HHS Public Access

Author manuscript

Psychiatry Res. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

Psychiatry Res. 2020 January ; 283: . doi:10.1016/j.psychres.2019.06.027.

Experimental and Quasi-Experimental Designs in Implementation Research

Christopher J. Miller[a,b,*], **Shawna N. Smith**[c,d], **Marianne Pugatch**[a]

[a]VA Boston Healthcare System, Center for Healthcare Organization and Implementation Research (CHOIR), United States Department of Veterans Affairs, Boston, MA, USA

[b]Department of Psychiatry, Harvard Medical School, Boston, MA, USA

[c]Department of Psychiatry, University of Michigan Medical School, Ann Arbor, MI, USA

[d]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA

Abstract

Implementation science is focused on maximizing the adoption, appropriate use, and sustainability of effective clinical practices in real world clinical settings. Many implementation science questions can be feasibly answered by fully experimental designs, typically in the form of randomized controlled trials (RCTs). Implementation-focused RCTs, however, usually differ from traditional efficacy- or effectiveness-oriented RCTs on key parameters. Other implementation science questions are more suited to quasi-experimental designs, which are intended to estimate the effect of an intervention in the absence of randomization. These designs include pre-post designs with a non-equivalent control group, interrupted time series (ITS), and stepped wedges, the last of which require all participants to receive the intervention, but in a staggered fashion. In this article we review the use of experimental designs in implementation science, including recent methodological advances for implementation studies. We also review the use of quasi-experimental designs in implementation science, and discuss the strengths and weaknesses of these approaches. This article is therefore meant to be a practical guide for researchers who are interested in selecting the most appropriate study design to answer relevant implementation science questions, and thereby increase the rate at which effective clinical practices are adopted, spread, and sustained.

Keywords

implementation; SMART design; quasi-experimental; pre-post with non-equivalent control group; interrupted time series; stepped wedge

[*]Corresponding Author: Christopher.Miller8@va.gov; (p) 857-364-5688 (fax) 857-364-6140.

1. Background

The first documented clinical trial was conducted in 1747 by James Lind, a royal navy physician, who tested the hypothesis that citrus fruit could cure scurvy. Since then, based on foundational work by Fisher and others (1935), the randomized controlled trial (RCT) has emerged as the gold standard for testing the efficacy of treatment versus a control condition for individual patients. Randomization of patients is seen as a crucial to reducing the impact of measured or unmeasured confounding variables, in turn allowing researchers to draw conclusions regarding causality in clinical trials.

As described elsewhere in this special issue, implementation science is ultimately focused on maximizing the adoption, appropriate use, and sustainability of effective clinical practices in real world clinical settings. As such, some implementation science questions may be addressed by experimental designs. For our purposes here, we use the term “experimental” to refer to designs that feature two essential ingredients: first, manipulation of an independent variable; and second, random assignment of subjects. This corresponds to the definition of randomized experiments originally championed by Fisher (1925). From this perspective, experimental designs usually take the form of RCTs—but implementation-oriented RCTs typically differ in important ways from traditional efficacy- or effectiveness-oriented RCTs. Other implementation science questions require different methodologies entirely: specifically, several forms of quasi-experimental designs may be used for implementation research in situations where an RCT would be inappropriate. These designs are intended to estimate the effect of an intervention despite a lack of randomization. Quasi-experimental designs include pre-post designs with a nonequivalent control group, interrupted time series (ITS), and stepped wedge designs. Stepped wedges are studies in which all participants receive the intervention, but in a staggered fashion. It is important to note that quasi-experimental designs are not unique to implementation science. As we will discuss below, however, each of them has strengths that make them particularly useful in certain implementation science contexts.

Our goal for this manuscript is two-fold. First, we will summarize the use of experimental designs in implementation science. This will include discussion of ways that implementation-focused RCTs may differ from efficacy- or effectiveness-oriented RCTs. Second, we will summarize the use of quasi-experimental designs in implementation research. This will include discussion of the strengths and weaknesses of these types of approaches in answering implementation research questions. For both experimental and quasi-experimental designs, we will discuss a recent implementation study as an illustrative example of one approach.

1. Experimental Designs in Implementation Science

RCTs in implementation science share the same basic structure as efficacy- or effectiveness-oriented RCTs, but typically feature important distinctions. In this section we will start by reviewing key factors that separate implementation RCTs from more traditional efficacy- or effectiveness-oriented RCTs. We will then discuss optimization trials, which are a type of experimental design that is especially useful for certain implementation science questions.

We will then briefly turn our attention to single subject experimental designs (SSEDs) and on-off-on (ABA) designs.

The first common difference that sets apart implementation RCTs from more traditional clinical trials is the primary research question they aim to address. For most implementation trials, the primary research question is not the extent to which a particular treatment or evidence-based practice is more effective than a comparison condition, but instead the extent to which a given implementation strategy is more effective than a comparison condition. For more detail on this pivotal issue, see Drs. Bauer and Kirchner in this special issue.

Second, as a corollary of this point, implementation RCTs typically feature different outcome measures than efficacy or effectiveness RCTs, with an emphasis on the extent to which a health intervention was successfully implemented rather than an evaluation of the health effects of that intervention (Proctor et al., 2011). For example, typical implementation outcomes might include the number of patients who receive the intervention, or the number of providers who administer the intervention as intended. A variety of evaluation-oriented implementation frameworks may guide the choices of such measures (e.g. RE-AIM; Gaglio et al., 2013; Glasgow et al., 1999). Hybrid implementation-effectiveness studies attend to both effectiveness and implementation outcomes (Curran et al., 2012); these designs are also covered in more detail elsewhere in this issue (Landes, this issue).

