Published in final edited form as:

Prev Sci. 2015 October ; 16(7): 967–977. doi:10.1007/s11121-015-0569-4.

Research Designs for Intervention Research with Small Samples II: Stepped Wedge and Interrupted Time-Series Designs

Carlotta Ching Ting Fok,

University of Alaska Fairbanks

David Henry, and

University of Illinois at Chicago

James Allen

University of Minnesota Medical School, Duluth Campus

Abstract

The stepped wedge design (SWD) and the interrupted time-series design (ITSD) are two alternative research designs that maximize efficiency and statistical power with small samples when contrasted to the operating characteristics of conventional randomized controlled trials (RCT). This paper provides an overview and introduction to previous work with these designs, and compares and contrasts them with the dynamic wait-list design (DWLD) and the regression point displacement design (RPDD), which were presented in a previous article (Wyman, Henry, Knoblauch, and Brown, 2015) in this Special Section. The SWD and the DWLD are similar in that both are intervention implementation roll-out designs. We discuss similarities and differences between the SWD and DWLD in their historical origin and application, along with differences in the statistical modeling of each design. Next, we describe the main design characteristics of the ITSD, along with some of its strengths and limitations. We provide a critical comparative review of strengths and weaknesses in application of the ITSD, SWD, DWLD, and RPDD as small samples alternatives to application of the RCT, concluding with a discussion of the types of contextual factors that influence selection of an optimal research design by prevention researchers working with small samples.

Keywords

research design; small samples; stepped wedge design; interrupted time-series design

The randomized controlled trial (RCT) has long been held out as the gold standard for evaluating the effectiveness of health care treatments and procedures, including preventive interventions, although thoughtful RCT proponents readily acknowledge numerous potential sources of bias introduced in even its most stringent design applications (Kaptchuk, 2001). One of the ways bias can occur is when intervention scale and cost, or the population at risk being small or difficult to access results in a relatively small number of participants. Wyman, Henry, Knoblauch, and Brown (2015) also note that group-based interventions introduce new challenges as individuals are nested within the group, and analyses are at the group level. Using RCT in these situations can result in low power and low external validity. Randomization of groups such as a community to a no treatment condition can prove

unfeasible. Ethical complexities can arise, and community willingness to engage can be negatively impacted when withholding a promising treatment, as in the case of a no-treatment or treatment as usual control group. These and other challenges lead prevention scientists to consider alternative research designs.

In light of these limitations, Wyman et al. (2015) proposed the dynamic wait-listed design (DWLD) and regression point displacement design (RPDD) as two promising research design alternatives. These designs increase efficiency and statistical power in contrast to RCTs, a crucial consideration when number of participating individuals or groups is small. Maximizing power and efficiency is a primary strategy to increase the capability to draw inferences from analyses with small samples. The DWLD extends the traditional wait-listed controlled design by increasing the number of time periods in which individuals or groups are randomized to receive intervention. This increases efficiency and power by lengthening the total proportion of time between intervention and no intervention comparisons. The RPDD compares outcomes from intervention units to their expected values obtained from archival or other existing data on a large number of non-intervention units. When archival data exists on a large number of nonintervention groups pre and post the intervention of interest, RPDD can be a cost effective approach when a few settings or even only one setting receives intervention. Statistical power for the RPDD increases as magnitude of the pretest-posttest correlations increases. The approach can be particularly well suited to assessing the effects of community interventions.

This paper extends the work of Wyman and colleagues to identify alternative research designs that hold particular promise for small sample research, providing overview of a related design to the DWLD, the stepped wedge design (SWD), as well as the interrupted time-series design (ITSD). In the sections to follow, we overview the SWD, then discuss similarities and differences between the SWD and DWLD, including differences in their statistical modeling. Randomization of start times for intervention make it possible to directly compare randomized groups receiving and not receiving intervention in both the DWLD and SWD. We then examine the ITSD, a quasi-experimental design with strengths in assessing longitudinal effects through multiple assessments prior to and after intervention (Wagner, Soumerai, Zhang, & Ross-Degnan, 2002). Included in this discussion, we describe how these multiple assessments allow the ITSD to better control for specific threats to internal validity and substantially increase statistical power, an important consideration in small samples. Strength of inference regarding intervention effectiveness in all quasi-experimental designs is dependent on their implementation, and includes such factors as randomization of start times, number and time length of the nonintervention unit comparisons, and ecological characteristics of the setting, groups, and intervention. In particular, when randomization is not possible or ethical, resulting inferences are not as strong as in true experimental design, such as RCT. We conclude with a comparative review of strengths and weaknesses in the application of these alternative designs, and discuss contextual influences on design selection by prevention researchers working with small samples.

The Stepped Wedge Design (SWD)

The stepped wedge design (SWD) is a crossover design very similar to the DWLD. Wyman et al. (2015) describe both designs as “roll-out” designs because intervention is rolled out sequentially across groups, and discuss similarities in the groups randomization schedule of each and researcher motivations for their selection. Like DWLD, SWD is a longitudinal design in which each cluster or group receives both a baseline condition—essentially a wait-list control condition—and an intervention condition. The timing of crossover from baseline control to intervention condition is randomized. Figure 1 displays the basic structure of a rollout design with six groups. Crossovers to intervention conditions are unidirectional, with sequential roll-out of the intervention to groups or clusters over multiple time periods (Brown & Lilford, 2006; Hussey & Hughes, 2007). Multiple groups can receive intervention at a given time period, and different groups receive different doses. The effectiveness of intervention is assessed by comparing the outcome variable in the control section with the intervention section of the wedge. As both of these roll-out designs involve a unidirectional move from control to intervention, intervention is not withdrawn once implemented. This can alleviate ethical concerns that can accompany ending intervention once implemented (Hussey & Hughes, 2007).

