INTRODUCTION
Dissemination and implementation research has evolved into an emerging field, implementation science, as exemplified by the launching of the journal *Implementation Science* in 2006, and the annual Dissemination and Implementation conferences, initiated in 2007. Not surprisingly, progress in this emerging science is uneven, with a greater volume of empirical studies mounted in the physical health care field than in the fields of mental health, substance use, or social services, due largely to medicine’s early focus on quality of care and, more recently, on comparative effectiveness. However, despite these important scientific efforts, there remains a lack of consensus on methodological approaches to the study of dissemination and implementation processes and especially tests of implementation strategies.\(^1\) To begin to address these deficiencies, this chapter reviews design issues for dissemination and implementation research, and also presents an overview of some of the analytic approaches to dissemination and implementation research, recognizing that this analytic work is still at an early stage of development. Finally, the chapter presents a case study of research that crosses three of four recognized dissemination and implementation phases and illustrates a number of design and analysis issues.

Note the following major changes in the 2nd edition as compared with the 1st edition. In the 2nd edition, the discussion of mixed methods designs has been removed and integrated in a new 2nd edition chapter 20 on mixed methods evaluation in dissemination and implementation science. Similarly, the discussion of system science approaches, specifically, system dynamics method and agent-based modeling, has been removed from this chapter and integrated with social network analysis into chapter 10. However, the role of general simulation in designing D&I studies has been retained and expanded. Third, a section on hybrid designs has been added in this 2nd edition chapter. Fourth, the concept of preimplementation designs has been linked with simulation modeling related to the exploration and adoption/preparation stages of D&I research. Finally, this chapter incorporates the classification of D&I designs recently put forth by Brown and colleagues,\(^2\) which encompasses three overarching categories of designs based on the type of comparison.

FOCUS OF DISSEMINATION AND IMPLEMENTATION RESEARCH
A useful organizing heuristic is to conceptualize dissemination and implementation studies (D&I) in relation to two other stages of research, efficacy and effectiveness. Nicely captured in the 2009 National Research Council and Institute of Medicine report on *Preventing Mental, Emotional, and Behavioral Disorders Among Young People* \(^3\) (shown in Figure 13.1) and adapted from that report and a recent typology;\(^3\) D&I studies are the last stage of research in the science to practice continuum, preceded by efficacy and effectiveness studies that are distinct
Dissemination and Implementation Research in Health

from and address different questions from D&I studies. The figure also demonstrates that distinct phases (albeit somewhat overlapping) exist within the D&I stage, characterized as exploration, adoption/preparation, implementation, and sustainment similar to the EPIS model proposed by Aarons and colleagues.\(^1\) In the exploratory phase, we focus on preimplementation factors including deciding on what evidence-based intervention would be most appropriate. In the adoption/preparation phase, we are interested in factors related to the formal decision to implement, or strategies to increase adoption of an intervention or program. The next phase is implementation (or implementation fidelity), which involves strategies for improving program fidelity in the field, and finally sustainment (and moving to scale), involving strategies to maintain delivery of the intervention or extend its use in communities or organizations. D&I trials are also distinct from efficacy and effectiveness trials as it concerns the independent variable that is manipulated, which are referred to as implementation strategies.

Note that this research model typology represented in Figure 13.1 is also reflected in the NIH Roadmap initiative for re-engineering the clinical research enterprise currently driving the translational research initiative at the NIH.\(^5-7\) The Roadmap initiative has identified three types of research leading to improvements in the public health of our nation, namely, basic research that informs the development of clinical interventions (e.g., biochemistry, neurosciences), treatment development that crafts the interventions and tests them in carefully controlled efficacy trials, and what has come to be known as service system and implementation research, where treatments and interventions are brought into and tested in usual care settings.\(^8\) Based on this tripartite division, the Roadmap further identified two translation steps that would be critical for moving from the findings of basic science to improvements in the quality of health care delivered in community, clinical, and other delivery settings. The first translation step brings together interdisciplinary teams that integrate the science work being done in the basic sciences and treatment development.

\[\text{FIGURE 13.1 Stages of research and phases of dissemination and implementation.}\]

Design and Analysis in Dissemination and Implementation Research

Science, such as translating neuroscience and basic behavior research findings into new treatments. The focus of the second translation phase is to translate evidence-based treatments into service delivery settings and sectors in local communities, and it is this second step that we identify as the D&I research enterprise.\(^7\)

Next, briefly defined and described are three facets of D&I research in order to better frame the discussion of evaluation designs and related analytic issues covered in the remainder of this chapter.