Third, given their focus, implementation RCTs are frequently cluster-randomized (i.e. with sites or clinics as the unit of randomization, and patients nested within those sites or clinics). For example, consider a hypothetical RCT that aims to evaluate the implementation of a training program for cognitive behavioral therapy (CBT) in community clinics. Randomizing at the patient level for such a trial would be inappropriate due to the risk of contamination, as providers trained in CBT might reasonably be expected to incorporate CBT principles into their treatment even to patients assigned to the control condition. Randomizing at the provider level would also risk contamination, as providers trained in CBT might discuss this treatment approach with their colleagues. Thus, many implementation trials are cluster randomized at the site or clinic level. While such clustering minimizes the risk of contamination, it can unfortunately create commensurate problems with confounding, especially for trials with very few sites to randomize. Stratification may be used to at least partially address confounding issues in cluster- randomized and more traditional trials alike, by ensuring that intervention and control groups are broadly similar on certain key variables. Furthermore, such allocation schemes typically require analytic models that account for this clustering and the resulting correlations among error structures (e.g., generalized estimating equations [GEE] or mixed-effects models; Schildcrout et al., 2018).

1.1. Optimization trials

Key research questions in implementation science often involve determining which implementation strategies to provide, to whom, and when, to achieve optimal implementation success. As such, trials designed to evaluate comparative effectiveness, or to optimize provision of different types or intensities of implementation strategies, may be more appealing than traditional effectiveness trials. The methods described in this section

are not unique to implementation science, but their application in the context of implementation trials may be particularly useful for informing implementation strategies.

While two-arm RCTs can be used to evaluate comparative effectiveness, trials focused on optimizing implementation support may use alternative experimental designs (Collins et al., 2005; Collins et al., 2007). For example, in certain clinical contexts, multi-component "bundles" of implementation strategies may be warranted (e.g. a bundle consisting of clinician training, technical assistance, and audit/feedback to encourage clinicians to use a new evidence-based practice). In these situations, implementation researchers might consider using factorial or fractional-factorial designs. In the context of implementation science, these designs randomize participants (e.g. sites or providers) to different *combinations* of implementation strategies, and can be used to evaluate the effectiveness of each strategy individually to inform an optimal combination (e.g. Coulton et al., 2009; Pellegrini et al., 2014; Wyrick, et al., 2014). Such designs can be particularly useful in informing multi-component implementation strategies that are not redundant or overly burdensome (Collins et al., 2014a; Collins et al., 2009; Collins et al., 2007).

Researchers interested in optimizing *sequences* of implementation strategies that adapt to ongoing needs over time may be interested in a variant of factorial designs known as the sequential, multiple-assignment randomized trial (SMART; Almirall et al., 2012; Collins et al., 2014b; Kilbourne et al., 2014b; Lei et al., 2012; Nahum-Shani et al., 2012; NeCamp et al., 2017). SMARTs are multistage randomized trials in which some or all participants are randomized more than once, often based on ongoing information (e.g., treatment response). In implementation research, SMARTs can inform optimal sequences of implementation strategies to maximize downstream clinical outcomes. Thus, such designs are well-suited to answering questions about what implementation strategies should be used, *in what order,* to achieve the best outcomes in a given context.

One example of an implementation SMART is the Adaptive Implementation of Effective Program Trial (ADEPT; Kilbourne et al., 2014a). ADEPT was a clustered SMART (NeCamp et al., 2017) designed to inform an adaptive sequence of implementation strategies for implementing an evidence-based collaborative chronic care model, Life Goals (Kilbourne et al., 2014c; Kilbourne et al., 2012a), into community-based practices. Life Goals, the clinical intervention being implemented, has proven effective at improving physical and mental health outcomes for patients with unipolar and bipolar depression by encouraging providers to instruct patients in self-management, and improving clinical information systems and care management across physical and mental health providers (Bauer et al., 2006; Kilbourne et al., 2012a; Kilbourne et al., 2008; Simon et al., 2006). However, in spite of its established clinical effectiveness, community-based clinics experienced a number of barriers in trying to implement the Life Goals model, and there were questions about how best to efficiently and effectively augment implementation strategies for clinics that struggled with implementation.

The ADEPT study was thus designed to determine the best *sequence* of implementation strategies to offer sites interested in implementing Life Goals. The ADEPT study involved use of three different implementation strategies. First, all sites received implementation

support based on Replicating Effective Programs (REP), which offered an implementation manual, brief training, and low- level technical support (Kilbourne et al., 2007; Kilbourne et al., 2012b; Neumann and Sogolow, 2000). REP implementation support had been previously found to be low-cost and readily scalable, but also insufficient for uptake for many community-based settings (Kilbourne et al., 2015). For sites that failed to implement Life Goals under REP, two additional implementation strategies were considered as augmentations to REP: External Facilitation (EF; Kilbourne et al., 2014b; Stetler et al., 2006), consisting of phone-based mentoring in strategic skills from a study team member; and Internal Facilitation (IF; Kirchner et al., 2014), which supported protected time for a site employee to address barriers to program adoption.

The ADEPT study was designed to evaluate the best way to augment support for these sites that were not able to implement Life Goals under REP, specifically querying whether it was better to augment REP with EF only or the more intensive EF/IF, and whether augmentations should be provided all at once, or staged. Intervention assignments are mapped in Figure 1. Seventy-nine community-based clinics across Michigan and Colorado were provided with initial implementation support under REP. After six months, implementation of the clinical intervention, Life Goals, was evaluated at all sites. Sites that had failed to reach an adequate level of delivery (defined as those sites enrolling fewer than ten patients in Life Goals, or those at which fewer than 50% of enrolled patients had received at least three Life Goals sessions) were considered non-responsive to REP and randomized to receive additional support through either EF or combined EF/IF. After six further months, Life Goals implementation at these sites was again evaluated. Sites surpassing the implementation response benchmark had their EF or EF/IF support discontinued. EF/IF sites that remained non-responsive continued to receive EF/IF for an additional six months. EF sites that remained non-responsive were randomized a second time to either continue with EF or further augment with IF. This design thus allowed for comparison of three different adaptive implementation interventions for sites that were initially non-responsive to REP to determine the best adaptive sequence of implementation support for sites that were initially non-responsive under REP:

- Provide EF for 6 months; continue EF for a further six months for sites that remain nonresponsive; discontinue EF for sites that are responsive;
- Provide EF/IF for 6 months; continue EF/IF for a further six months for sites that remain non-responsive; discontinue EF/IF for sites that are responsive; and
- Provide EF for 6 months; step up to EF/IF for a further six months for sites that remain non-responsive; discontinue EF for sites that are responsive.

