R. D., Hur, K., Bhaumik, D. K., & Bell, C. C. (2007). Profiling of county-level foster care placements using random-effects Poisson regression models. *Health Services and Outcomes Research Methodology, 7*, 97–108.
- Henry, D., Allen, J., Fok, C. C., Rasmus, S., Charles, B., & People Awakening Team. (2012). Patterns of protective factors in an intervention for the prevention of suicide and alcohol abuse with Yup'ik Alaska native youth (Early Online), 1–7. *The American Journal of Drug and Alcohol Abuse*. doi:10.3109/00952990.2012.704460.
- Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials, 28*, 182–191.
- Hutton, J. L. (2001). Are distinctive ethical principles required for cluster randomized controlled trials? *Statistics in Medicine, 20*, 473–488.
- Kegeles, S. M., Hays, R. B., & Coates, T. J. (1996). The Mpowerment Project: A community level HIV prevention intervention for young gay men. *American Journal of Public Health, 86*, 1129–1136.
- Linden, A., Trochim, W. M. K., & Adams, J. L. (2006). Evaluating program effectiveness using the regression point displacement design. *Evaluation and the Health Professions, 29*, 407–423.
- 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.
- Quinby, R. K., Hanson, K., Brooke-Weiss, B., Arthur, M. W., Hawkins, J. D., & Fagan, A. A. (2008). Installing the communities that care prevention system: implementation progress and fidelity in a randomized controlled trial. *Journal of Community Psychology, 36*, 313–332. doi:10.1002/jcop.20194.
- Quinnett, P. (1995). *QPR: Ask a question, save a life*. Spokane: QPR Institute and Suicide Awareness/Voices of Education.
- Rosenbaum, P. A., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.
- Skogan, W. G., Hartnett, S. M., Bump, N., & Dubois, J. (2009). *Evaluation of CeaseFire-Chicago*. Evanston: Northwestern University.
- Trochim, W. M. K., & Campbell, D. T. (1996). The regression point displacement design for evaluating community-based pilot programs and demonstration projects. Retrieved from <http://www.socialresearchmethods.net/research/RPD/RPD.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. doi:10.1037/0022-006X.76.1.104.
- 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. doi:10.2105/AJPH.2009.190025.