Historical Origins and Applications

The SWD and DWLD are similar designs with different historical origins. While the DWLD originated in prevention research in educational settings (Brown, Wyman, Guo, & Pena, 2006), SWD origins were in medicine (Brown & Lilford, 2006). The DWLD was first introduced in suicide prevention research with youth in secondary schools (Brown et al., 2006), where intervention aimed to increase number of referrals of suicidal youth to mental health professionals. As the intervention was school-wide, random assignment occurred at the level of schools. Participating schools were randomly assigned to intervention at different time points, in a roll-out, wait-listed manner, such that all schools received intervention at the final time period.

The SWD was first initiated in a medical study in West Africa (Brown & Lilford, 2006) to examine the effectiveness of Hepatitis B vaccination (HBV) in preventing liver cancer and other chronic liver disease (The Gambia Hepatitis Study Group, 1987). Ethical concerns with withholding immunization made an RCT impossible, and logistical difficulties in contacting and scheduling individual immunization made individual randomization not possible. This led to group-wide intervention and use of the SWD, where HBV vaccination was progressively administered by immunization teams to groups of children in Gambia at about three-month intervals, such that after a four-year period national coverage was achieved.

Given these historical, disciplinary, and geographical differences, two variants emerged of a similar design. The DWLD found its niche in prevention studies primarily conducted in the United States (Brown et al., 2006; Brown, Wyman, Brinales, & Gibbons, 2007; Wyman et al., 2008; Wyman et al., 2010), while the SWD was primarily used in medicine, and in HIV studies in developing countries in particular (Brown & Lilford, 2006; The Gambia Hepatitis Study Group, 1987; Grant et al., 2005; Levy et al., 2004; Mdege, Man, Taylor Nee Brown,

& Torgerson, 2011). The primary differences between the designs lie mostly in application, with regards to blocking and research at the individual level. As the DWLD originated in educational contexts, balancing of units (in this case, schools) was often achieved by blocking prior to randomization, increasing the efficiency of the design. Work with the SWD, on the other hand, has made more limited use of these possibilities associated with blocking. Additionally, unlike DWD, work with SWD has not focused upon its potential application at the individual level.

Despite differences in origin, both the DWLD and SWD involve the same unidirectional, sequential intervention roll-out to clusters over different time periods that can be randomized. Accordingly, length of intervention for each group is different, with the last group to start intervention receiving intervention for only one time point. In settings with geographically remote or difficult to access populations, simultaneous implementation of intervention to all treatment group participants may not prove feasible. This feature maximizes efficiency when resources are limited, and researchers face logistics and cost-effectiveness considerations.

Statistical Modeling of the SWD

Statistical modeling of the SWD is a linear mixed-effects model (LMM; Hussey & Hughes, 2007) that includes fixed effects for time, and for intervention status at each particular time point. Random intercepts account for clustering of repeated observations within units, where units could be at either the individual or group level. A random effect for time may also be included when multiple units are assigned to the same time “step” or intervention “roll out”

Statistical Power

Power for the SWD is related to the number of time points constituting a step, and the number of participants or clusters randomized at each step. Hussey and Hughes (2007) empirically examined the effect of number of randomization steps on power. While number of clusters at each time point is a strong determinant of power, they found greater power was achieved when each cluster was randomized to its own time step, over the case when multiple clusters were randomized to shared time steps. In other words, optimal power was obtained when each cluster had its own randomization step. In interventions that have stronger effects with increased dosage, there is some loss in ability to show intervention impact given units randomized to start later receive less intervention dosage, though this deficit can be offset to some extent by increasing the number of measurement points.

SWD Case Example: Prevention of Infection by Antibiotic-resistant Bacteria in Long Term Acute Care Hospitals

Klebsiella pneumoniae carbapenemase-producing enterobacteriaceae (KPC) is an antibiotic-resistant bacteria strain declared an “immediate public health threat requiring urgent and aggressive action” by the Centers for Disease Control and Prevention (2013). Hayden et al. (2014) used a SWD to evaluate the effects of an intervention to prevent KPC infection. Four long-term care hospitals were randomly assigned start times for a four component, bundled intervention that consisted of patient screening, contact isolation of KPC-positive patients in ward cohorts or private rooms, bathing all patients daily with

chlorhexidine gluconate, and healthcare worker education and adherence monitoring. Start times were approximately every 2-months, thus at 7 months following initial roll-out, intervention was in effect at all four hospitals. Because of variability in availability of historical clinical data and the different dates of adoption of the intervention, the pre-intervention period at each hospital ranged from 16–29 months, and the intervention period from 12–19 months. Patients were screened for KPC within three days of admission, then re-screened every two weeks for the duration of the study.