While analyses of this study are still ongoing, including the comparison of these three adaptive sequences of implementation strategies, results have shown that patients at sites that were randomized to receive EF as the initial augmentation to REP saw more improvement in clinical outcomes (SF-12 mental health quality of life and PHQ-9 depression scores) after 12 months than patients at sites that were randomized to receive the more intensive EF/IF augmentation.

1.2. Single Subject Experimental Designs and On-Off-On (ABA) Designs

We also note that there are a variety of Single Subject Experimental Designs (SSEDs; Byiers et al., 2012), including withdrawal designs and alternating treatment designs, that can be used in testing evidence-based practices. Similarly, an implementation strategy may be used to encourage the use of a specific treatment at a particular site, followed by that strategy's withdrawal and subsequent reinstatement, with data collection throughout the process (on-off-on or ABA design). A weakness of these approaches in the context of implementation science, however, is that they usually require reversibility of the intervention (i.e. that the withdrawal of implementation support truly allows the healthcare system to revert to its pre-implementation state). When this is not the case—for example, if a hypothetical study is focused on training to encourage use of an evidence-based psychotherapy—then these designs may be less useful.

2. Quasi-Experimental Designs in Implementation Science

In some implementation science contexts, policy-makers or administrators may not be willing to have a subset of participating patients or sites randomized to a control condition, especially for high-profile or high-urgency clinical issues. Quasi-experimental designs allow implementation scientists to conduct rigorous studies in these contexts, albeit with certain limitations. We briefly review the characteristics of these designs here; other recent review articles are available for the interested reader (e.g. Handley et al., 2018).

2.1. Pre-Post with Non-Equivalent Control Group

The pre-post with non-equivalent control group uses a control group in the absence of randomization. Ideally, the control group is chosen to be as similar to the intervention group as possible (e.g. by matching on factors such as clinic type, patient population, geographic region, etc.). Theoretically, both groups are exposed to the same trends in the environment, making it plausible to decipher if the intervention had an effect. Measurement of both treatment and control conditions classically occurs pre- and post-intervention, with differential improvement between the groups attributed to the intervention. This design is popular due to its practicality, especially if data collection points can be kept to a minimum. It may be especially useful for capitalizing on naturally occurring experiments such as may occur in the context of certain policy initiatives or rollouts—specifically, rollouts in which it is plausible that a control group can be identified. For example, Kirchner and colleagues (2014) used this type of design to evaluate the integration of mental health services into primary care clinics at seven US Department of Veterans Affairs (VA) medical centers and seven matched controls.

One overarching drawback of this design is that it is especially vulnerable to threats to internal validity (Shadish, 2002), because pre-existing differences between the treatment and control group could erroneously be attributed to the intervention. While unmeasured differences between treatment and control groups are always a possibility in healthcare research, such differences are especially likely to occur in the context of these designs due to the lack of randomization. Similarly, this design is particularly sensitive to secular trends that may differentially affect the treatment and control groups (Cousins et al., 2014; Pape et

al., 2013), as well as regression to the mean confounding study results (Morton and Torgerson, 2003). For example, if a study site is selected for the experimental condition precisely because it is underperforming in some way, then regression to the mean would suggest that the site will show improvement regardless of any intervention; in the context of a pre-post with non-equivalent control group study, however, this improvement would erroneously be attributed to the intervention itself (Type I error).

There are, however, various ways that implementation scientists can mitigate these weaknesses. First, as mentioned briefly above, it is important to select a control group that is as similar as possible to the intervention site(s), which can include matching at both the health care network and clinic level (e.g. Kirchner et al., 2014). Second, propensity score weighting (e.g. Morgan, 2018) can statistically mitigate internal validity concerns, although this approach may be of limited utility when comparing secular trends between different study cohorts (Dimick and Ryan, 2014). More broadly, qualitative methods (e.g. periodic interviews with staff at intervention and control sites) can help uncover key contextual factors that may be affecting study results above and beyond the intervention itself.

2.2. Interrupted Time Series

Interrupted time series (ITS; Shadish, 2002; Taljaard et al., 2014; Wagner et al., 2002) designs represent one of the most robust categories of quasi-experimental designs. Rather than relying on a non-equivalent control group, ITS designs rely on repeated data collections from intervention sites to determine whether a particular intervention is associated with improvement on a given metric relative to the pre-intervention secular trend. They are particularly useful in cases where a comparable control group cannot be identified—for example, following widespread implementation of policy mandates, quality improvement initiatives, or dissemination campaigns (Eccles et al., 2003). In ITS designs, data are collected at multiple time points both before and after an intervention (e.g., policy change, implementation effort), and analyses explore whether the intervention was associated with the outcome beyond any pre-existing secular trend. More formally, ITS evaluations focus on identifying whether there is discontinuity in the trend (change in slope or level) after the intervention relative to before the intervention, using segmented regression to model pre- and post-intervention trends (Gebski et al., 2012; Penfold and Zhang, 2013; Taljaard et al., 2014; Wagner et al., 2002). A number of recent implementation studies have used ITS designs, including an evaluation of implementation of a comprehensive smoke-free policy in a large UK mental health organization to reduce physical assaults (Robson et al., 2017); the impact of a national policy limiting alcohol availability on suicide mortality in Slovenia (Pridemore and Snowden, 2009); and the effect of delivery of a tailored intervention for primary care providers to increase psychological referrals for women with mild to moderate postnatal depression (Hanbury et al., 2013).

ITS designs are appealing in implementation work for several reasons. Relative to uncontrolled pre-post analyses, ITS analyses reduce the chances that intervention effects are confounded by secular trends (Bernal et al., 2017; Eccles et al., 2003). Time-varying confounders, such as seasonality, can also be adjusted for, provided adequate data (Bernal et al., 2017). Indeed, recent work has confirmed that ITS designs can yield effect estimates

similar to those derived from cluster-randomized RCTs (Fretheim et al., 2013; Fretheim et al., 2015). Relative to an RCT, ITS designs can also allow for a more comprehensive assessment of the longitudinal effects of an intervention (positive or negative), as effects can be traced over all included time points (Bernal et al., 2017; Penfold and Zhang, 2013).