**Implementation Strategies**

Central to understanding the distinctiveness of D&I research is the concept of implementation strategies. Implementation strategies are multi-component interventions on and within the service delivery system that are aimed at increasing the adoption of new practices into routine care. Powell and colleagues\(^10\) identified 73 discrete components of implementation strategies currently in use in the literature that have at least minimal support for their effectiveness. Concept mapping was then used to categorize these strategies into nine broad domains with varying numbers of discrete strategies in each.\(^11\) These domains are: Engage consumers, Use evaluative and iterative strategies, Change infrastructure, Adapt and tailor to the context, Develop stakeholder interrelationships, Utilize financial strategies, Support clinicians, Provide interactive assistance, and Train and educate stakeholders. It is rare that a single discrete component of a strategy is used in implementation research. Rather, a package of component strategies is typically selected to address multiple barriers to adoption and implementation. However, when testing and evaluating the effects of implementation, key differences in strategies are the focus. For example, multiple strategies might be used to implement a behavioral intervention to increase testing for HIV among high-risk populations seen in community-based primary care clinics. These could include engaging key stakeholders, training clinicians, and incentivizing use of the intervention.

In the context of a trial to evaluate the penetration rate of HIV testing across clinics, the researcher might vary the amount or type of training the clinicians receive or change the way in which the intervention is embedded within the system (e.g., colocated testing vs. referral to an external STI testing center), with all other strategies held equal across systems. A comparison of these two alternatives would then isolate the effects of a strategy that is manipulated by the research team across units (i.e., the independent variable). It is the effects of these implementation strategies on proximal outcomes, processes, and outputs of the service delivery system and, in some cases that are explained in detail later, the distal patient-level outcomes, that are commonly under investigation in D&I research.

**Evaluating Implementation Outcomes, Processes, and Outputs**

Outcomes of implementation research are covered more fully in other chapters of this second edition, but an overview is useful here in the context of our discussion of research designs. Inherent in the definition of implementation strategies is the goal of increasing adoption, which can be conceptualized as a finite event (e.g., a new innovation is implemented) or an ongoing process (e.g., the intent of a service system to adopt a new innovation). However, outcomes in implementation research extend into many areas that affect the adoption and sustained delivery of a new innovation and the efficiency by which adoption occurs. To date, Proctor and colleagues\(^12\) provide the most comprehensive taxonomy of implementation outcomes and place them in the greater context of service delivery system outcomes, using the Institute of Medicine’s Standards of Care\(^13\) and the distal clinical outcomes at the level of the individual. Among the implementation outcomes are acceptability, adoption, appropriateness, cost, feasibility, fidelity, reach or penetration, and sustainment. These are viewed as the direct outcomes resulting from the use of implementation strategies, which, in turn, affect service system outcomes and patient-level clinical outcomes, such as symptom severity.

Not explicit in such a taxonomy is that some outcomes are more germane to particular phases of implementation research.\(^14\) For example, in the early phases of the EPIS model in Figure 13.1, primary outcomes might be acceptability, appropriateness, and feasibility, whereas during implementation, assessments of cost, fidelity, and penetration are possible and considered key outcomes. Likewise outcomes of acceptability, fidelity, and cost, which are relevant to adoption/preparation and implementation phases, all can have lasting effect on sustainment, the 4th and final phase.
Beyond the Proctor taxonomy are measures that evaluate the efficiency and success of the implementation process itself. A prime example of this type of measurement system is the Stages of Implementation Completion (SIC). The SIC is intended to be individualized to specific innovations and service contexts, as it concerns the stages to be completed and the specific activities within each stage. Some examples of stages include, readiness planning, hiring and training of staff, service delivery, and consultation. Each stage contains three or more activities defined by the implementation broker and implementing site, often in consultation with the innovation developer, as being important within a given stage. The SIC measure is focused on critical, observable indicators involving speed, quality, and quantity. These include the completion of key implementation stages, the time spent in each stage, the numbers of the population served, and the proportion of activities completed in each stage. These dimensions can then be used to compare the outcomes between sites. An example of this type of application is provided in the context of the CAL-OH study later in this chapter.

**Research Questions in D&I**

The four phases of D&I research in Figure 13.1 correspond to fundamentally different research questions. In particular, the exploration phase focuses on identifying or enlarging the set of organizations or communities that express interest in using or making available a particular innovation (e.g., intervention or program). One may be interested in the sheer number of settings that express interest through a passive dissemination process, or we may want to identify whether some communities, say those serving high proportions of minority or poor populations, are differentially interested in using a certain program. Early D&I research focused extensively on the characterization of barriers and facilitators to implementation, which remains a key consideration today but is often not viewed as a novel research question given the sheer amount of literature in this area to inform future research. Today, when new research begins, implementation readiness and capacity assessment that occurs during the adoption/preparation phase typically addresses this question in a particular study and informs the choice of which implementation strategies are needed.

