

UCSF

UC San Francisco Previously Published Works

Title

Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research

Permalink

<https://escholarship.org/uc/item/52b4j4w8>

Journal

Annual Review of Public Health, 39(1)

ISSN

0163-7525

Authors

Handley, Margaret A

Lyles, Courtney R

McCulloch, Charles

et al.

Publication Date

2024-04-01

DOI

10.1146/annurev-publhealth-040617-014128

Peer reviewed



Published in final edited form as:

Annu Rev Public Health. 2018 April 01; 39: 5–25. doi:10.1146/annurev-publhealth-040617-014128.

Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research

Margaret A. Handley^{1,2}, Courtney Lyles², Charles McCulloch¹, Adithya Cattamanchi³

¹Department of Epidemiology and Biostatistics, Division of Infectious Disease Epidemiology, University of California, San Francisco, CA

²General Internal Medicine and UCSF Center for Vulnerable Populations, San Francisco Zuckerberg General Hospital and Trauma Center, University of California, San Francisco, CA, 1001 Potrero Avenue, Box 1364, San Francisco, CA 94110

³Division of Pulmonary and Critical Care Medicine and UCSF Center for Vulnerable Populations, San Francisco Zuckerberg General Hospital and Trauma Center, University of California, San Francisco, CA, 1001 Potrero Avenue, San Francisco, CA 94110

Abstract

Interventional researchers face many design challenges when assessing intervention implementation in real-world settings. Intervention implementation requires ‘holding fast’ on internal validity needs while incorporating external validity considerations (such as uptake by diverse sub-populations, acceptability, cost, sustainability). Quasi-experimental designs (QEDs) are increasingly employed to achieve a better balance between internal and external validity. Although these designs are often referred to and summarized in terms of logistical benefits versus threats to internal validity, there is still uncertainty about: (1) how to select from among various QEDs, and (2) strategies to strengthen their internal and external validity. We focus on commonly used QEDs (pre-post designs with non-equivalent control groups, interrupted time series, and stepped wedge designs) and discuss several variants that maximize internal and external validity at the design, execution, and analysis stages.

Keywords

Quasi-experimental design; stepped wedge; interrupted time series; pre-post; implementation science; external validity

INTRODUCTION

Public health practice involves implementation or adaptation of evidence-based interventions into new settings in order to improve health for individuals and populations. Such interventions typically include on one or more of the “7 Ps” (programs, practices, principles, procedures, products, pills, and policies) (9). Increasingly, both public health and clinical research have sought to generate practice-based evidence on a wide range of interventions,

which in turn has led to a greater focus on intervention research designs that can be applied in real-world settings (2, 8, 9, 20, 25, 26, 10, 2).

Randomized controlled trials (RCTs) in which individuals are assigned to intervention or control (standard-of-care or placebo) arms are considered the gold standard for assessing causality and as such are a first choice for most intervention research. Random allocation minimizes selection bias and maximizes the likelihood that measured and unmeasured confounding variables are distributed equally, enabling any difference in outcomes between intervention and control arms to be attributed to the intervention under study. RCTs can also involve random assignment of groups (e.g., clinics, worksites or communities) to intervention and control arms, but a large number of groups are required in order to realize the full benefits of randomization. Traditional RCTs strongly prioritize internal validity over external validity by employing strict eligibility criteria and rigorous data collection methods.

Alternative research methods are needed to test interventions for their effectiveness in many real-world settings—and later when evidence-based interventions are known, for spreading or scaling up these interventions to new settings and populations (23,40). In real-world settings, random allocation of the intervention may not be possible or fully under the control of investigators because of practical, ethical, social, or logistical constraints. For example, when partnering with communities or organizations to deliver a public health intervention, it might not be acceptable that only half of individuals or sites receive an intervention. As well, the timing of intervention roll-out might be determined by an external process outside the control of the investigator, such as a mandated policy. Also, when self-selected groups are expected to participate in a program as part of routine care, there would arise ethical concerns associated with random assignment – for example, the withholding or delaying of a potentially effective treatment or the provision of a less effective treatment for one group of participants (49). As described by Peters et al “implementation research seeks to understand and work within real world conditions, rather than trying to control for these conditions or to remove their influence as causal effects. “ (40). For all of these reasons, a blending of the design components of clinical effectiveness trials and implementation research is feasible and desirable, and this review covers both. Such blending of effectiveness and implementation components within a study can provide benefits beyond either research approach alone (14), for example by leading to faster uptake of interventions by simultaneously testing implementation strategies.

Since assessment of intervention effectiveness and implementation in real-world settings requires increased focus on external validity (including consideration of factors enhancing intervention uptake by diverse sub-populations, acceptability to a wide range of stakeholders, cost, and sustainability) (34), interventional research designs are needed that are more relevant to the potential, ‘hoped for’ treatment population than a RCT, and that achieve a better balance between internal and external validity. Quasi-experimental designs (QEDs), which first gained prominence in social science research (11), are increasingly being employed to fill this need. [BOX 1 HERE: Definitions used in this review].

QEDs test causal hypotheses but, in lieu of fully randomized assignment of the intervention, seek to define a comparison group or time period that reflects the counter-factual (*i.e.*,

outcomes if the intervention had not been implemented) (43). QEDs seek to identify a comparison group or time period that is as similar as possible to the treatment group or time period in terms of baseline (pre-intervention) characteristics. QEDs can include partial randomization such as in stepped wedge designs (SWD) when there is pre-determined (and non-random) stratification of sites, but the order in which sites within each strata receive the intervention is assigned randomly. For example, strata that are determined by size or perceived ease of implementation may be assigned to receive the intervention first. However, within those strata the specific sites themselves are randomly selected to receive the intervention across the time intervals included in the study). In all cases, the key threat to internal validity of QEDs is a lack of similarity between the comparison and intervention groups or time periods due to differences in characteristics of the people, sites, or time periods involved.

Previous reviews in this journal have focused on the importance and use of QEDs and other methods to enhance causal inference when evaluating the impact of an intervention that has already been implemented (4,8,9,18). Design approaches in this case often include creating a post-hoc comparison group for a natural experiment or identifying pre and post-intervention data to then conduct an interrupted time series study. Analysis phase approaches often utilize techniques such as pre-post, regression adjustment, scores, difference-in-differences, synthetic controls, interrupted time series, regression discontinuity, and instrumental variables (4,9,18). Although these articles summarize key components of QEDs (e.g. interrupted time series), as well as analysis-focused strategies (regression adjustment, propensity scores, difference-in-differences, synthetic controls, and instrumental variables) there is still uncertainty about: (1) how to select from among various QEDs in the pre-implementation design phase, and (2) strategies to strengthen internal and external validity before and during the implementation phase.

In this paper we discuss the a priori choice of a QED when evaluating the impact of an intervention or policy for which the investigator has some element of design control related to 1) order of intervention allocation (including random and non-random approaches); 2) selecting sites or individuals; and/or 3) timing and frequency of data collection. In the next section, we discuss the main QEDs used for prospective evaluations of interventions in real-world settings and their advantages and disadvantages with respect to addressing threats to internal validity [BOX 2 HERE Common Threats to Internal Validity of Quasi-Experimental Designs Evaluating Interventions in ‘Real World’ Settings]. Following this summary, we discuss opportunities to strengthen their internal validity, illustrated with examples from the literature. Then we propose a decision framework for key decision points that lead to different QED options. We conclude with a brief discussion of incorporating additional design elements to capture the full range of relevant implementation outcomes in order to maximize external validity.

QUASI-EXPERIMENTAL DESIGNS FOR PROSPECTIVE EVALUATION OF INTERVENTIONS

Table 1 summarizes the main QEDs that have been used for prospective evaluation of health intervention in real-world settings; pre-post designs with a non-equivalent control group, interrupted time series and stepped wedge designs. We do not include pre-post designs without a control group in this review, as in general, QEDs are primarily those designs that identify a comparison group or time period that is as similar as possible to the treatment group or time period in terms of baseline (pre-intervention) characteristics (50). Below, we describe features of each QED, considering strengths and limitations and providing examples of their use.