ITS designs also present a number of challenges. First, the segmented regression approach requires clear delineation between pre- and post-intervention periods; interventions with indeterminate implementation periods are likely not good candidates for ITS. While ITS designs that include multiple 'interruptions' (e.g. introductions of new treatment components) are possible, they will require collection of enough time points between interruptions to ensure that each intervention's effects can be ascertained individually (Bernal et al., 2017). Second, collecting data from sufficient time points across all sites of interest, especially for the pre-intervention period, can be challenging (Eccles et al., 2003): a common recommendation is at least eight time points both pre- and post-intervention (Penfold and Zhang, 2013). This may be onerous, particularly if the data are not routinely collected by the health system(s) under study. Third, ITS cannot protect against confounding effects from other interventions that begin contemporaneously and may impact similar outcomes (Eccles et al., 2003).

2.3. Stepped Wedge Designs

Stepped wedge trials are another type of quasi-experimental design. In a stepped wedge, all participants receive the intervention, but are assigned to the timing of the intervention in a staggered fashion (Betran et al., 2018; Brown and Lilford, 2006; Hussey and Hughes, 2007), typically at the site or cluster level. Stepped wedge designs have their analytic roots in balanced incomplete block designs, in which all pairs of treatments occur an equal number of times within each block (Hanani, 1961). Traditionally, all sites in stepped wedge trials have outcome measures assessed at all time points, thus allowing sites that receive the intervention later in the trial to essentially serve as controls for early intervention sites. A recent special issue of the journal *Trials* includes more detail on these designs (Davey et al., 2015), which may be ideal for situations in which it is important for all participating patients or sites to receive the intervention during the trial. Stepped wedge trials may also be useful when resources are scarce enough that intervening at all sites at once (or even half of the sites as in a standard treatment-versus-control RCT) would not be feasible. If desired, the administration of the intervention to sites in waves allows for lessons learned in early sites to be applied to later sites (via formative evaluation; see Elwy et al., this issue).

The Behavioral Health Interdisciplinary Program (BHIP) Enhancement Project is a recent example of a stepped-wedge implementation trial (Bauer et al., 2016; Bauer et al., 2019). This study involved using blended facilitation (including internal and external facilitators; Kirchner et al., 2014) to implement care practices consistent with the collaborative chronic care model (CCM; Bodenheimer et al., 2002a, b; Wagner et al., 1996) in nine outpatient mental health teams in VA medical centers. Figure 2 illustrates the implementation and stepdown periods for that trial, with black dots representing primary data collection points.

The BHIP Enhancement Project was conducted as a stepped wedge for several reasons. First, the stepped wedge design allowed the trial to reach nine sites despite limited

implementation resources (i.e. intervening at all nine sites simultaneously would not have been feasible given study funding). Second, the stepped wedge design aided in recruitment and retention, as all participating sites were certain to receive implementation support during the trial: at worst, sites that were randomized to later- phase implementation had to endure waiting periods totaling about eight months before implementation began. This was seen as a major strength of the design by its operational partner, the VA Office of Mental Health and Suicide Prevention. To keep sites engaged during the waiting period, the BHIP Enhancement Project offered a guiding workbook and monthly technical support conference calls.

Three additional features of the BHIP Enhancement Project deserve special attention. First, data collection for late-implementing sites did not begin until immediately before the onset of implementation support (see Figure 2). While this reduced statistical power, it also significantly reduced data collection burden on the study team. Second, onset of implementation support was staggered such that wave 2 began at the end of month 4 rather than month 6. This had two benefits: first, this compressed the overall amount of time required for implementation during the trial. Second, it meant that the study team only had to collect data from one site at a time, with data collection periods coming every 2–4 months. More traditional stepped wedge approaches typically have data collection across sites temporally aligned (e.g. Betran et al., 2018). Third, the BHIP Enhancement Project used a balancing algorithm (Lew et al., 2019) to assign sites to waves, retaining some of the benefits of randomization while ensuring balance on key site characteristics (e.g. size, geographic region).

Despite their utility, stepped wedges have some important limitations. First, because they feature delayed implementation at some sites, stepped wedges typically take longer than similarly-sized parallel group RCTs. This increases the chances that secular trends, policy changes, or other external forces impact study results. Second, as with RCTs, imbalanced site assignment can confound results. This may occur deliberately in some cases—for example, if sites that develop their implementation plans first are assigned to earlier waves. Even if sites are randomized, however, early and late wave sites may still differ on important characteristics such as size, rurality, and case mix. The resulting confounding between site assignment and time can threaten the internal validity of the study—although, as above, balancing algorithms can reduce this risk. Third, the use of formative evaluation (Elwy, this issue), while useful for maximizing the utility of implementation efforts in a stepped wedge, can mean that late-wave sites receive different implementation strategies than early-wave sites. Similarly, formative evaluation may inform midstream adaptations to the clinical innovation being implemented. In either case, these changes may again threaten internal validity. Overall, then, stepped wedges represent useful tools for evaluating the impact of health interventions that (as with all designs) are subject to certain weaknesses and limitations.

3. Conclusions and Future Directions

Implementation science is focused on maximizing the extent to which effective healthcare practices are adopted, used, and sustained by clinicians, hospitals, and systems. Answering questions in these domains frequently requires different research methods than those

employed in traditional efficacy- or effectiveness-oriented randomized clinical trials (RCTs). Implementation-oriented RCTs typically feature cluster or site-level randomization, and emphasize implementation outcomes (e.g. the number of patients receiving the new treatment as intended) rather than traditional clinical outcomes. Hybrid implementation-effectiveness designs incorporate both types of outcomes; more details on these approaches can be found elsewhere in this special issue (Landes, this issue). Other methodological innovations, such as factorial designs or sequential, multiple-assignment randomized trials (SMARTs), can address questions about multi-component or adaptive interventions, still under the umbrella of experimental designs. These types of trials may be especially important for demystifying the "black box" of implementation—that is, determining what components of an implementation strategy are most strongly associated with implementation success. In contrast, pre-post designs with non-equivalent control groups, interrupted time series (ITS), and stepped wedge designs are all examples of quasiexperimental designs that may serve implementation researchers when experimental designs would be inappropriate. A major theme cutting across each of these designs is that there are relative strengths and weaknesses associated with any study design decision. Determining what design to use ultimately will need to be informed by the primary research question to be answered, while simultaneously balancing the need for internal validity, external validity, feasibility, and ethics.