We can summarize how D&I research is distinct from other research stages. In contrast to the traditional research questions of efficacy research, which routinely examines overall impact in a relatively homogeneous target population and of effectiveness research, which routinely asks who benefits and for how long in more realistic settings, D&I research questions focus primarily on whether different strategies for informing communities or delivery of an intervention increase the speed of implementation, the quality of program delivery, and/or the quantity or degree of access or penetration of the intervention. One can view these characteristics of speed, quality, and quantity as leading to measurable quantities that can be used to monitor the implementation process. Implementation success would generally be measured by attainment of certain milestones, such as a decision that a community or organization adopts a program, certification that an agency has been credentialed, or other appropriate milestones that can be measured using standardized measures of implementation. Through these milestone measures, we can assess the speed with which implementation takes place. The other two dimensions of quality, and quantity are also considered as critical to evaluating implementation strategies. For example, quality can refer to the fidelity or competence in program delivery, and quantity can refer to how many of the target population are served (similar to concepts of reach covered in chapter 19). Consequently, it is recommended that the study designs, assessment instruments, analytical strategies, and analytic tools for D&I research all should relate to speed, quality, or quantity of implementation. Given the different purposes for D&I research compared with efficacy or effectiveness, it is likely that they may require different research designs or different emphases as they navigate the tension all evaluations have between internal and external validity.

While specific research questions regarding D&I research may most efficiently be addressed by a unique research design, there may remain questions about effectiveness of newly implemented interventions that are worth answering anew. Indeed, hybrid designs, which address research questions related to both implementation and effectiveness simultaneously in one study, are being used much more frequently than they have in the past. Hybrid designs are discussed more fully later in this chapter. Finally, note that the primary focus of this 2nd edition chapter is on implementation without a separate discussion about designs in dissemination research, because dissemination research has not developed a body
of research designs distinct from or as extensive as that for implementation research.

**PREIMPLEMENTATION METHODS**

The process of implementing new innovations in real-world systems is complex and involves numerous decisions on the part of the implementation evaluators and key decision makers in the system. Factors that are commonly considered during preimplementation, which coincides with the exploration and adoption/preparation phases of the EPIS framework in Figure 13.1, include the selection of the innovation to be implemented, the potential implementation strategies used, the population of both providers and patients to be targeted, the expected reach of an innovation within a particular setting given such factors as the characteristics of the population being served, the number of providers that will deliver the innovation, and other resources being allocated. Each of these factors has an influence on the overall impact of the effort and the speed that maximal impact is achieved. Given the plethora of potential options facing implementers, and the need to consider costs and efficiency, certain modeling approaches can be useful during preimplementation, and rarely will there be much reliance on new, rather than existing empirical studies. Policymakers face similar questions when crafting legislation, regulations, and other endorsements of specific and general practices. Attention to these factors occurs during exploration when needs and capacities are evaluated for different alternatives.

At times the number of choices in the exploratory phase is small, such as a situation where policymakers have dictated a certain evidence-based program is to be used. One such example is the recent requirement that states must use set-aside funding to implement the Coordinated Specialty Care model, based on the RAISE project to address the mental health needs of adolescents and young adults who are experiencing psychotic symptoms.

When there are a relatively limited number of possible intervention and implementation delivery choices, some straightforward tools can help guide effective decision-making. Often decision analysis is combined with an economic analysis, supporting not only overall decision analysis, but also cost-effectiveness analyses or cost-benefit analyses that explicitly acknowledge the reality of budget constraints and other limited resources, identifying those interventions that can feasibly maximize the decision makers’ objectives. Economic calculations for different HIV prevention programs, for example, can be compared to provide guidance on efficient use of limited funds or other resources. Linear programming tools have been developed to aid health departments in allocating limited resources for HIV prevention, for example.