1. Pre-Post With Non-Equivalent Control Group

The first type of QED highlighted in this review is perhaps the most straightforward type of intervention design: the pre-post comparison study with a non-equivalent control group. In this design, the intervention is introduced at a single point in time to one or more sites, for which there is also a pre-test and post-test evaluation period. The pre-post differences between these two sites is then compared. In practice, interventions using this design are often delivered at a higher level, such as to entire communities or organizations¹ [Figure 1 here]. In this design the investigators identify additional site(s) that are similar to the intervention site to serve as a comparison/control group. However, these control sites are different in some way than the intervention site(s) and thus the term “non-equivalent” is important, and clarifies that there are inherent differences in the treatment and control groups (15).

The strengths of pre-post designs are mainly based in their simplicity, such as data collection is usually only at a few points (although sometimes more). However, pre-post designs can be affected by several of the threats to internal validity of QEDs presented here. The largest challenges are related to 1) ‘history bias’ in which events unrelated to the intervention occur (also referred to as secular trends) before or during the intervention period and have an effect on the outcome (either positive or negative) that are not related to the intervention (39); and 2) differences between the intervention and control sites because the non-equivalent control groups are likely to differ from the intervention sites in a number of meaningful ways that impact the outcome of interest and can bias results (selection bias).

At this design stage, the first step at improving internal validity would be focused on selection of a non-equivalent control group(s) for which some balance in the distribution of known risk factors is established. This can be challenging as there may not be adequate information available to determine how ‘equivalent’ the comparison group is regarding relevant covariates.

It can be useful to obtain pre-test data or baseline characteristics to improve the comparability of the two groups. In the most controlled situations within this design, the

¹It is important to note that if such randomization would be possible at the site level based on similar sites, a cluster randomized control trial would be an option.

investigators might include elements of randomization or matching for individuals in the intervention or comparison site, to attempt to balance the covariate distribution. Implicit in this approach is the assumption that the greater the similarity between groups, the smaller the likelihood that confounding will threaten inferences of causality of effect for the intervention (33, 47). Thus, it is important to select this group or multiple groups with as much specificity as possible.

In order to enhance the causal inference for pre-post designs with non-equivalent control groups, the best strategies improve the comparability of the control group with regards to potential covariates related to the outcome of interest but are not under investigation. One strategy involves creating a cohort, and then using targeted sampling to inform matching of individuals within the cohort. Matching can be based on demographic and other important factors (e.g. measures of health care access or time-period). This design in essence creates a matched, nested case-control design.

Collection of additional data once sites are selected cannot in itself reduce bias, but can inform the examination of the association of interest, and provide data supporting interpretation consistent with the reduced likelihood of bias. These data collection strategies include: 1) extra data collection points at additional pre- or post- time points (to get closer to an interrupted time series design in effect and examine potential threats of maturation and history bias), and 2) collection of data on other dependent variables with a priori assessment of how they will 'react' with time dependent variables. A detailed analysis can then provide information on the potential affects on the outcome of interest (to understand potential underlying threats due to history bias).

Additionally, there are analytic strategies that can improve the interpretation of this design, such as: 1) analysis for multiple non-equivalent control groups, to determine if the intervention effects are robust across different conditions or settings (.e.g. using sensitivity analysis), 2) examination within a smaller critical window of the study in which the intervention would be plausibly expected to make the most impact, and 3) identification of subgroups of individuals within the intervention community who are known to have received high vs. low exposure to the intervention, to be able to investigate a potential "dose-response" effect. Table 2 provides examples of studies using the pre-post non-equivalent control group designs that have employed one or more of these improvement approaches to improve the internal study's validity.

Cousins et al utilized a non-equivalent control selection strategy to leverage a recent cross-sectional survey among six universities in New Zealand regarding drinking among college-age students (16). In the original survey, there were six sites, and for the control group, five were selected to provide non-equivalent control group data for the one intervention campus. The campus intervention targeted young adult drinking-related problems and other outcomes, such as aggressive behavior, using an environmental intervention with a community liaison and a campus security program (also know as a Campus Watch program). The original cross-sectional survey was administered nationally to students using a web-based format, and was repeated in the years soon after the Campus Watch intervention was implemented in one site. Benefits of the design include: a consistent sampling frame at each

control sites, such that sites could be combined as well as evaluated separately and collection of additional data on alcohol sales and consumption over the study period, to support inference. In a study by Wertz et al (48), a non-equivalent control group was created using matching for those who were eligible for a health coaching program and opted out of the program (to be compared with those who opted in) among insured patients with diabetes and/or hypertension. Matching was based on propensity scores among those patients using demographic and socioeconomic factors and medical center location and a longitudinal cohort was created prior to the intervention (see Basu et al 2017 for more on this approach).

In the pre-post malaria-prevention intervention example from Gambia, the investigators were studying the introduction of bed nets treated with insecticide on malaria rates in Gambia, and collected additional data to evaluate the internal validity assumptions within their design (1). In this study, the investigators introduced bed nets at the village level, using communities not receiving the bed nets as control sites. To strengthen the internal validity they collected additional data that enabled them to: 1) determine whether the reduction in malaria rates were most pronounced during the rainy season within the intervention communities, as this was a biologically plausible exposure period in which they could expect the largest effect size difference between intervention and control sites, and 2) examine use patterns for the bed nets, based on how much insecticide was present in the bed nets over time (after regular washing occurred), which aided in calculating a “dose-response” effect of exposure to the bed net among a subsample of individuals in the intervention community.

2. Interrupted Time Series

An interrupted time series (ITS) design involves collection of outcome data at multiple time points before and after an intervention is introduced at a given point in time at one or more sites (6, 13). The pre-intervention outcome data is used to establish an underlying trend that is assumed to continue unchanged in the absence of the intervention under study (*i.e.*, the counterfactual scenario). Any change in outcome level or trend from the counter-factual scenario in the post-intervention period is then attributed to the impact of the intervention. The most basic ITS design utilizes a regression model that includes only three time-based covariates to estimate the pre-intervention slope (outcome trend before the intervention), a “step” or change in level (difference between observed and predicted outcome level at the first post-intervention time point), and a change in slope (difference between post- and pre-intervention outcome trend) (13, 32) [Figure 2 here].

Whether used for evaluating a natural experiment or, as is the focus here, for prospective evaluation of an intervention, the appropriateness of an ITS design depends on the nature of the intervention and outcome, and the type of data available. An ITS design requires the pre- and post-intervention periods to be clearly differentiated. When used prospectively, the investigator therefore needs to have control over the timing of the intervention. ITS analyses typically involve outcomes that are expected to change soon after an intervention is introduced or after a well-defined lag period. For example, for outcomes such as cancer or incident tuberculosis that develop long after an intervention is introduced and at a variable rate, it is difficult to clearly separate the pre- and post-intervention periods. Last, an ITS

analysis requires at least three time points in the pre- and post-intervention periods to assess trends. In general, a larger number of time points is recommended, particularly when the expected effect size is smaller, data are more similar at closer together time points (*i.e.*, auto-correlation), or confounding effects (*e.g.*, seasonality) are present. It is also important for investigators to consider any changes to data collection or recording over time, particularly if such changes are associated with introduction of the intervention.

In comparison to simple pre-post designs in which the average outcome level is compared between the pre- and post-intervention periods, the key advantage of ITS designs is that they evaluate for intervention effect while accounting for pre-intervention trends. Such trends are common due to factors such as changes in the quality of care, data collection and recording, and population characteristics over time. In addition, ITS designs can increase power by making full use of longitudinal data instead of collapsing all data to single pre- and post-intervention time points. The use of longitudinal data can also be helpful for assessing whether intervention effects are short-lived or sustained over time.

While the basic ITS design has important strengths, the key threat to internal validity is the possibility that factors other than the intervention are affecting the observed changes in outcome level or trend. Changes over time in factors such as the quality of care, data collection and recording, and population characteristics may not be fully accounted for by the pre-intervention trend. Similarly, the pre-intervention time period, particularly when short, may not capture seasonal changes in an outcome.