New innovations in study design are constantly being developed and refined. Several such innovations are covered in other articles within this special issue (e.g. Kim et al., this issue). One future direction relevant to the study designs presented in this article is the potential for adaptive trial designs, which allow information gleaned during the trial to inform the adaptation of components like treatment allocation, sample size, or study recruitment in the later phases of the same trial (Pallmann et al., 2018). These designs are becoming increasingly popular in clinical treatment (Bhatt and Mehta, 2016) but could also hold promise for implementation scientists, especially as interest grows in rapid-cycle testing of implementation strategies or efforts. Adaptive designs could potentially be incorporated into both SMART designs and stepped wedge studies, as well as traditional RCTs to further advance implementation science (Cheung et al., 2015). Ideally, these and other innovations will provide researchers with increasingly robust and useful methodologies for answering timely implementation science questions.

Acknowledgments

Funding

This work was supported by Department of Veterans Affairs grants QUE 15–289 (PI: Bauer) and CIN 13403 and National Institutes of Health grant RO1 MH 099898 (PI: Kilbourne).

References

Almirall D, Compton SN, Gunlicks-Stoessel M, Duan N, Murphy SA, 2012 Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. Stat Med 31 (17), 1887–1902. [PubMed: 22438190]

Bauer MS, McBride L, Williford WO, Glick H, Kinosian B, Altshuler L, Beresford T, Kilbourne AM, Sajatovic M, Cooperative Studies Program 430 Study, T., 2006 Collaborative care for bipolar

disorder: Part II. Impact on clinical outcome, function, and costs. Psychiatr Serv 57 (7), 937–945. [PubMed: 16816277]

Bauer MS, Miller C, Kim B, Lew R, Weaver K, Coldwell C, Henderson K, Holmes S, Seibert MN, Stolzmann K, Elwy AR, Kirchner J, 2016 Partnering with health system operations leadership to develop a controlled implementation trial. Implement Sci 11, 22. [PubMed: 26912342]

Bauer MS, Miller CJ, Kim B, Lew R, Stolzmann K, Sullivan J, Riendeau R, Pitcock J, Williamson A, Connolly S, Elwy AR, Weaver K, 2019 Effectiveness of Implementing a Collaborative Chronic Care Model for Clinician Teams on Patient Outcomes and Health Status in Mental Health: A Randomized Clinical Trial. JAMA Netw Open 2 (3), e190230.

Bernal JL, Cummins S, Gasparrini A, 2017 Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 46 (1), 348–355. [PubMed: 27283160]

Betran AP, Bergel E, Griffin S, Melo A, Nguyen MH, Carbonell A, Mondlane S, Merialdi M, Temmerman M, Gulmezoglu AM, 2018 Provision of medical supply kits to improve quality of antenatal care in Mozambique: a stepped-wedge cluster randomised trial. Lancet Glob Health 6 (1), e57–e65. [PubMed: 29241615]

Bhatt DL, Mehta C, 2016 Adaptive Designs for Clinical Trials. N Engl J Med 375 (1), 65–74. [PubMed: 27406349]

Bodenheimer T, Wagner EH, Grumbach K, 2002a Improving primary care for patients with chronic illness. JAMA 288 (14), 1775–1779. [PubMed: 12365965]

Bodenheimer T, Wagner EH, Grumbach K, 2002b Improving primary care for patients with chronic illness: the chronic care model, Part 2. JAMA 288 (15), 1909–1914. [PubMed: 12377092]

Brown CA, Lilford RJ, 2006 The stepped wedge trial design: a systematic review. BMC medical research methodology 6 (1), 54. [PubMed: 17092344]

Byiers BJ, Reichle J, Symons FJ, 2012 Single-subject experimental design for evidence-based practice. Am J Speech Lang Pathol 21 (4), 397–414. [PubMed: 23071200]

Cheung YK, Chakraborty B, Davidson KW, 2015 Sequential multiple assignment randomized trial (SMART) with adaptive randomization for quality improvement in depression treatment program. Biometrics 71 (2), 450–459. [PubMed: 25354029]

Collins LM, Dziak JJ, Kugler KC, Trail JB, 2014a Factorial experiments: efficient tools for evaluation of intervention components. Am J Prev Med 47 (4), 498–504. [PubMed: 25092122]

Collins LM, Dziak JJ, Li R, 2009 Design of experiments with multiple independent variables: a resource management perspective on complete and reduced factorial designs. Psychol Methods 14 (3), 202–224. [PubMed: 19719358]

Collins LM, Murphy SA, Bierman KL, 2004 A conceptual framework for adaptive preventive interventions. Prev Sci 5 (3), 185–196. [PubMed: 15470938]

Collins LM, Murphy SA, Nair VN, Strecher VJ, 2005 A strategy for optimizing and evaluating behavioral interventions. Ann Behav Med 30 (1), 65–73. [PubMed: 16097907]

Collins LM, Murphy SA, Strecher V, 2007 The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med 32 (5 Suppl), S112–118. [PubMed: 17466815]

Collins LM, Nahum-Shani I, Almirall D, 2014b Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assignment, randomized trial (SMART). Clin Trials 11 (4), 426–434. [PubMed: 24902922]

Coulton S, Perryman K, Bland M, Cassidy P, Crawford M, Deluca P, Drummond C, Gilvarry E, Godfrey C, Heather N, Kaner E, Myles J, Newbury-Birch D, Oyefeso A, Parrott S, Phillips T, Shenker D, Shepherd J, 2009 Screening and brief interventions for hazardous alcohol use in accident and emergency departments: a randomised controlled trial protocol. BMC Health Serv Res 9, 114. [PubMed: 19575791]