However, when there are no strict limits on programs or implementation strategies, the exploratory phase can consist of many different choices, as the number of factors to be considered during pre-implementation yields a parameter space whose size is determined by the product of the number of levels considered in each factor. On top of this factorial explosion in possibilities, models of real-world implementation behavior generally need to take into account interacting processes and agents, resulting in behaviors that are often nonlinear and highly context specific. A full description of such a system with complex behavior requires a scope that is generally infeasible to achieve using traditional experimental methods. There are several simulation approaches, under the heading of systems science, that have been used to model complex behavior in implementation, including system dynamics, network science, and agent-based modeling (see chapter 10). These methods have informed us about strategies to prevent, for example, the spread of HIV through sexual networks. A wide range of questions can be addressed using these simulation models, yet they mostly consider a variety of scenarios of how an implementation scales from the local to the system level. As these computer simulations can be easily scaled up, they enable capturing long-term, systemic impacts of a wide range of specified alternative scenarios, and comparison among them to provide recommendations about the best actions to take. It is important to note that such analyses provide their greatest value when they are employed in an iterative fashion, allowing policymakers to consider a variety of what-if scenarios and to evaluate multiple decisions holistically.

This section on preimplementation simulation designs focuses on two simulation approaches showing how decision analysis with microsimulations and agent-based modeling can be used to address this mismatch in scope and guide implementation research during the preimplementation stages. Decision analysis represents a significant advance to project the needs...
of an agency and potential of a novel intervention to address a given service need. It employs computer-based models that simulate behaviors within a system. Using either microsimulations or agent-based modeling, this behavior is specified on the micro-level, in contrast to system-dynamics models, which consider dynamics on the system level alone. Specifying behavior on the micro level allows the heterogeneity in the system to be embraced, rather than to be controlled for or averaged out.

Microsimulation as a decision-support tool is well suited to supporting human service policymakers (such as child welfare directors and managers) as they confront the challenges of complex, real-world operations, as it can help identify policies and interventions that are most likely to achieve a set of desired objectives given current uncertainties. Indeed, decision analyses have been used successfully to consider complex decisions involving health conditions with long time courses and multiple outcomes. Such analyses have considered the prevention and management of HIV/AIDS, cardiovascular disease, diabetes, and HPV and cervical cancer. For example, relying in part on model-based cost-effectiveness analyses, the Institute of Medicine recommended a shift from funding programs based on high AIDS prevalence to targeting prevention efforts to subgroups at high risk of infection.

In the context of ongoing Child Welfare services, evidence-based interventions’ superior effectiveness must be considered with respect to how well they enhance safety, permanence, and well-being and to any additional resources required to implement and maintain them at levels ensuring effectiveness. As decision-makers contemplate these interventions, they require actionable information to overcome uncertainty as to whether the long-term benefits and averted costs of evidence-based interventions justify the investment. With over 1 million children served by the US child welfare system at a cost of $20 billion annually, the use of evidence-based interventions has the potential to improve the health and well-being of a large, vulnerable population in a more cost-effective manner. However, substantial investments may be required to incorporate such interventions into child welfare agencies requiring evidence to support these decisions as they weigh the various trade-offs.

An excellent illustration of the use of decision analysis is provided by Goldhaber-Fiebert and colleagues. The authors used a computer-based microsimulation model to evaluate the effect of implementing one such evidence-based foster parent training intervention: KEEP (Keeping Foster Parents Trained and Supported). The microsimulation computed policy-relevant outcomes such as increased rates of adoption and reunification (positive exits) along with improved foster care placement stability (e.g., reduced lateral foster placement changes and reduced negative exits to group care) resulting from the application of KEEP. The microsimulation incorporated data on children in foster care from randomized controlled trials of KEEP as well as large, population-representative longitudinal studies (e.g., National Survey of Child and Adolescent Well-Being), using multivariate Cox proportional hazard models and bootstrapping to provide estimates of the rates of foster care placement change, the main covariates that determine these rates, and their associated uncertainty.

The detailed microsimulation developed for this analysis simulated large cohorts of individual children whose characteristics matched those of the actual foster care populations within the US child welfare system. The model then followed these “simulated individuals” on their paths through the system, tracking their placement changes and allowing past experiences to influence their future risks of placement change and exit. This approach permitted the consideration of the rich, complex effects of each individual’s experience in the system of over time, identifying cumulative benefits to KEEP, emphasizing higher-risk groups of children who may differentially benefit from the application of the intervention, and gauging the heterogeneous mediating effects that different state child welfare systems could have on KEEP. The paper demonstrated decision analytic methods to employ existing data to project policy-relevant child welfare outcomes related to permanence and stability. Decision-analytic microsimulation modeling is a feasible and useful methodology to inform challenging child welfare policy decisions and can be extended to consider multiple evidence-based interventions and outcomes.