The unit of randomization was hospital, and data on infections were aggregated by month for analysis. The statistical model was similar to that suggested by Hussey and Hughes (2007):

$$Y_{ij} = \beta_{00} + \beta_{01}(\text{time}) + \beta_{02}(\text{time} * \text{intervention}) + u_{0j} + r_{ij},$$

where u_{0j} is a random error term to account for the clustering of multiple observations within hospitals and β_{02} estimates the intervention effect. A main effect for intervention is not included because intervention status of each unit is completely nested within time.

In Figure 2, the solid straight line represents the linear trend of infection in the comparison phase, while the segmented straight line represents the linear trend in the intervention phase. The seven-month roll-out period occurred when some hospitals were in the baseline phase and others were in the intervention phase, and demarcates the pre- and post-intervention phases. Figure 2 displays an immediate reduction in infection rates with the introduction of the intervention, followed by a negative slope, whereas the trend prior to introduction of the intervention was increasing rate of infection. In response to intervention, KPC infection rates fell from 3.7 to 2.5 events per 1000 patient-days ($p < .001$).

The Interrupted Time-Series Design (ITSD)

A time series is a sequence of observations or values of a measure taken consecutively over a period of time. In an interrupted time series design (ITSD), multiple observations are assessed for a number of consecutive points in time before and after intervention within the same individual or group. Intervention is introduced at one or more time points, in which the time series is “interrupted,” that is, divided or segmented into two or more portions (Wagner et al., 2002). The intervention effect is assessed by comparing the pattern of change post-intervention to the pre-intervention pattern. As individuals or groups may serve as their own control, measurement at multiple pre- and post-intervention time points allows the separation of true intervention effects from other extraneous factors, such as threats associated with pre-existing differences across units or groups, and diffusion of intervention effects from treatment to control groups, thus reducing common threats to internal validity and increasing statistical power.

Design Considerations for the ITSD

Selection of treatment settings—ITSD is appropriate to a broad variety of applications wherein intervention effects are assessed by changes in growth patterns when individuals move from no-intervention to intervention or vice versa. In each of these potential

applications, the researcher has control over the start time of intervention. There may or may not be a comparison group in an ITSD. One potential application of ITSD is in policy research. Using archival data, the ITSD can examine outcome variables of interest before and after policy change across a large number of time points. The design is particularly useful when sample sizes are limited; as comparisons occur within the unit, the ITSD is feasible with a single group or even single individual. In addition, the ITSD allows for considerable design flexibility, and researchers can make pragmatic responses to emergent logistical challenges that can delay the course of work with difficult to reach populations. In settings with multiple cases or groups, the time at which the baseline and the treatment phases begin can be the same or can differ. Intervention initiation time can be staggered across cases or groups, such as in the multiple baseline design.

More complex ITSDs—In its simplest form, the ITSD is an A-B design, where multiple observations occur over the baseline phase (A) and the intervention, also described as the treatment phase (B). Intervention effects are assessed by comparing the series of observations prior to intervention to those following intervention. More complex ITSD designs involve more than one baseline phase. For example, in its simplest form, withdrawing the treatment sometime in the treatment phase is described as an A-B-A, or reversal or withdrawal design. In this design, the initial baseline phase (A) is followed by the treatment phase (B) and then subsequently, a second baseline phase (A) without treatment (Glass, 1997). More complicated reversal designs can involve extensions to this basic sequence, such as the ABAB or the ABABAB designs.

Number of observations in time series—As noted above, comparison of growth within each unit pre- to post-intervention reduces specific threats to internal validity and increases statistical power. Number of observations required for each time series segment depends on number of cases or groups, and variability of the response to intervention. Statistical power increases with increased number of observations. For a single case or group with a small number of observations, researchers may need to reduce variability at pre-intervention, or take into account different sources of this variability when making inferences regarding treatment effects.

Statistical Modeling of the ITSD

Segmented regression analysis is a powerful technique to analyze ITSDs (Wagner et al, 2002). As the pre- and post-intervention measurement points constitute separate segments of the time series, this technique provides the estimation of the level of y-intercept at the beginning of each segment, as well as the trend, or slope (rate of change) within each segment. Changes in level and trend of outcomes in the post-intervention compared to the pre-intervention segment assesses the extent to which change occurs as result of intervention. With a single subject or unit measured over multiple occasions, segmented regression can be represented as follows:

$$Y_t = \beta_0 + \beta_1(\text{time}) + \beta_2(\text{phase}[\text{pre- vs. post- intervention}]) + \beta_3(\text{time} * \text{phase}) + \varepsilon_t$$

where Y_{it} is estimated outcome for person i at time t ; $time$ is a continuous variable; $phase$ is a variable indicating before intervention (0) vs after intervention (1); $time*phase$ is the interaction between time and phase, and ε_t is a residual for time t . β_0 is an intercept; β_1 is the slope that estimates the rate of change for the outcome variable over time; β_2 is the change in level from pre to post-intervention; and β_3 is the interaction of the time by pre to post-intervention slopes measuring the effect of the intervention, which is the effect of interest in the analysis. If multiple units are involved, one or more random effects to account for clustering of measurement points within individuals may be added to the model.