Detailed reviews have been published of variations on the basic ITS design that can be used to enhance causal inference. In particular, the addition of a control group can be particularly useful for assessing for the presence of seasonal trends and other potential time-varying confounders (52). Zombre et al (52) maintained a large number of control number of sites during the extended study period and were able to look at variations in seasonal trends as well as clinic-level characteristics, such as workforce density and sustainability. In addition to including a control group, several analysis phase strategies can be employed to strengthen causal inference including adjustment for time varying confounders and accounting for auto correlation.

3. Stepped Wedge Designs

Stepped wedge designs (SWDs) involve a sequential roll-out of an intervention to participants (individuals or clusters) over several distinct time periods (5, 7, 22, 24, 29, 30, 38). SWDs can include cohort designs (with the same individuals in each cluster in the pre and post intervention steps), and repeated cross-sectional designs (with different individuals in each cluster in the pre and post intervention steps) (7). In the SWD, there is a unidirectional, sequential roll- out of an intervention to clusters (or individuals) that occurs over different time periods. Initially all clusters (or individuals) are unexposed to the intervention, and then at regular intervals, selected clusters cross over (or ‘step’) into a time period where they receive the intervention [Figure 3 here]. All clusters receive the intervention by the last time interval (although not all individuals within clusters necessarily receive the intervention). Data is collected on all clusters such that they each contribute data during both control and intervention time periods. The order in which clusters receive the

intervention can be assigned randomly or using some other approach when randomization is not possible. For example, in settings with geographically remote or difficult-to-access populations, a non-random order can maximize efficiency with respect to logistical considerations.

The practical and social benefits of the stepped wedge design have been summarized in recent reviews (5, 22, 24, 27, 29, 36, 38, 41, 42, 45, 46, 51). In addition to addressing general concerns with RCTs discussed earlier, advantages of SWDs include the logistical convenience of staggered roll-out of the intervention, which enables a smaller staff to be distributed across different implementation start times and allows for multi-level interventions to be integrated into practice or 'real world' settings (referred to as the feasibility benefit). This benefit also applies to studies of de-implementation, prior to a new approach being introduced. For example, with a staggered roll-out it is possible to build in a transition cohort, such that sites can adjust to the integration of the new intervention, and also allow for a switching over in sites to de-implementing a prior practice. For a specified time period there may be 'mixed' or incomplete data, which can be excluded from the data analysis. However, associated with a longer duration of roll-out for practical reasons such as this switching, are associated costs in threats to internal validity, discussed below.

There are several limitations to the SWD. These generally involve consequences of the trade-offs related to having design control for the intervention roll-out, often due to logistical reasons on the one hand, but then having 'down the road' threats to internal validity. These roll-out related threats include potential lagged intervention effects for non-acute outcomes; possible fatigue and associated higher drop-out rates of waiting for the cross-over among clusters assigned to receive the intervention later; fidelity losses for key intervention components over time; and potential contamination of later clusters (22). Another drawback of the SWD is that it involves data assessment at each point when a new cluster receives the intervention, substantially increasing the burden of data collection and costs unless data collection can be automated or uses existing data sources. Because the SWD often has more clusters receiving the intervention towards the end of the intervention period than in previous time periods, there is a potential concern that there can be temporal confounding at this stage. The SWD is also not as suited for evaluating intervention effects on delayed health outcomes (such as chronic disease incidence), and is most appropriate when outcomes that occur relatively soon after each cluster starts receiving the intervention. Finally, as logistical necessity often dictates selecting a design with smaller numbers of clusters, there are relatedly challenges in the statistical analysis. To use standard software, the common recommendation is to have at least 20 to 30 clusters (35).

Stepped wedge designs can embed improvements that can enhance internal validity, mimicking the strength of RCTs. These generally focus on efforts to either reduce bias or achieve balance in covariates across sites and over time; and/or compensate as much as possible for practical decisions made at the implementation stage, which affect the distribution of the intervention over time and by sites. The most widely used approaches are discussed in order of benefit to internal validity: 1) partial randomization; 2) stratification and matching; 3) embedding data collection at critical points in time, such as with a phasing-in of intervention components, and 4) creating a transition cohort or wash-out period. The

most important of these SWD elements is random assignment of clusters as to when they will cross over into the intervention period. As well, utilizing data regarding time-varying covariates/confounders, either to stratify clusters and then randomize within strata (partial randomization) or to match clusters on known covariates in the absence of randomization, are techniques often employed to minimize bias and reduce confounding. Finally, maintaining control over the number and timing of data collection points over the study period can be beneficial in several ways. First, it can allow for data analysis strategies that can incorporate cyclical temporal trends (such as seasonality-mediated risk for the outcome, such as with flu or malaria) or other underlying temporal trends. Second, it can enable phased interventions to be studied for the contribution of different components included in the phases (e.g. passive then active intervention components), or can enable ‘pausing’ time, as when there is a structured wash out or transition cohort created for practical reasons (e.g. one intervention or practice is stopped/de-implemented, and a new one is introduced) (see Figure 4).

Table 2 provides examples of studies using SWD that have used one or more of the design approaches described above to improve the internal validity of the study. In the study by Killam et al 2010 (31), a non-randomized SWD was used to evaluate a complex clinic-based intervention for integrating anti-retro viral (ART) treatment into routine antenatal care in Zambia for post-partum women. The design involved matching clinics by size and an inverse roll-out, to balance out the sizes across the four groups. The inverse roll-out involved four strata of clinics, grouped by size with two clinics in each strata. The roll-out was sequenced across these eight clinics, such that one smaller clinics began earlier, with three clinics of increasing size getting the intervention afterwards. This was then followed by a descending order of clinics by size for the remaining roll-out, ending with the smallest clinic. This inverse roll-out enabled the investigators to start with a smaller clinic, to work out the logistical considerations, but then influence the roll-out such as to avoid clustering of smaller or larger clinics in any one step of the intervention.

A second design feature of this study involved the use of a transition cohort or wash-out period (see Figure 4) (also used in the Morrison et al 2015 study)(19, 37). This approach can be used when an existing practice is being replaced with the new intervention, but there is ambiguity as to which group an individual would be assigned to while integration efforts were underway. In the Killam study, the concern was regarding women who might be identified as ART-eligible in the control period but actually enroll into and initiate ART at an antenatal clinic during the intervention period. To account for the ambiguity of this transition period, patients with an initial antenatal visit more than 60 days prior to the date of implementing the ART in the intervention sites were excluded. For analysis of the primary outcome, patients were categorized into three mutually exclusive categories: a referral to ART cohort, an integrated ART in the antenatal clinics cohort, and a transition cohort. It is important to note that the time period for a transition cohort can add considerable time to an intervention roll-out, especially when there is to be a de-implementation of an existing practice that involves a wide range or staff or activities. As well, the exclusion of the data during this phase can reduce the study’s power if not built into the sample size considerations at the design phase.

Morrison et al 2015 (37) used a randomized cluster design, with additional stratification and randomization within relevant sub-groups to examine a two-part quality improvement intervention focusing on clinician uptake of patient cooling procedures for post-cardiac care in hospital settings (referred to as Targeted Temperature Management). In this study, 32 hospitals were stratified into two groups based on intensive care unit size (< 10 beds vs 10 beds), and then randomly assigned into four different time periods to receive the intervention. The phased intervention implementation included both passive (generic didactic training components regarding the intervention) and an active (tailored support to site-specific barriers identified in passive phase) components. This study exemplifies some of the best uses of SWD in the context of QI interventions that have either multiple components of for which there may be a passive and active phase, as is often the case with interventions that are layered onto systems change requirements (e.g. electronic records improvements/customization) or relate to sequenced guidelines implementation (as in this example).

Studies using a wait-list partial randomization design are also included in Table 2 (24, 27, 42). These types of studies are well-suited to settings where there is routine enumeration of a cohort based on a specific eligibility criteria, such as enrolment in a health plan or employment group, or from a disease-based registry, such as for diabetes (27, 42). It has also been reported that this design can increase efficiency and statistical power in contrast to cluster-based trials, a crucial consideration when the number of participating individuals or groups is small (22).