Cousins K, Connor JL, Kypri K, 2014 Effects of the Campus Watch intervention on alcohol consumption and related harm in a university population. Drug Alcohol Depend 143, 120–126. [PubMed: 25108584]

Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C, 2012 Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 50 (3), 217–226. [PubMed: 22310560]

Davey C, Hargreaves J, Thompson JA, Copas AJ, Beard E, Lewis JJ, Fielding KL, 2015 Analysis and reporting of stepped wedge randomised controlled trials: synthesis and critical appraisal of published studies, 2010 to 2014. Trials 16 (1), 358. [PubMed: 26278667]

Dimick JB, Ryan AM, 2014 Methods for evaluating changes in health care policy: the difference-in-differences approach. JAMA 312 (22), 2401–2402. [PubMed: 25490331]

Eccles M, Grimshaw J, Campbell M, Ramsay C, 2003 Research designs for studies evaluating the effectiveness of change and improvement strategies. Qual Saf Health Care 12 (1), 47–52. [PubMed: 12571345]

Fisher RA, 1925, July Theory of statistical estimation In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 22, No. 5, pp. 700–725). Cambridge University Press.

Fisher RA, 1935 The design of experiments. Oliver and Boyd, Edinburgh.

Fretheim A, Soumerai SB, Zhang F, Oxman AD, Ross-Degnan D, 2013 Interrupted time-series analysis yielded an effect estimate concordant with the cluster-randomized controlled trial result. Journal of Clinical Epidemiology 66 (8), 883–887. [PubMed: 23810027]

Fretheim A, Zhang F, Ross-Degnan D, Oxman AD, Cheyne H, Foy R, Goodacre S, Herrin J, Kerse N, McKinlay RJ, Wright A, Soumerai SB, 2015 A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation. J Clin Epidemiol 68 (3), 324–333. [PubMed: 25499983]

Gaglio B, Shoup JA, Glasgow RE, 2013 The RE-AIM framework: a systematic review of use over time. Am J Public Health 103 (6), e38–46.

Gebski V, Ellingson K, Edwards J, Jernigan J, Kleinbaum D, 2012 Modelling interrupted time series to evaluate prevention and control of infection in healthcare. Epidemiol Infect 140 (12), 2131–2141. [PubMed: 22335933]

Glasgow RE, Vogt TM, Boles SM, 1999 Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 89 (9), 1322–1327. [PubMed: 10474547]

Hanani H, 1961 The existence and construction of balanced incomplete block designs. The Annals of Mathematical Statistics 32 (2), 361–386.

Hanbury A, Farley K, Thompson C, Wilson PM, Chambers D, Holmes H, 2013 Immediate versus sustained effects: interrupted time series analysis of a tailored intervention. Implement Sci 8, 130. [PubMed: 24188718]

Handley MA, Lyles CR, McCulloch C, Cattamanchi A, 2018 Selecting and Improving Quasi-Experimental Designs in Effectiveness and Implementation Research. Annu Rev Public Health 39, 5–25. [PubMed: 29328873]

Hussey MA, Hughes JP, 2007 Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials 28 (2), 182–191. [PubMed: 16829207]

Kilbourne AM, Almirall D, Eisenberg D, Waxmonsky J, Goodrich DE, Fortney JC, Kirchner JE, Solberg LI, Main D, Bauer MS, Kyle J, Murphy SA, Nord KM, Thomas MR, 2014a Protocol: Adaptive Implementation of Effective Programs Trial (ADEPT): cluster randomized SMART trial comparing a standard versus enhanced implementation strategy to improve outcomes of a mood disorders program. Implement Sci 9, 132. [PubMed: 25267385]

Kilbourne AM, Almirall D, Goodrich DE, Lai Z, Abraham KM, Nord KM, Bowersox NW, 2014b Enhancing outreach for persons with serious mental illness: 12-month results from a cluster randomized trial of an adaptive implementation strategy. Implement Sci 9, 163. [PubMed: 25544027]

Kilbourne AM, Bramlet M, Barbaresso MM, Nord KM, Goodrich DE, Lai Z, Post EP, Almirall D, Verchinina L, Duffy SA, Bauer MS, 2014c SMI life goals: description of a randomized trial of a collaborative care model to improve outcomes for persons with serious mental illness. Contemp Clin Trials 39 (1), 74–85. [PubMed: 25083802]

Kilbourne AM, Goodrich DE, Lai Z, Clogston J, Waxmonsky J, Bauer MS, 2012a Life Goals Collaborative Care for patients with bipolar disorder and cardiovascular disease risk. Psychiatr Serv 63 (12), 1234–1238. [PubMed: 23203358]

Kilbourne AM, Goodrich DE, Nord KM, Van Poppelen C, Kyle J, Bauer MS, Waxmonsky JA, Lai Z, Kim HM, Eisenberg D, Thomas MR, 2015 Long-Term Clinical Outcomes from a Randomized

Controlled Trial of Two Implementation Strategies to Promote Collaborative Care Attendance in Community Practices. Adm Policy Ment Health 42 (5), 642–653. [PubMed: 25315181]

Kilbourne AM, Neumann MS, Pincus HA, Bauer MS, Stall R, 2007 Implementing evidence-based interventions in health care: application of the replicating effective programs framework. Implement Sci 2, 42. [PubMed: 18067681]

Kilbourne AM, Neumann MS, Waxmonsky J, Bauer MS, Kim HM, Pincus HA, Thomas M, 2012b Public-academic partnerships: evidence-based implementation: the role of sustained community-based practice and research partnerships. Psychiatr Serv 63 (3), 205–207. [PubMed: 22388527]

Kilbourne AM, Post EP, Nossek A, Drill L, Cooley S, Bauer MS, 2008 Improving medical and psychiatric outcomes among individuals with bipolar disorder: a randomized controlled trial. Psychiatr Serv 59 (7), 760–768. [PubMed: 18586993]

Kirchner JE, Ritchie MJ, Pitcock JA, Parker LE, Curran GM, Fortney JC, 2014 Outcomes of a partnered facilitation strategy to implement primary care-mental health. J Gen Intern Med 29 Suppl 4, 904–912. [PubMed: 25355087]

Lei H, Nahum-Shani I, Lynch K, Oslin D, Murphy SA, 2012 A "SMART" design for building individualized treatment sequences. Annu Rev Clin Psychol 8, 21–48. [PubMed: 22224838]

Lew RA, Miller CJ, Kim B, Wu H, Stolzmann K, Bauer MS, 2019 A robust method to reduce imbalance for site-level randomized controlled implementation trial designs. Implementation Sci, 14, 46.