ITSD Case Example: Prevention Policy Research in Youth Alcohol Use

Dumsha, DiTomasso, Gomez, Melucci, and Stouch (2011) examined the pattern of high school student drinking behavior following the introduction of “alcopops” to the U.S. market in 1999. Alcopops are sweetened, flavored alcoholic beverages that gained rapid popularity following market introduction, especially among teenage girls (Franson, 2002; Keane, 2003). Dumsha et al. used cross-sectional ITSD to conduct a secondary data analysis of Youth Risk Behavior Survey (YRBS) data collected on a total of 60,426 U.S. youth grades 9 to 12 by the Centers for Disease Control and Prevention (CDC). The CDC has conducted a biennial national YRBS survey to assess the health-risk behaviors of students since 1991. The outcome of interest was current (last 30 day) alcohol use on the YRBS administered biennially from 1997 to 2003.

Selection of time points—YRBS data were available from 1991 to 2013 for this re-analysis. As alcopops were launched in late 1999, we selected YRBS data for 1995, 1997, and 1999 to represent the period before alcopops introduction (pre-intervention), allowing three time points to estimate a linear slope. Data for 2001, 2003, and 2005 estimated the trend after introduction of alcopops (post-intervention).

Type of ITSD—Dumsha et al. and our replication both used an A-B design. The time period before alcopops entered the market constituted the baseline period (A) and the intervention or treatment phase (B) began with the launch of alcopops; intervention (introduction of alcopops) was not withdrawn after it was introduced. As the launch of the beverage occurred at the same time for all individuals throughout the U.S., baseline and intervention phases occurred at the same time for every person. In contrast to our replication across six time points, Dumsha et al. explored effects of the introduction of alcopops through three time points, using only one pre-intervention time point and two post-intervention time segments to estimate a slope.

Statistical modeling of the ITSD—The three-part segmented regression model used by Dumsha et al. can be represented as:

$$Y_t = \beta_0 + \beta_1(\text{pre-intervention time}) + \beta_2(\text{immediate post-intervention time}) + \beta_3(\text{long-term post-intervention time}) + \varepsilon_t$$

where Y_t is estimated alcohol use level at time t ; *pre-intervention time* is a continuous variable measuring time before intervention start (1997–1999); *immediate post-intervention* is a continuous variable indicating time immediately after alcopops introduction (1999–2001); *long-term post-intervention time* is a continuous variable measuring the long-term effects of alcopops (2001–2003); and ε_t is a residual for time t . β_0 is an intercept, β_1 is the slope that estimates the rate of change for the outcome variable before the intervention; β_2 is change in level immediately after alcopops launch, and β_3 is the slope measuring long-term alcopops effects.

Small sample re-analysis—For the purposes of illustration, we present here a small sample re-analysis using YRBS data. We created an aggregated case data set of current alcohol use. Each “case” represented one of the 47 possible combinations of gender, ethnic groups, and ages assessed in the YRBS data set. This resulted in a small sample data set of 47 “cases.” Each case included at least 20 individual observations used to create an aggregated score, assessed at 6 points in time (3 prior to the introduction of alcopops and 3 following introduction) analyzed according to the model specified above. The results we obtained using only these 47 cases were consistent with those obtained by Dumsha et al. using the entire sample.

Figure 3 shows the trend in past 30-day alcohol use exhibited a nonsignificant negative slope during the pre-alcopops or pre-intervention period, ($\beta = -.09$, $SE=.04$, ns), with a substantial drop from 1999 to 2001 ($\beta = -.61$, $SE=.61$, $p < .01$). After 2001, the post-alcopops or post-intervention period, the trend reversed, and this change (interruption) in measurements witnessed 2005 levels return to those of 1995 ($\beta = +.11$, $SE=.06$, $p = .05$). Use of a smaller sample likely led to reduced sensitivity to change and stability in the results, as the estimates obtained differ from Dumsha et al. However, study effects were sufficiently robust and the design sufficiently sensitive to produce comparable results regarding direction of change and statistical significance. Adjustment of some parameters in the design, for example, inclusion of a longer time span with a larger number of time points would add to the strength of inferences.

In summary, despite limiting the data set to only 47 “cases,” our small sample re-analysis had sufficient power to detect changes similar to those found in the Dumsha et al. large sample study. However, as in all uses of the ITSD, in interpreting casual effects in both the original study and our re-analysis, precise details regarding the specific implementation of the ITSD must be carefully considered; one threat to internal validity for both the current studies is that the individuals comprising the surveyed units were not the same over time.

Design Selection in Small Samples Research

Prevention scientists conducting health disparities research with small and often culturally distinct communities face numerous challenges. Among these challenges are (1) reduced statistical power related to small samples, (2) threats to internal and external validity linked to properties associated with a sample that is small, (3) concerns connected with health disparities research including the often pressing needs of a disparities community, and (4) ethical issues attendant to research in high need and high risk settings. This article and

Wyman et al. (2015) together describe the SWD, DWLD, ITSD, and RPDD as design alternatives to RCT. These alternative designs possess features that can address discrete elements of these challenges. Table 1 provides a summary for each of these four alternative designs describing the key attributes that define each design, their strengths and limitations, and factors that contribute to internal and external validity, and provide enhanced statistical power. Next we summarize these strengths and weaknesses, and conclude with certain contextual considerations prevention researchers often face in design selection with small samples.