The study by Grant et al et al uses a variant of the SWD for which individuals within a setting are enumerated and then randomized to get the intervention. In this example, employees who had previously screened positive for HIV at the company clinic as part of mandatory testing, were invited in random sequence to attend a workplace HIV clinic at a large mining facility in South Africa to initiate a preventive treatment for TB during the years prior to the time when ARTs were more widely available. Individuals contributed follow-up time to the “pre-clinic” phase from the baseline date established for the cohort until the actual date of their first clinic visit, and also to the “post- clinic” phase thereafter. Clinic visits every 6 months were used to identify incident TB events. Because they were looking at reduction in TB incidence among the workers at the mine and not just those in the study, the effect of the intervention (the provision of clinic services) was estimated for the entire study population (incidence rate ratio), irrespective of whether they actually received isoniazid.

CONSIDERATIONS IN CHOOSING BETWEEN QED

We present a decision ‘map’ approach based on a Figure 5 to assist in considering decisions in selecting among QEDs and for which features you can pay particular attention to in the design [Figure 5 here].

First, at the top of the flow diagram (1), consider if you can have multiple time points you can collect data for in the pre and post intervention periods. Ideally, you will be able to select more than two time points. If you cannot, then multiple sites would allow for a non-

equivalent pre-post design. If you can have more than the two time points for the study assessments, you next need to determine if you can include multiple sites (2). If not, then you can consider a single site point ITS. If you can have multiple sites, you can choose between a SWD and a multiple site ITS based on whether or not you observe the roll-out over multiple time points, (SWD) or if you have only one intervention time point (controlled multiple site ITS)

STRATEGIES TO STRENGTHEN EXTERNAL VALIDITY

In a recent article in this journal (26), the following observation was made that there is an unavoidable trade-off between these two forms of validity such that with a higher control of a study, there is stronger evidence for internal validity but that control may jeopardize some of the external validity of that stronger evidence. Nonetheless, there are design strategies for non-experimental studies that can be undertaken to improve the internal validity while not eliminating considerations of external validity. These are described below across all three study designs.

1. Examine variation of acceptability and reach among diverse sub-populations

One of the strengths of QEDs is that they are often employed to examine intervention effects in real world settings and often, for more diverse populations and settings. Consequently, if there is adequate examination of characteristics of participants and setting-related factors it can be possible to interpret findings among critical groups for which there may be no existing evidence of an intervention effect for. For example in the Campus Watch intervention (16), the investigator over-sampled the Maori indigenous population in order to be able to stratify the results and investigate whether the program was effective for this under-studied group. In the study by Zombré et al (52) on health care access in Burkina Faso, the authors examined clinic density characteristics to determine its impact on sustainability.

2. Characterize fidelity and measures of implementation processes

Some of the most important outcomes for examination in these QED studies include whether the intervention was delivered as intended (i.e., fidelity), maintained over the entire study period (i.e., sustainability), and if the outcomes could be specifically examined by this level of fidelity within or across sites. As well, when a complex intervention is related to a policy or guideline shift and implementation requires logistical adjustments (such as phased roll-outs to embed the intervention or to train staff), QEDs more truly mimic real world constraints. As a result, capturing processes of implementation are critical as they can describe important variation in uptake, informing interpretation of the findings for external validity. As described by Prost et al (41), for example, it is essential to capture what occurs during such phased intervention roll-outs, as with following established guidelines for the development of complex interventions including efforts to define and protocolize activities before their implementation (17,18, 28). However, QEDs are often conducted by teams with strong interests in adapting the intervention or 'learning by doing', which can limit interpretation of findings if not planned into the design. As done in the study by Baillet et al (3), the investigators refined intervention, based on year 1 data, and then applied in years 2–

3, at this later time collecting additional data on training and measurement fidelity. This phasing aspect of implementation generates a tension between protocolizing interventions and adapting them as they go along. When this is the case, additional designs for the intervention roll-out, such as adaptive or hybrid designs can also be considered.

3. Conduct community or cohort-based sampling to improve inference

External validity can be improved when the intervention is applied to entire communities, as with some of the community-randomized studies described in Table 2 (12, 21). In these cases, the results are closer to the conditions that would apply if the interventions were conducted ‘at scale’, with a large proportion of a population receiving the intervention. In some cases QEDs also afford greater access for some intervention research to be conducted in remote or 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.

4. Employ a model or framework that covers both internal and external validity

Frameworks can be helpful to enhance interpretability of many kinds of studies, including QEDs and can help ensure that information on essential implementation strategies are included in the results (44). Although several of the case studies summarized in this article included measures that can improve external validity (such as sub-group analysis of which participants were most impacted, process and contextual measures that can affect variation in uptake), none formally employ an implementation framework. Green and Glasgow (2006) (25) have outlined several useful criteria for gauging the extent to which an evaluation study also provides measures that enhance interpretation of external validity, for which those employing QEDs could identify relevant components and frameworks to include in reported findings.

CONCLUSION

It has been observed that it is more difficult to conduct a good quasi-experiment than to conduct a good randomized trial (43). Although QEDs are increasingly used, it is important to note that randomized designs are still preferred over quasi-experiments except where randomization is not possible. In this paper we present three important QEDs and variants nested within them that can increase internal validity while also improving external validity considerations, and present case studies employing these techniques.

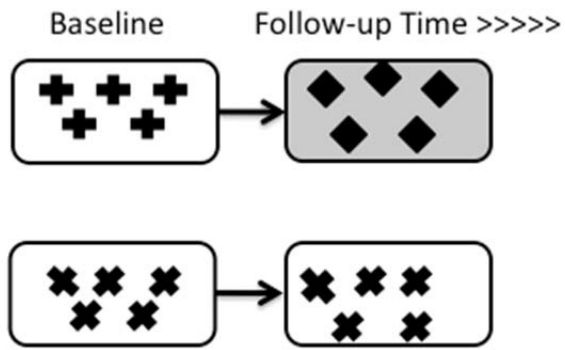
LITERATURE CITED

1. Alonso PL, Lindsay WS, Armstrong-Schellenberg JRM, Konteh M, Keita MK et al. Malaria Control Trial using Insecticide-Treated Bed Nets and Targeted Chemoprophylaxis in a Rural Area of The Gambia, West Africa. 1993. Transactions of the Royal Society of Tropical Medicine and Hygiene Volume 87(2):1–60
2. Ammerman A, Smith T, Calancie L. 2014. Practice-based evidence in public health: improving reach, relevance, and results. *Annu. Rev. Public Health* 35:47–63 [PubMed: 24641554]
3. Bailet LL, Repper K, Murphy S, Piasta S, Zettler-Greeley C. 2013. Emergent Literacy Intervention for Pre-kindergarteners at Risk for Reading Failure: Years 2 and 3 of a Multiyear Study. *Journal of Learning Disabilities*. 46:133–153. [PubMed: 21685354]