Morgan CJ, 2018 Reducing bias using propensity score matching. J Nucl Cardiol 25 (2), 404–406. [PubMed: 28776312]

Morton V, Torgerson DJ, 2003 Effect of regression to the mean on decision making in health care. BMJ 326 (7398), 1083–1084. [PubMed: 12750214]

Nahum-Shani I, Qian M, Almirall D, Pelham WE, Gnagy B, Fabiano GA, Waxmonsky JG, Yu J, Murphy SA, 2012 Experimental design and primary data analysis methods for comparing adaptive interventions. Psychol Methods 17 (4), 457–477. [PubMed: 23025433]

NeCamp T, Kilbourne A, Almirall D, 2017 Comparing cluster-level dynamic treatment regimens using sequential, multiple assignment, randomized trials: Regression estimation and sample size considerations. Stat Methods Med Res 26 (4), 1572–1589. [PubMed: 28627310]

Neumann MS, Sogolow ED, 2000 Replicating effective programs: HIV/AIDS prevention technology transfer. AIDS Educ Prev 12 (5 Suppl), 35–48. [PubMed: 11063068]

Pallmann P, Bedding AW, Choodari-Oskooei B, Dimairo M, Flight L, Hampson LV, Holmes J, Mander AP, Odondi L.o., Sydes MR, Villar SS, Wason JMS, Weir CJ, Wheeler GM, Yap C, Jaki T, 2018 Adaptive designs in clinical trials: why use them, and how to run and report them. BMC medicine 16 (1), 29–29. [PubMed: 29490655]

Pape UJ, Millett C, Lee JT, Car J, Majeed A, 2013 Disentangling secular trends and policy impacts in health studies: use of interrupted time series analysis. J R Soc Med 106 (4), 124–129. [PubMed: 23564896]

Pellegrini CA, Hoffman SA, Collins LM, Spring B, 2014 Optimization of remotely delivered intensive lifestyle treatment for obesity using the Multiphase Optimization Strategy: Opt-IN study protocol. Contemp Clin Trials 38 (2), 251–259. [PubMed: 24846621]

Penfold RB, Zhang F, 2013 Use of Interrupted Time Series Analysis in Evaluating Health Care Quality Improvements. Academic Pediatrics 13 (6, Supplement), S38–S44. [PubMed: 24268083]

Pridemore WA, Snowden AJ, 2009 Reduction in suicide mortality following a new national alcohol policy in Slovenia: an interrupted time-series analysis. Am J Public Health 99 (5), 915–920. [PubMed: 19299669]

Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, Griffey R, Hensley M, 2011 Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 38 (2), 65–76. [PubMed: 20957426]

Robson D, Spaducci G, McNeill A, Stewart D, Craig TJK, Yates M, Szatkowski L, 2017 Effect of implementation of a smoke-free policy on physical violence in a psychiatric inpatient setting: an interrupted time series analysis. Lancet Psychiatry 4 (7), 540–546. [PubMed: 28624180]

Schildcrout JS, Schisterman EF, Mercaldo ND, Rathouz PJ, Heagerty PJ, 2018 Extending the Case-Control Design to Longitudinal Data: Stratified Sampling Based on Repeated Binary Outcomes. Epidemiology 29 (1), 67–75. [PubMed: 29068838]

Shadish WR, Cook Thomas D., Campbell Donald T, 2002 Experimental and quasi-experimental designs for generalized causal inference. Houghton Miffflin Company, Boston, MA.

Simon GE, Ludman EJ, Bauer MS, Unutzer J, Operskalski B, 2006 Long-term effectiveness and cost of a systematic care program for bipolar disorder. Arch Gen Psychiatry 63 (5), 500–508. [PubMed: 16651507]

Stetler CB, Legro MW, Rycroft-Malone J, Bowman C, Curran G, Guihan M, Hagedorn H, Pineros S, Wallace CM, 2006 Role of "external facilitation" in implementation of research findings: a qualitative evaluation of facilitation experiences in the Veterans Health Administration. Implement Sci 1, 23. [PubMed: 17049080]

Taljaard M, McKenzie JE, Ramsay CR, Grimshaw JM, 2014 The use of segmented regression in analysing interrupted time series studies: an example in pre-hospital ambulance care. Implement Sci 9, 77. [PubMed: 24943919]