Strengths

A major strength of all four alternative designs is every individual or group in a study receives intervention. Wyman et al. (2015) note prevention researchers' responsiveness to such concerns can be extremely important for maintaining partnerships in research programs where community involvement is crucial, as is often the case with health disparities and other disenfranchised groups. As one example, in health disparities communities the prevalence of a problem or disorder that is a prevention target is often unacceptably high. The potential for positive change from an intervention is often a primary reason communities might be willing to participate in a research study, especially in situations of past negative experiences with research. Thus not receiving intervention, or being placed on an extended waitlist control group until another community has completed the entire intervention, as often is the case in an RCT design, might prove unacceptable to community partners.

Another strength of the SWD and DWLD is their potential to address logistical difficulties that arise when prevention researchers must implement interventions in multiple remote or otherwise difficult to access communities. These designs turn these logistical challenges into a strength. By randomizing start times, a smaller staff can be sequenced across different implementation start times in different intervention settings/conditions.

Weaknesses

One potential limitation in use of the SWD and ITSD, also observed for the DWLD and RPDD by Wyman et al. (2015), is despite their potential to increase power in small samples analyses, their incorrect application can result in lower power than an RCT. This can occur whenever the unique design features that maximize power in these designs are not correctly understood and implemented, or if logistical or ethical realities preclude their correct implementation. Issues associated with potentially long wait periods, inadequate intervention dose in some conditions, and randomization can pose challenges to correct implementation.

For example, although the SWD and DWLD involve all research participants in the intervention, a potential weakness of these designs is some communities or groups only receive intervention during a few and, in at least one condition, a single time period; this can raise ethical concerns if the amount of intervention dose over the time period is potentially inadequate to have reasonable expectation of impact. An additional weakness in both the SWD and DWLD is the potential for long wait periods for intervention in some of the later

roll-out groups; this may prove unacceptable to communities, particularly in high need or high risk settings.

In cases of randomization of intervention start times in the SWD and DWLD, a host of logistical and ethical reasons can make randomization not possible. As one example, community or school leadership may change, and new leadership at one site may withdraw permission for the research part way through the study, removing one of the clusters from the randomized rollout protocol for intervention implementation. In another example, communities with high needs in the problem area the intervention seeks to address may be more motivated and mobilized, and more capable of starting immediately. Because of their greater need, it may also potentially be ethically questionable to withhold intervention to a later date as specified by the randomization protocol. In the case of a cluster of youth suicides in a school or community that has volunteered to participate in a suicide prevention intervention research study using a SWD or DWLD, their leadership may ask for immediate implementation of the intervention in response to a real world crisis and ongoing risk for additional suicide. It should be noted that many of these same logistical and potential ethical challenges associated with randomization in the SWD and DWLD can impose similar limitations on the applicability or implementability of the RCT. An additional weakness of all four designs, in contrast to the RCT, is they require numerous measurement points to achieve their heightened statistical power and internal validity with small samples. Depending on the nature, length, and complexity of assessments, this requirement may not be conceptually, logistically, or financially feasible, and can place a level of burden on research participants that may prove unacceptable.

Interaction of Contextual Factors with Research Design

Contextual factors associated with the intervention setting interact with design considerations to impact internal validity, external validity, and power. These types of contextual factors pose important considerations for prevention scientists.

Internal Validity

Internal validity refers to the extent to which a study is capable of establishing causality is related to the degree it minimizes error or bias. Although a strength of the ITSD is it permits finely gauged pre- and post-intervention comparisons within each unit or group, the design is also susceptible to the effects of instrumentation (Campbell, Stanley, & Gage, 1963). As a threat to internal validity, instrumentation effects refer to study participants basing current responses to measurement instruments upon their past experiences with the measures. An example of an instrumentation effect can occur in longitudinal studies of sensitive or illegal behaviors, where the first assessment can produce lower reported rates of the behavior than subsequent assessments (Campbell et al., 1963). This is often because as participants grow to trust the investigators, they become more forthcoming in reporting these behaviors. This effect can result in a prevention study finding an increase in such behaviors, and it becomes difficult to disentangle instrumentation effects from response to intervention. A related, but somewhat different threat can emerge in complex ITSD design variants, where treatment is introduced, then withdrawn, and then introduced again. Effects gained from a previous treatment may still be present and carried over in the next treatment; these are termed

carryover effects. Instrumentation and carryover effects can be a particularly problematic threat to internal validity for the ITSD, where there is no control group and when counterbalancing is not always possible.

However, in contrast to an RCT, internal validity may be enhanced when using an ITSD in a small, close-knit community. This is because in the ITSD, participants serve as their own controls. This eliminates the problem of diffusion, where the intervention effects inadvertently are spread from treatment to control groups. When diffusion is a concern, the ITSD may be a preferred design over an RCT, or an SWD or DWLD. There is also a reduced risk in an ITSD that the participating units, particularly those that might otherwise be assigned to a control group condition, will seek out intervention outside of the prevention research intervention.

Staggered start times in the SWD and DWLD do have the benefit of enhancing internal validity by reducing the influence of historical events. The closer the arrangement of intervention start times is to a random selection, the lower is the potential for historical events to pose a threat to interval validity. Therefore, the SWD and DWLD is a good design choice in settings where contextual factors allow for randomization of intervention start times for the units.