4. Basu S, Meghani A and Siddiqi A 2017 Evaluating the Health Impact of Large-Scale Public Policy Changes: Classical and Novel Approaches *Annu. Rev. Public Health* 38:351–370 [PubMed: 28384086]
5. Beard E, Lewis JJ, Copas A, Davey C, Osrin D et al. 2015 Stepped wedge randomised controlled trials: systematic review of studies published between 2010 and 2014. *Trials* 16:353.1–14. [PubMed: 26278881]
6. Bernal JL, Cummins S, Gasparrini A. 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int J Epidemiol* 2017; 46 (1): 348–355. [PubMed: 27283160]
7. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. 2006. *BMC Med Res Methodol* 2006;6:54. [PubMed: 17092344]
8. Brown CH, Ten Have TR, Jo B, Dagne G, Wyman PA, et al. 2009. Adaptive Designs for Randomized Trials in Public Health. *Annu. Rev. Public Health* 2009. 30:1–25 [PubMed: 19296774]
9. Brown CH, Curran G, Palinkas LA, Aarons GA, Wells KB, Jones L, Collins LM, Duan N, et al. 2017. An Overview of Research and Evaluation Designs for Dissemination and Implementation. *Annu Rev Public Health*. 3 20;38:1–22. [PubMed: 28384085]
10. Brownson RC, Diez Roux AV, Swartz k. 2014. Commentary: Generating Rigorous Evidence for Public Health: The Need for New Thinking to Improve Research and Practice. *Annu. Rev. Public Health* 35:1–7 [PubMed: 24328987]
11. Campbell DT, and Stanley JC, “Experimental and Quasi-Experimental Designs for Research on Teaching.” In Gage NL (ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally, 1963.
12. Cisse B, Ba EH, Sokhna C, NDiaye JL, Gomis JF, Dial Y, et al. 2016. Effectiveness of Seasonal Malaria Chemoprevention in Children under Ten Years of Age in Senegal: A Stepped- Wedge Cluster-Randomised Trial. *PLoS Med* 13:1–18.
13. Cochrane Reviews. 2012. Interrupted time series (ITS) analyses
14. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. *Med Care*. 3 2012;50(3):217–226 [PubMed: 22310560]
15. Cook Thomas D., and Campbell Donald T.. “The design and conduct of quasi-experiments and true experiments in field settings.” *Handbook of industrial and organizational psychology* 223 (1976): 336.
16. Cousins K, Connor JL, Kypri K Effects of the Campus Watch intervention on alcohol consumption and related harm in a university population 2014. *Drug and Alcohol Dependence* 143–126
17. Craig P, Dieppe P, Macintyre S, Nazareth I, Petticrew M. 2008. Developing and evaluating complex interventions: the new Medical Research Council guidance. *Br Med J*. 337:1655.
18. Craig P, Katikireddi SV, Leyland A, Popham F. 2017. Natural Experiments: An Overview of Methods, Approaches, and Contributions to Public Health Intervention Research. *Annu Rev Public Health* 38:39–56 [PubMed: 28125392]
19. Dainty KN, Scales D, Brooks S, Needham DM, Dorian P et al. 2011. A knowledge translation collaborative to improve the use of therapeutic hypothermia in post-cardiac arrest patients: protocol for a stepped wedge randomized trial. *Implementation Science* 6:4. [PubMed: 21235799]
20. Fan E, Laupacis A, Pronovost P, et al. 2010. How to use an article about quality improvement. *JAMA* 304: 2279–87 –94 [PubMed: 21098772]
21. Fernald L, Gertler PJ, Neufeld LM. 2008 Role of cash in conditional cash transfer programmes for child health, growth, and development: an analysis of Mexico’s Oportunidades *Lancet* 371: 828–37 [PubMed: 18328930]
22. Fok CCT, Henry D, Allen A. 2015. Research Designs for Intervention Research with Small Samples II: Stepped Wedge and Interrupted Time-Series Designs *Prev Sci*. 16:967–977 [PubMed: 26017633]
23. Glasgow RE, Vinson C, Chambers D, Khoury MJ, Kaplan RM, Hunter C. National Institutes of Health Approaches to Dissemination and Implementation Science: Current and Future Directions. *Am J Public Health*. Jul 2012;102(7):1274–1281.

24. Grant AD, Charalambous S, Fielding KL, Day JH, Corbett EL, Chaisson RE et al. 2005. Effect of Routine Isoniazid Preventive Therapy on Tuberculosis Incidence Among HIV-Infected Men in South Africa *JAMA* 293: 2719–2725 [PubMed: 15941800]
25. Green LW, Glasgow RE. Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. *Eval Health Prof.* 3 2006;29(1):126–153 [PubMed: 16510882]
26. Green LW, Brownson RC, Fielding JE. 2017 Introduction: How Is the Growing Concern for Relevance and Implementation of Evidence-Based Interventions Shaping the Public Health Research Agenda? *Annu Rev Public Health* 20: 38
27. Handley MA, Schillinger D, Shiboski S. 2011. Quasi-Experimental Designs in Practice-based Research Settings: Design and Implementation Considerations *J Am Board Fam Med.* 24: 589–596. [PubMed: 21900443]
28. Hawe P. Lessons from Complex Interventions to Improve Health. 2015. *Annual Rev of Public Health.*
29. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. 2015 The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *Br Med J* 350:
30. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. 2007 *Contemp Clin Trials* 28:182–91 [PubMed: 16829207]
31. Killam WP, Tambatamba BC, Chintu N, et al. 2010. Antiretroviral therapy in antenatal care to increase treatment initiation in HIV-infected pregnant women: a stepped-wedge evaluation. *AIDS* 24:85–9 [PubMed: 19809271]
32. Kontopantelis E, Doran T, Springate DA, Buchian I, Reeves D 2015 Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis *BMJ* 2015;350:h2750 [PubMed: 26058820]
33. Krass I Quasi experimental designs in pharmacist intervention research 2016. *Int J Clin Pharm* 38:647–654 [PubMed: 26825756]
34. Leviton LC. 2017. Generalizing about Public Health Interventions: A Mixed-Methods Approach to External Validity. *Annu Rev Public Health.* 38:371–391. [PubMed: 28125391]
35. McNeish DM and Harring JR. 2014. Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communication in Statistics* 46: 855–69.
36. Mdege ND, Man MS, Taylor CA, Torgerson DJ. 2011. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation *Journal of Clinical Epidemiology* 64:936–48. [PubMed: 21411284]
37. Morrison LJ, Brooks SC, Dainty KN, Dorian P, Needham DM, Ferguson ND, Rubenfeld GD, Slutsky AS, Wax RS, Zwarenstein M, Thorpe K, Zhan C, Scales DC. 2015. Improving use of targeted temperature management after out-of-hospital cardiac arrest: a stepped wedge cluster randomized controlled trial. *Strategies for Post-Arrest Care Network. Crit Care Med.* 43:954–64 [PubMed: 25654175]
38. Murray DM, Pennell M, Rhoda D, Hade EM, Paskett ED. 2010. Designing Studies That Would Address the Multilayered Nature of Health Care *J Natl Cancer Inst Monogr* 2010;40:90–96
39. Naci H, Soumerai SB. 2016. History Bias, Study Design, and the Unfulfilled Promise of Pay-for-Performance Policies in Health Care. *Prev Chronic Dis* 13:160133
40. Peters DH, Adam T, Alonge O, Agyepong IA, Tran N. Implementation research: what it is and how to do it. *BMJ.* 2013;347:f6753 [PubMed: 24259324]
41. Prost A, Binik A, Abubakar I, Roy A, De Allegri M et al. 2015. Logistic, ethical, and political dimensions of stepped wedge trials: critical review and case studies *Trials* 16:351 [PubMed: 26278521]
42. Ratanawongsa N, Handley MA, Quan J, Sarkar U, Pfeifer K, Soria C, Schillinger D. 2012. Quasi-experimental trial of diabetes Self-Management Automated and Real-Time Telephonic Support (SMARTSteps) in a Medicaid managed care plan: study protocol. *BMC Health Serv Res* :22

43. Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin
44. Tabak RG, Khoong EC, Chambers DA, Brownson RC. Bridging research and practice: models for dissemination and implementation research. *Am J Prev Med*. 9 2012;43(3):337–350 [PubMed: 22898128]
45. Turner EL, Li F, Gallis JA, Prague M, Murray DM. 2017. Review of recent methodological developments in group- randomized trials: part 1—design. *Am J Public Health* 107:
46. Turner ET, Prague M, Gallis JA, Fan L. 2017. Review of Recent Methodological Developments in Group-Randomized Trials: Part 2—Analysis. *Am J Public Health*. published online ahead of print 5 18, 2017
47. Weihart LS, Galva LW, Yn AF, Stevens P, Mwenyekonde TN et al. 2017. Mixed-Method Quasi-Experimental Study of Outcomes of a Large-Scale Multilevel Economic and Food Security Intervention on HIV Vulnerability in Rural Malawi. *AIDS Behav* 21:712–723 [PubMed: 27350305]
48. Wertz D, Hou L, Devries A et al. Clinical and Economic Outcomes of the Cincinnati Pharmacy Coaching Program for Diabetes and Hypertension 2012. *Managed Care*. 44–54.
49. West SG, Duan N, Pequegnat W, Gaist P, Des Jarlais DC, et al. 2008. Alternatives to the randomized controlled trial. *Am J Public Health*. 98:1359–66. [PubMed: 18556609]
50. White H, & Sabarwal S. *Quasi-experimental Design and Methods, Methodological Briefs: Impact Evaluation* 8. 2014. UNICEF Office of Research, Florence.
51. Wyman PA, Henry D, Knoblauch S, Brown CH. 2015. Designs for Testing Group-Based Interventions with Limited Numbers of Social Units: The Dynamic Wait-Listed and Regression Point Displacement Designs *Prev Sci*. 16:956–966 [PubMed: 25481512]
52. Zombré D, Allegri MD, Ridde V. 2017. Immediate and sustained effects of user fee exemption on healthcare utilization among children under five in Burkina Faso: A controlled interrupted time-series analysis *Social Science & Medicine* 179: 27e35 [PubMed: 28242542]