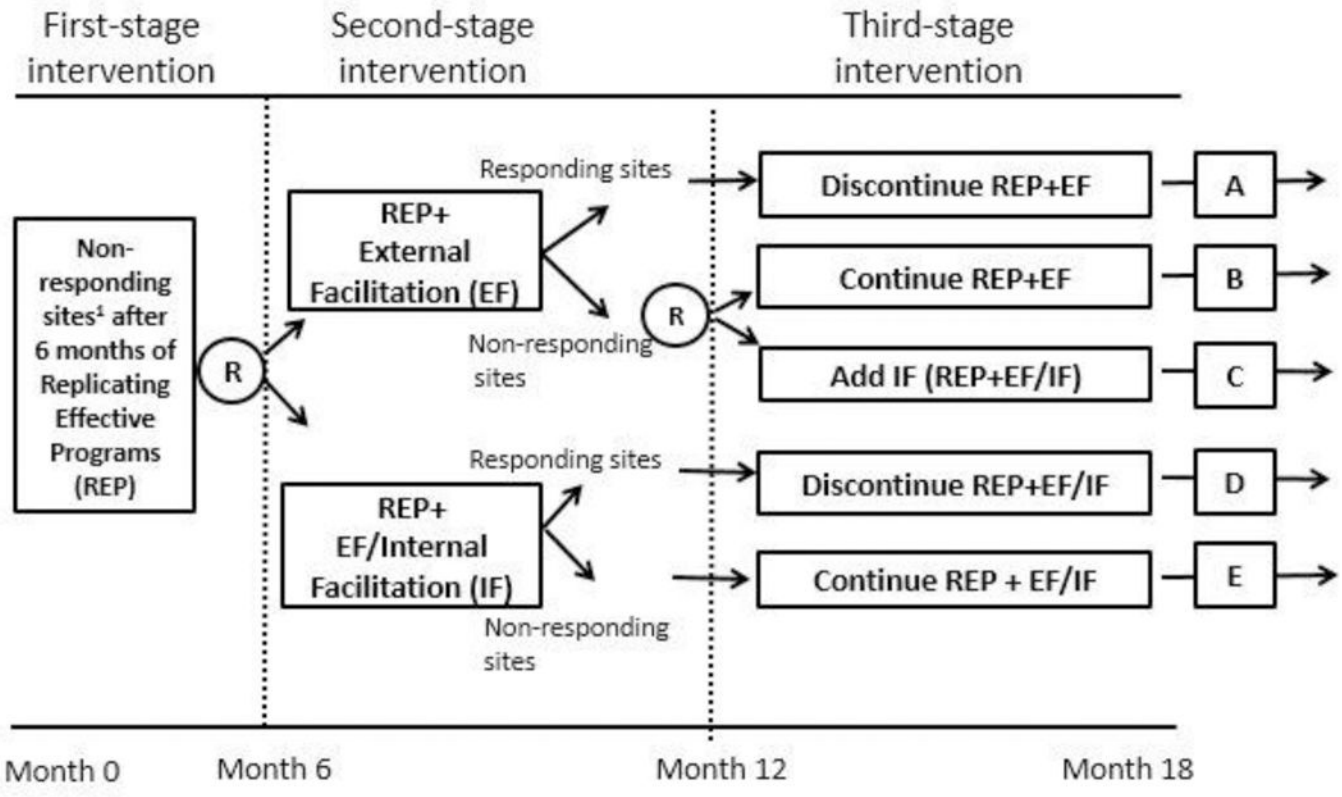
Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D, 2002 Segmented regression analysis of interrupted time series studies in medication use research. J Clin Pharm Ther 27 (4), 299–309. [PubMed: 12174032]

Wagner EH, Austin BT, Von Korff M, 1996 Organizing care for patients with chronic illness. Milbank Q 74 (4), 511–544. [PubMed: 8941260]

Wyrick DL, Rulison KL, Fearnow-Kenney M, Milroy JJ, Collins LM, 2014 Moving beyond the treatment package approach to developing behavioral interventions: addressing questions that arose during an application of the Multiphase Optimization Strategy (MOST). Transl Behav Med 4 (3), 252–259. [PubMed: 25264465]

Highlights

- Many implementation science questions can be addressed by fully experimental designs (e.g. randomized controlled trials [RCTs]).
- Implementation trials differ in important ways, however, from more traditional efficacy- or effectiveness-oriented RCTs.
- Adaptive designs represent a recent innovation to determine optimal implementation strategies within a fully experimental framework.
- Quasi-experimental designs can be used to answer implementation science questions in the absence of randomization.
- The choice of study designs in implementation science requires careful consideration of scientific, pragmatic, and ethical issues.



[1]Response/non-response: Sites were considered non-responsive at both 6 months and 12 months if either <10 patients were receiving Life Goal or <50% of patients receiving life goals had received ≤ 3 sessions.

Fig. 1.
SMART design from ADEPT trial.

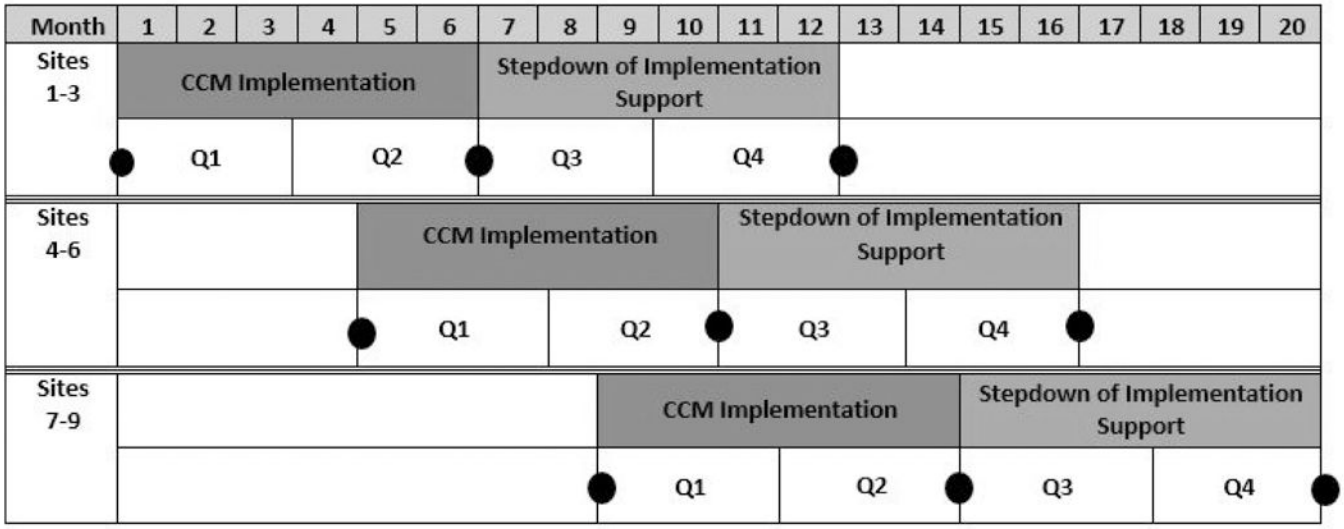


Fig. 2.
BHIP Enhancement Project stepped wedge (adapted form Bauer et al., 2019).