Another potential threat to internal validity in the SWD and DWLD is attrition related to long wait times for intervention in later implementation units, and in particular, potential for differential attrition rates dependent upon longer wait times. Differential attrition across groups in the SWD and DWLD can be reduced through efforts to shorten wait periods. Because of this, the SWD and DWLD is most promising as a design choice when contextual factors related to staffing adequacy as well as limited intervention complexity and/or time duration allow for short wait times across those groups that receive intervention later.

External Validity

External validity describes the extent to which a research conclusion can be generalized to the population or to other settings. The DWLD, SWD, ITSD, and RPDD, when applied to interventions conducted in entire communities, result in studies closer to the conditions that would apply if the interventions were conducted “at scale,” meaning implemented in the fashion they are in population level, large scale application. These designs also afford greater facility for intervention research in remote and difficult to reach communities, where the cost and logistical requirements of an RCT may become prohibitive, or may require alteration of the intervention or staffing support to levels that would never be feasible in real world application.

Statistical Power

A primary strategy to allow for causal inferences from small samples is to enhance the statistical power of the analysis through research designs that maximize efficiency in ability to make use of all available information in the data. The alternative designs described here possess this capability, though in different ways that interact differentially with features of the intervention context. For example, the SWD and DWLD extend the traditional wait-listed controlled design by increasing the time periods in which individuals or groups are

randomized. This increases efficiency and power by increasing the proportion in comparisons between intervention and control conditions. Any characteristics of the context that allow for increasing time intervals, as well as number of clusters, participants, clusters randomized at each time step, and participants within clusters further contributes to power in these designs. The ITSD uses this same logic, increasing the number of time periods through which the pattern of pre-intervention scores can be contrasted with the pattern of post-intervention scores. Characteristics of the intervention setting that allow for increasing the number of time periods of assessment argue for possible selection of the ITSD. The RPDD instead compares intervention units to their expected values. These expected values are obtained from archival or other existing data. Statistical power for the RPDD increases with the number of time periods of existing data along with the magnitude of the pretest-posttest correlations between outcomes. Settings for which data is available on large numbers of units prior to and following the time period of the intervention, and with outcomes that have high pretest-posttest correlations are most promising for the RPDD.

Interaction of Intervention Factors with Research Design

Additional considerations related to the intervention itself may also guide design selection. For example, the SWD and DWLD are most appropriate for interventions that groups or communities experience as highly needed, or for which questions exist regarding length and dose necessary for efficacy (Brown & Lilford, 2006; Gerritsen et al., 2011). Because of this, the SWD and DWLD are particularly beneficial in study of an intervention with demonstrated effectiveness applied to a new setting (Brown & Lilford, 2006). However, as wait time for some of the groups to start intervention in both the SWD and DWLD can prove lengthy, the integrity of the design could break down if certain groups or communities seek out other remedies and interventions on their own before the official start time of the intervention under study. Finally, in the case of interventions that ethically should not be withdrawn once implemented, the SWD and DWLD are most suitable. The ITSD may be used for interventions with a defined length, but in prevention studies of serious health risks, it is doubtful that complex ITSD designs with multiple bidirectional crossover points (e.g., ABAB designs) would ever be ethically justifiable.

Past researchers have found the ITSD to instead be extremely useful in prevention policy studies or in intervention implementation studies where the assignment of groups to a control condition is inappropriate (Pridemore & Snowden, 2009). Stallings-Smith, Zeka, Goodman, Kabir, & Clancy (2013) used the ITSD to examine the effectiveness of a smoking ban in reducing rate of heart disease and mortality. The ITSD can also be applied in evaluating the effects of a critical or catastrophic event on health outcomes; Pridemore, Rahan, and Chamlin (2009) evaluated the impact of the September 11 World Trade Center attacks and the Oklahoma City bombing on the rate of suicide. The ITSD is especially useful when there is a clearly identified time point of intervention or policy change, and less useful with outcomes that are more time continuous, and without such clear, critical time points.

Conclusions

In prevention research with small populations, culturally distinct groups, and community interventions, small samples are the norm, and conventional RCT designs may prove underpowered or impractical, and in some circumstances, can negatively impact partnership or even prove potentially unethical. The SWD and ITSD, along with the RPDD and DWLD, provide alternative designs. A number of contextual considerations regarding intervention setting and considerations regarding the intervention itself become important in guiding informed selection of one design over another. These designs provide additional methods for distilling evidence from small samples of critical importance to health promotion and prevention research.

Acknowledgements

The authors declare that they have no conflict of interest.