Each  represents a cluster and the shaded site with a  is receiving the intervention. The non-equivalent control group is depicted with 

Figure 1.
Illustration of the Pre-Post Non-Equivalent Control Group Design

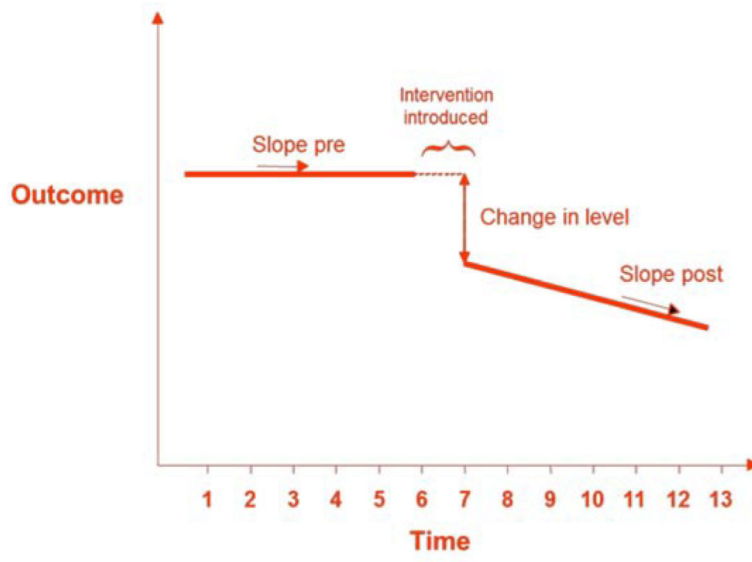
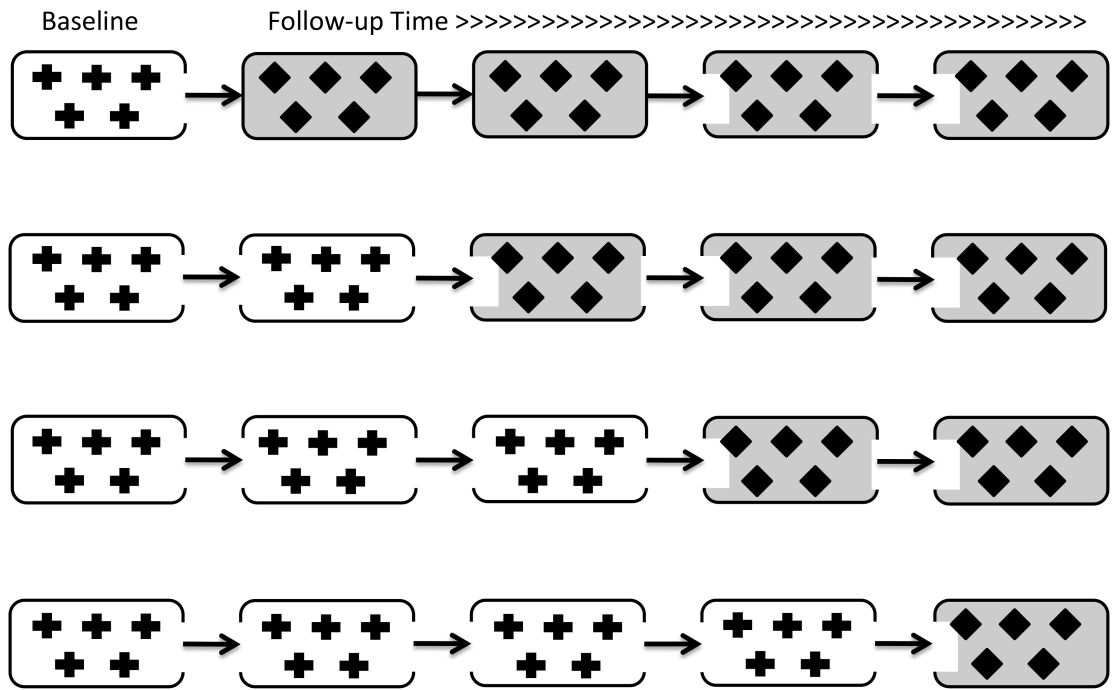


Figure 2.
Interrupted Time Series Design






Each  represents a cluster and the shaded sites are receiving the intervention. The unshaded sites are acting as control sites for the time period. Unexposed individuals  can be sampled and cross-over to the intervention  in wait list variants of the stepped wedge design

Figure 3.
 Illustration of the stepped wedge study design-Intervention Roll-Out Over Time*
 * Adapted from Turner et al 2017

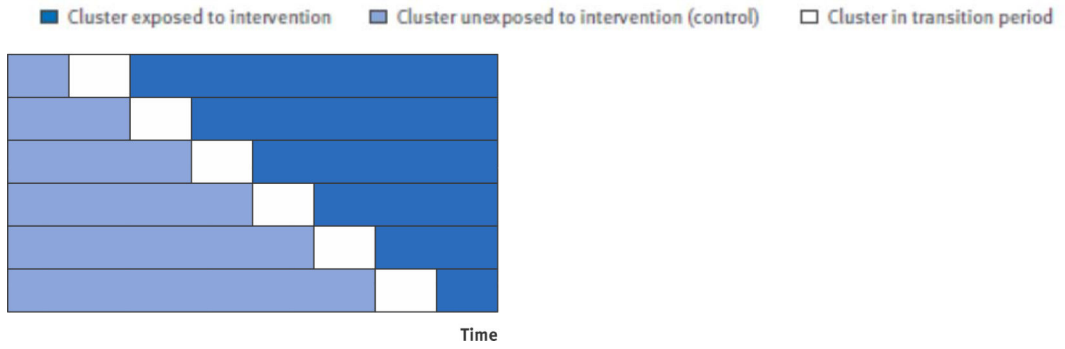


Figure 4.
Illustration of the stepped wedge study design- Summary of Exposed and Unexposed Cluster Time*
Adapted from Hemming 2015

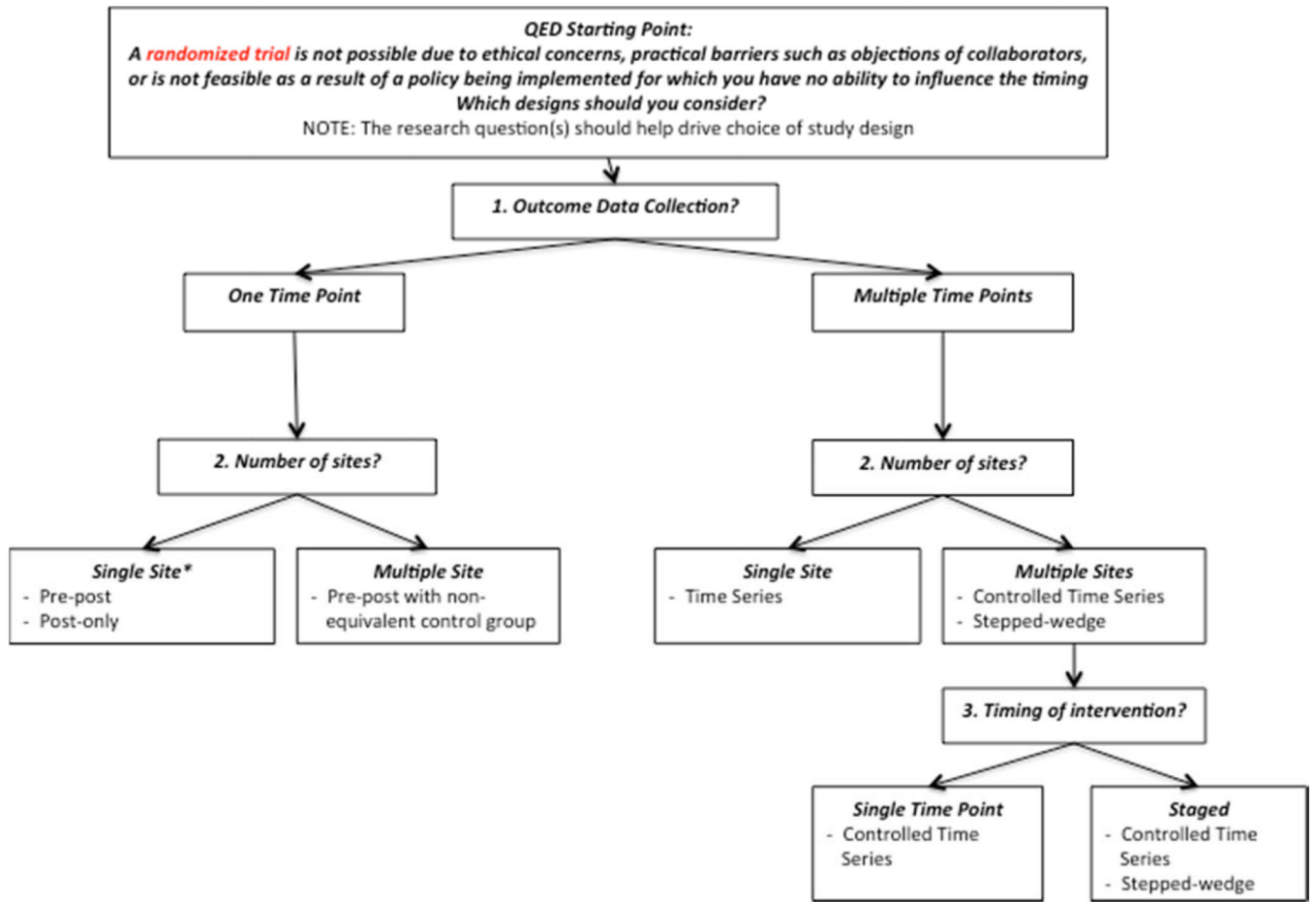


Fig 5.
 Quasi-Experimental Design Decision-Making Map

Table 1

Overview of Commonly Used QED in Intervention Research*

QED Design	Key Design Elements	Advantages	Disadvantages
Pre-Post with Non-equivalent control group	Comparison of those receiving the intervention with those not receiving it. Analysis is usually based on estimating the difference in the amount of change over time in the outcome of interest between the two groups, beginning with the intervention and moving forward in time; The two groups can also be a different group examined using a before and after intervention cohort	Simplicity of data collection, when smaller number of time points, and associated lower cost; less cumbersome to implement than other designs	Temporal biases are a substantial risk and may result in regression to the mean or over-interpretation of intervention effects; quality of data may vary in different time periods resulting in measurement error; non-equivalent sites may not be comparable for important covariates
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	Useful for when there is a small number of communities or groups, as each group acts as their own control May be only option for studying impacts of large scale health policies	Requires a large number of measurements, may not be feasible for geographically dispersed areas
Stepped Wedge Design	Intervention is rolled out over time, usually at the site level. Participants who initially do not receive the intervention later-cross over to receive the intervention. Those that wait, provide control data during the time others receive the intervention, reducing the risk of bias due to time and time-dependent covariates. Can either be based on serial cross-sectional data collected by sites for different time periods (sites cross over) or by following a cohort of same individuals over time (individuals cross over)	All clusters or wait list groups eventually receives the intervention; Do not need to supply intervention in all sites in a short time frame “staggered implementation”	May not be able to randomly assign roll-out of sites, thereby potentially jeopardizing internal validity Cannot guarantee everyone in each cluster or list will receive the intervention during the time that cluster is receiving the intervention -Often takes longer than other designs to implement -Control data must be collected or ascertained from sites or participants -Site differences and implementation processes can vary significantly over time -Risk of contamination in later sites or intervention fatigue – both can wash out potential intervention effects

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Improving Quasi-Experimental Designs-Internal and External Validity Considerations

Study/General Design	Intervention	Design Strategy to Improve Internal Validity	Design Strategy to Improve External Validity
Pre-Post Designs with Non-Equivalent Control Group			
Cousins et al 2016	Campus Watch program targeting problem drinking and violence at 1 university campus with 5 control campuses in New Zealand	<p>Consistent Sampling Across and Within Control sites:</p> <ul style="list-style-type: none"> Standardization of independent repeat sampling, survey and follow-up methods across all sites (5 control and 1 intervention site) <p>Multiple non-equivalent control groups</p> <ul style="list-style-type: none"> 5 sites as controls studies aggregate and individually as controls <p>Collection of additional data to support inference:</p> <ul style="list-style-type: none"> Consumption and harms data from national surveys to compare data trends over time 	Over-sampling of indigenous groups to extend interpretation of findings
Wertz et al 2012	Chronic disease management program with pharmacist-based patient coaching within a health care insurance plan in Cincinnati, US	<p>Prospective cohort sampling to improve control group similarity:</p> <ul style="list-style-type: none"> Matching of participants with non-participants on demographic and health care access measures (using propensity score matching) 	
Alonso 1993	Distribution of bed nets to prevent malaria and reduce malaria mortality in Gambia 41 sites receiving intervention compared to external villages (which differed by size and ethnic distribution)	<p>Collection of additional data to support inferences:</p> <ul style="list-style-type: none"> Examination of data trends during the highest infection times of the year (i.e., rainy season vs dry season) to see if rates were higher then. <p>Subgroup analysis:</p> <ul style="list-style-type: none"> Detailed study of those using bed nets within intervention villages (i.e., guaranteed exposure "dose", to examine dose-response in intervention arm 	
Interrupted Time Series			
Study/General Design	Intervention	Design Strategy to Improve Internal Validity	Design Strategy to Improve External Validity
Pellegrin 2016 Interrupted time series with comparison group	Formal transfer of high-risk patients being discharged from hospital to a community-based pharmacist follow-up program for up to 1 year post-hospitalization (6 intervention and 5 control sites)	<p>Extended period of data collection:</p> <ul style="list-style-type: none"> Long baseline period (12 pre-intervention data points) <p>Control Group Inverse Roll-Out/Covariate Balance:</p> <ul style="list-style-type: none"> Intervention roll-out staggered based on staff availability (site 1 had eight post-intervention data points while site 8 had two) 	Detailed implementation-related process measures monitored (and provided to individual community-based pharmacists regarding their performance) over entire study period
Robinson 2015 Interrupted time series without control group	New hospital discharge program to support high-risk patients with nurse telephone follow-up and referral to specific services (such as pharmacists for medication reconciliation and review)	<p>Regression discontinuity analysis:</p> <ul style="list-style-type: none"> Additionally examined regression discontinuity during the intervention period to determine if the risk score used to determine eligibility for the program influenced the outcome 	Measured implementation outcomes of whether the intervention was delivered with high fidelity to the protocols
Zombré 2017 Interrupted time series with comparison group	Removal of direct payment at point of health care services for children under 5, very low income individuals and pregnant women re: consultations, medications and hospitalizations	<p>Extended period of data collection:</p> <ul style="list-style-type: none"> Built into a pilot to collect control data, and then extend this work to include additional districts, one intervention and one non-intervention district, along with 6 additional years of observation. <p>Control Group</p>	Examined sustainability over 72 months of follow-up, and associations with clinic characteristics, such as density of workforce.
Stepped Wedge Design			