References

- Brown CA, Lilford RJ. The stepped wedge trial design: A systematic review. *BMC Medical Research Methodology*. 2006; 6:1–9. [PubMed: 16412232]
- Brown CH, Wyman PA, Guo J, Pena J. Dynamic wait-listed designs for randomized trials: New designs for prevention of youth suicide. *Clinical Trials*. 2006; 3:259–271. [PubMed: 16895043]
- Brown CH, Wyman PA, Brinales JM, Gibbons RD. The role of randomized trials in testing interventions for the prevention of youth suicide. *International Review of Psychiatry*. 2007; 19:617–631. [PubMed: 18092240]
- Campbell DT. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*. 1957; 54:297–312. [PubMed: 13465924]
- Campbell, DT.; Stanley, JC.; Gage, NL. *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton, Mifflin and Company; 1963.
- Centers for Disease Control and Prevention. Antibiotic resistance threats in the United States, 2013. 2013. <http://www.cdc.gov/features/antibioticresistancethreats/>
- Dumsha JZ, DiTomasso RA, Gomez FC, Melucci NJ, Stouch BC. Changes in self-reported drinking behaviors among US teenagers associated with the introduction of flavored malt beverages: An interrupted time series quasi-experiment. *Addiction Research and Theory*. 2011; 19:199–212.
- Fisher, RA. *The design of experiments*. Edinburgh, Scotland: Oliver & Boyd; 1935.
- Franson, P. Fast-growing malternatives threaten grape expectations. *Wine Business Online*. 2002. Retrieved from <http://www.winebusiness.com/SalesMarketing/Webarticle.cfm?AID=64002&IssueId=63975>
- Gerritsen DL, Smalbrugge M, Teerenstra S, Leontjevas Adang EM, Vernooij-Dassen MJ, Koopmans RTCM. Act In case of Depression: The evaluation of a care program to improve the detection and treatment of depression in nursing homes. *Study Protocol. BMC Psychiatry*. 2011; 11(91):1–7. [PubMed: 21194496]
- Glass, GV. Interrupted time-series quasi-experiments. In: Jaeger, RM., editor. *Complementary Methods for Research in Education*. Washington D. C: American Educational Research Association; 1997. p. 589–608.
- Golden MR, Whittington WLH, Handsfield HH, Hughes JP, Stamm WE, Hogben M, Holmes KK. Effect of expedited treatment of sex partners on recurrent or persistent gonorrhea or chlamydial infection. *New England Journal of Medicine*. 2005; 352:676–85. [PubMed: 15716561]
- Grant AD, Charalambous S, Fielding KL, Day JH, Corbett EL, Chaisson RE, Churchyard GJ. Effect of routine Isoniazid preventative therapy on tuberculosis incidence among HIV-infected men in South Africa: a novel randomized incremental recruitment study. *Journal of the American Medical Association*. 2005; 22:2719–2725. [PubMed: 15941800]

- Hawkins NG, Sanson-Fisher RW, Shakeshaft A, D'Este C, Green LW. The multiple baseline design for evaluating population-based research. *American Journal of Preventive Medicine*. 2007; 33:163–168.
- Hayden MK, Lin MK, Lolans K, Blom D, Weiner S, Lyles R, Weinstein RA. Prevention of colonization and infection by *Klebsiella Pneumoniae Carbapenemase-Producing Enterobacteriaceae* in long term acute care hospitals. Manuscript submitted for publication. 2014
- Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*. 2007; 28:182–191. Retrieved from <http://faculty.washington.edu/peterg/Vaccine2006/articles/HusseyHughes.2007.pdf>. [PubMed: 16829207]
- Kaptschuk J. The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology*. 2001; 54:541–549. [PubMed: 11377113]
- Keane, R. Malternative maximization. Adams Beverage Group. 2003. Retrieved from <http://www.beveragenet.net/bd/2003/0307/0307mlt.asp>
- Levy RW, Rayner CR, Fairley CK, Kong DCM, Mijch A, Costello K, McArthur C. Multidisciplinary HIV adherence intervention: A randomized study. *AIDS Patient Care and STDs*. 2004; 18:728–735. [PubMed: 15659884]
- Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology*. 2011; 64:936–948. [PubMed: 21411284]
- Pinheiro, JC.; Bates, MD. Mixed-effects models in S and S-PLUS. New York: Springer; 2004.
- Pridemore AWA, Trahan A, Chamlin MB. No evidence of suicide increase following terrorist attacks in the United States: An interrupted time-series analysis of September 11 and Oklahoma City. *Suicide and Life Threatening Behavior*. 2009; 39:659–670. [PubMed: 20121329]
- Pridemore AW, Snowden AJ. Reduction in suicide mortality following a new national alcohol policy in Slovenia: An interrupted time-series analysis. *American Journal of Public Health*. 2009; 99:915–920. [PubMed: 19299669]
- Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Application and data analysis methods. 2nd ed.. Thousand Oaks, CA: Sage Publications; 2002.
- Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974; 66:688–701.
- Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of American Statistical Association*. 2005; 100:322–331.
- Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton-Mifflin; 2002.
- Stallings-Smith S, Zeka A, Goodman P, Kabir Z, Clancy L. Reductions in cardiovascular, cerebrovascular, and respiratory mortality following the National Irish Smoking Ban: Interrupted time-series analysis. *PLOSOne*. 2013; 8(4):1–7.
- The Gambia Hepatitis Study Group. The Gambia hepatitis intervention study. *Cancer Research*. 1987; 47:5782–5787. [PubMed: 2822233]
- Wagner AK, Soumerai SB, Zhang F, Ross Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*. 2002; 27:299–309. [PubMed: 12174032]
- West SG, Duan N, Pequegnat W, Gaist P, Des Jarlais DC, Holtgrave D, Mullen PD. Alternatives to the randomized controlled trial. *American Journal of Public Health*. 2008; 98:1359–1366. [PubMed: 18556609]
- Woertman W, Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology*. 2013; 66:752–758. [PubMed: 23523551]
- Wyman PA, Henry D, Knoblauch S, Brown CH. Designs for testing group-based interventions with limited numbers of social units: The dynamic wait-listed and regression point displacement designs. *Prevention Science*. 2015