Study/General Design	Intervention	Design Strategy to Improve Internal Validity	Design Strategy to Improve External Validity
Killam et al 2010 Non-randomized stepped wedge cluster trial	Site-level roll out of integrated antiretroviral treatment (ART) intervention in 8 public sector clinics, to achieve more rapid treatment initiation among women with HIV in Zambia, than the existing referral method used for initiation of treatment.	<p>Site Matching:</p> <ul style="list-style-type: none"> The 8 sites were matched into four pairs based on the number of HIV-infected pregnant women expected in each site. <p>Inverse Roll-Out/Covariate Balance:</p> <ul style="list-style-type: none"> The intervention roll out was done for one member of the least busy pair, one member of the second busiest pair, one member of the third busiest pair, and one member of the busiest pair. Rollout to the remaining pairs proceeded in reverse order. <p>Wash-out Period/ Transition Cohort/:</p> <ul style="list-style-type: none"> A transition cohort was established that was later excluded from the analysis. It included women who were identified as eligible in the control period of time close to the time the intervention was starting. 	
Morison et al 2015 See also: Dainty et al 2011 Randomized stepped wedge cluster trial	Multi-faceted quality improvement intervention with a passive and an active phase among 6 regional emergency medical services systems and 32 academic and community hospitals in Ontario, Canada. The intervention focused on comparing interventions to improve the implementation of targeted temperature management following out-of-hospital cardiac arrest through passive (education, generic protocol, order set, local champions) versus additional active quality improvement interventions (nurse specialist providing site-specific interventions, monthly audit-feedback, network educational events, internet blog) versus no intervention (baseline standard of care).	<p>Cluster randomization with constrained randomization for size:</p> <ul style="list-style-type: none"> Randomization at the level of the hospital, rather than the patient to minimize contamination, since the intervention targeted groups of clinicians. Hospitals were stratified by number of Intensive Care Unit beds (< 10 beds vs 10 beds as a proxy for hospital size). Randomization was done within strata. <p>Transition Cohort/ Phased Roll-Out of Intervention Components:</p> <ul style="list-style-type: none"> Formalized a transition cohort for which a more passive intervention strategy was tested. This also allowed more time for sites to adopt all elements of the complex intervention before crossing over to the active intervention group. 	Characterization of system and organizational factors that might affect adoption: Collection of longitudinal data relevant to implementation processes that could impact interpretation of findings such as academic vs community affiliation, urban vs rural (bed size)
Cissé et al 2016 Randomized stepped wedge cluster trial	Seasonal malaria prophylaxis for children up to age 10 in central Senegal given to households monthly through health system staff led home visits during the malaria season. The first two phases of implementation focused on children under age 5 years and the last phase included children up to age 10 years, and maintained a control only group of sites during this period.	<p>Constrained Randomization of Clusters by Geographic Indicators and Time Period:</p> <ul style="list-style-type: none"> Constrained randomization of program roll-out across 54 health posts catchment areas and center-covered regions, More sites received the intervention later stages (n=18) than in beginning (n=9). To achieve balance within settings for potential confounders (since they did not have data on malaria incidence), such as distance from river, distance from health center, population size and number of villages, assessment of ability to implement. <p>Control-Only Clusters:</p> <ul style="list-style-type: none"> Included nine clinics as control sites throughout the study period. 	Characterization of factors that might affect usage and adherence made with longitudinal data: Independent evaluations of malaria prophylaxis usage, adherence, and acceptance were included prospectively, using routine health cards at family level and with external assessments from community surveys. In-depth interviews conducted across community levels to understand acceptability and other responses to the intervention Included an embedded study broadening inclusion criteria, to focus on a wider age group of at risk children
Grant et al 2005 Wait-list randomized stepped wedge design	Enrollment of 1,655 male mine employees with HIV infection randomized over a short period of time into an intervention to prevent TB infection (use of isoniazid preventive therapy), among individuals with HIV. Treatment was self-administered for 6 months or for 12 months and results were based on cohort analyses.	<p>Wait-list Individual Randomization:</p> <ul style="list-style-type: none"> Employees were invited in random sequence to attend a workplace HIV clinic. 	Enumeration of at risk cohort and estimation of spill-over effect beyond those enrolled: Since they used an enrollment list, they were able to estimate the effect of the intervention (the provision of clinic services) among the entire eligible population, not just those

			enrolled in the intervention over the study period.
Ratanawongsa et al; Handley et al 2011 Wait-list randomized stepped wedge design	Enrollment of 362 patients with diabetes into a health-IT enabled self-management support telephone coaching program, using a wait-list generated from a regional health plan, delivered in 3 languages.	Wait-list for Eligible Randomization: • Patients were identified from an actively maintained diabetes registry covering 4 safety net health clinics in the United States, and randomized to receive the coaching intervention immediately or after 6 months. Constrained Randomization by Language • Patients were randomized to balance enrolment for English, Cantonese, and Spanish, over the study period.	External validity-related measures for acceptability among patients as well as fidelity measures, for the health IT-enabled health coaching intervention were assessed using a fidelity framework.
Bailet et al 2011	Literacy intervention for pre-kindergarten children at risk for reading failure in a southern US city administered in child care and pre-school sites, delivered twice a week for 9 weeks. For large sites, did not randomize at site level, but split the schools, so all children could be taught in the intervention period, either fall or spring. At-risk children in these “split” schools received intervention at only one of the two time points (as did their “non-split school” peers); however, the randomization to treatment group occurred at the child level.	Constrained randomization for size, then cluster randomization within strata. • Random assignment of clusters (schools). Site Matching: • Matched pairs of child care centers by zip code and percentage of children receiving a state-sponsored financial subsidy. Within these groups random assignment to receive either immediate or deferred enrolment into the intervention.	External validity was enhanced in years 2–3 with a focus on teacher training for ensuring measures fidelity, completion of each week of the curriculum to enhance assessment of a potential dose-response. Refined intervention applied in years 2–3, based on initial data.
Fernald et al 2008	Mexican Government randomly chose 320 early intervention and 186 late (approximately one year later) intervention communities in seven states for Oportunidades, which provided cash transfers to families conditional on children attending school and family members obtaining preventive medical care and attending <i>pláticas</i> —education talks on health-related topics.	Constrained Randomization of Clusters Time Period: • More communities randomized to an early intervention period	

BOX 1.**DEFINITIONS AND TERMS USED IN PAPER**

Terms and Definitions	
Quasi-Experimental Design:	QEDs include a wide range of nonrandomized or partially randomized pre-post intervention studies
Pre-Post Design	A QED with data collected before and after an intervention is introduced, and then the compared. An added control group can be added for a Pre-Post Design with a Non-Equivalent control group
Non-Equivalent Control Group	A control group that is not randomly assigned to receive or not receive the intervention. Usually, an intact group is selected that is thought to be similar to the intervention group.
Interrupted Time Series Design	Multiple observations are evaluated for several consecutive points in time before and after intervention within the same individual or group
Stepped Wedge Design	A type of crossover design where the time of crossover is randomized
Wash out period	Time period for which a prior practice or intervention is stopped, and a new one is implemented, for which both interventions may be operating, and thus the data is excluded.
Inverse Roll-Out	Sites are rolled out to receive the intervention using a structured approach to create balance between the sites over the roll-out time period, using a sample characteristic that is ordered (and then reverse ordered). Commonly size or geography may be used. (e.g. 1,2,3,4 for size followed by 4,3,2,1)
Partial Randomization	A type of stratified randomization, with strata constructed for potential confounding variables and randomization occurs separately within each stratum (also called blocked randomization)
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
External Validity	External validity describes the extent to which a research conclusion can be generalized to the population or to other settings

BOX 2:**Common Threats to Internal Validity of Quasi-Experimental Designs Evaluating Interventions in ‘Real World’ Settings**

History Bias	Events other than the intervention occurring at the same time may influence the results
Selection Bias	Systematic differences in subject characteristics between intervention and control groups that are related to the outcome
Maturation Bias	Occurs when changes occur to individuals in the groups, differently, over time resulting in effects, in addition to (or rather than) the treatment condition, that may change the performance of participants in the post-test relative to the pre-test
Lack of Blinding	Awareness of group assignment can influence those delivering or receiving the intervention
Differential Drop-Out	Attrition that may affect either intervention or control groups differently and result in selection bias and/or loss of statistical power
Variability in interactive effects	Implementation of intervention with multiple components may vary across the implementation process and by sites

* adapted from White et al