- Wyman PA, Brown CH, Inman J, Cross W, Schmeelk-Cone K, Guo J, Pena JB. Randomized trial of a gatekeeper program for suicide prevention: 1-year impact on secondary school staff. *Journal of Consulting and Clinical Psychology*. 2008; 76(1):104–115. [PubMed: 18229988]
- Wyman PA, Brown CH, LoMurray M, Schmeelk-Cone K, Petrova M, Yu Q, Wang W. An outcome evaluation of the Sources of Strength suicide prevention program delivered by adolescent peer leaders in high schools. *American Journal of Public Health*. 2010; 100(9):1653–1661. [PubMed: 20634440]

Figure 1.
Example of a six-group stepped wedge design (SWD). X = intervention condition; 0 = control condition.

Figure 2. Stepped Wedge Design used by Hayden et al. (2014) to evaluate the effects of an intervention to prevent *klebsiella pneumoniae carbapenemase-producing enterobacteriaceae* (KPC) colonization in four long-term care hospitals. Solid straight line estimates the baseline and comparison trend line, and the segmented straight line estimates the trend after initiation of the intervention. Used with permission.

Figure 3.

Re-analysis of the aggregated alcopop data set displaying past 30 day alcohol use reported on the Youth Risk Behavior Survey (YRBS). A total of 47 “cases” were assessed at 6 points in time (1995, 1997, 1999, 2001, 2003, and 2005). Year on the x-axis is centered at 2000, around the time of the alcopops (flavored malt beverages) marketing launch.

Table A1

Comparison of Stepped Wedge, Dynamic Wait-Listed, Interrupted Time Series, and Regression Point Displacement Designs

Design	Brief Description	Strengths	Limitations	Internal and External Validity	Power
Stepped-wedge	<ul style="list-style-type: none"> - A type of crossover design where time of crossover is randomized - Crossover is unidirectional, typically from control to intervention. - Important similarities to the dynamic wait-listed design, which evolved out of different disciplines using somewhat different analytic approaches 	<ul style="list-style-type: none"> - Easier logistically, especially for geographically dispersed areas - All participants receive intervention, though some receive only small dose 	<ul style="list-style-type: none"> - Long wait times for some groups - Dose is different across groups - Potential partnership and ethical issues as some groups must wait for a long time for intervention and receive intervention at fewer time points (receive less dose) - Requires capacity to deliver intervention in all settings during final time period - Potential confounds with developmental effects 	<ul style="list-style-type: none"> - The use of multiple time intervals reduces measurement error - Increases feasibility for researchers to work in different communities that are difficult to access because start times are not all at once - Randomization occurs both at level of time and dose 	<ul style="list-style-type: none"> - Increased power through multiple time intervals - Optimal power is achieved when randomization occurs at each cluster
Dynamic wait-listed	<ul style="list-style-type: none"> - Extension of the traditional wait-listed design - The study period is divided into a larger number of time units in which intervention is implemented 	<ul style="list-style-type: none"> - Increased length of time intervention could be compared with control - Logistics of intervention become more manageable - All participants receive intervention 	<ul style="list-style-type: none"> - The design should not be used in situations where schedules for intervention implementation cannot be varied 	<ul style="list-style-type: none"> - Randomization at level of time provides improvements in internal validity over traditional wait-listed design - Shorter waiting periods reduces threat due to differential attrition - Greater number of groups reduces influence of historical events and increases generalizability of intervention effect 	<ul style="list-style-type: none"> - Increased power through more time blocks that intervention can be compared with control condition
Interrupted time-series	Multiple observations are assessed for a number of consecutive points in time before and after intervention within the same individual or group	Feasible for small number of communities or groups, as each group acts as their own control	<ul style="list-style-type: none"> - Requires a large number of measurement points, high cost, potentially high participant burden, potentially not financially or logistically feasible for 	<ul style="list-style-type: none"> - Comparison of pre- and post-intervention within each unit reduces threats to internal validity 	<ul style="list-style-type: none"> - Power increases with the increase in the number of observations

Design	Brief Description	Strengths	Limitations	Internal and External Validity	Power
Regression point displacement	-	The pre-post results of a single or multiple treatment groups are compared to a large sample of individuals, groups, or communities	-	geographically dispersed areas geographically dispersed areas Intervention cessation within case is not always ethical With many effective interventions, effects persist over time, not allowing for A-B-A-B designs Potential confounds with developmental effects	-
		Ease of implementation, can implement intervention with as few as one (or more) groups	-	Low power with small sample, incorrect estimates if data are nonlinear	Depends upon how intervention settings are selected and how pretest and posttest measures are gathered
		Well suited to situations where a single community receives intervention and archival data are available	-	Not everyone receives intervention, only one or more groups does	Increased internal validity if strongly correlated pretest and posttest measures are chosen
		High degree of flexibility in the analysis	-	Requires similar data across a larger set of individuals or groups that do not receive intervention	Increased external validity with random selection of intervention units
					Closely approximates a naturally occurring experiment with potential high ecological validity
					Increased power through large number of comparison communities or groups