Another recent example highlights how agent-based modeling can be used to study policy impacts. In particular it focuses on the effects of various CDC guidelines for PrEP (preexposure prophylaxis) prescription on HIV prevalence. In this study the authors use existing field data to create a realistic system interacting agents, men who have sex with men, and simulate not only
Design and Analysis in Dissemination and Implementation Research

how HIV spreads among that population but also how different implementations of CDC guidelines for PrEP prescription affect this spreading process. The simulation incorporates a heterogeneous population of 10,000 agents with variations in sexual activity, risk behaviors, testing frequencies, and adherence levels. A second agent-based modeling approach highlights the importance of STI testing as a critical factor determining intervention success.57

These examples highlight the power of simulation for implementation during the pre-implementation stage. It highlights this method as a scalable tool for doing virtual experiments, scanning the parameter space, and conducting consequent scenario analysis based on the outcomes, but the potential goes well beyond that. Decision analysis using simulation can also identify the factors critical for successful implementation, inform which populations to target and the differences to be expected for various target populations, and identify tipping points in the system, which can inform the amount of resources needed to achieve the desired impact.

RANDOMIZED AND NONRANDOMIZED IMPLEMENTATION DESIGNS FOR ADOPTION/ PREPARATION, IMPLEMENTATION, AND SUSTAINMENT PHASES

Historically, basic science and treatment research as partners in the first translation step have relied heavily on what has come to be known as the “gold standard” of designs, namely, the randomized controlled trial (RCT), involving randomization at the person level. In the efficacy phase the primary aim is to determine whether an intervention has impact on its intended target. A great deal of methods development has been devoted to the use of RCTs to evaluate program efficacy for medical48 and behavioral research.49 While effectiveness research also has played an important role in the science to practice continuum, group-based randomized trials are generally needed for these more complex longitudinal designs that often include multiple levels in the analysis.50,51 In one of the few comprehensive discussions of the distinction between efficacy and effectiveness trials, Flay in 198652 noted that “whereas efficacy trials are concerned with testing whether a treatment or procedure does more good than harm when delivered under optimum conditions, effectiveness trials are concerned with testing whether a treatment does more good than harm when delivered via a real-world program.”52

It is not accidental that Flay's language on D&I includes a discussion of random assignment to different approaches for delivery. Flay’s perspective then was that randomized trials could be used for such research, but such designs would need to differ from individually based RCTs.2 There are circumstances where randomization may not be feasible or acceptable53 and alternatives may be proposed to the randomized design such as "interrupted time series,” “multiple baseline across settings” or “regression-discontinuity” designs. For example, Brown and colleagues have argued that incorporating randomization across time and place in roll-out trials can be acceptable for both communities and researchers.50,54

This section reviews the major issues in designs for D&I research, with a particular focus on randomization and alternatives to randomized designs, and also discusses the emerging development of hybrid, adaptive, and staging designs as well as mixed method designs. This is followed by a section that discusses power calculations in multilevel implementation designs and the use of analytic strategies such as mediation and moderation. Although little work in the use of mediation and moderation analyses in D&I research exists to date, the section lays out the need for this approach if we are to better understand the mechanisms and limitations of implementation strategies. Finally, this chapter addresses a rapidly emerging set of tools for dissemination and implementation research under the label of system science and then concludes with a case example of a rigorous and complex study in California of the impact of two implementation strategies for a 51-county roll-out of the robust, evidence-based intervention “Treatment Foster Care Oregon.”

This emphasis on randomized designs in implementation research is supported by a paper that reviewed types of designs published on D&I research in nonmedical care service systems, namely child welfare and child mental health.55 Using standardized search strategies, nine relevant studies were identified, all involving randomization, eight at a single level and one at 2 levels (treatment intervention vs. control, and intervention strategy vs. control). Note that randomized designs constitute the majority of implementation studies that met the criteria of some kind of control or comparison condition. Also
note that approximately 1 in 10 studies reviewed by EPOC used an interrupted time series design as an alternative to randomization. This is one of the four alternatives to randomization reviewed by Glasgow and colleagues.55

The classification of D&I designs put forth by Brown and colleagues’ encompasses three overarching categories of designs based on the type of comparison.

Within-Site Designs: that is, evaluation of an implementation project focused on change within a single site. This classification includes two types of within-site designs. The weaker design is the post design, where changes in the system’s health care processes and utilization are evaluated after the introduction of an evidence-based practice. The more rigorous variant, the pre-post design, adds an evaluation of preimplementation data, which can then be compared to postimplementation data to infer effects. Interrupted time series designs in the EPOC framework56 are a slight methodological improvement on the within-site design that simply requires additional measurement and analytic techniques that account for trend, serial dependence, and other characteristics of such data.57 All these within-site design variants are considered quasiexperimental.

Between-Site Designs: that is, evaluation that compares outcomes, outputs, and processes between two or more sites where a novel intervention is being implemented. A basic type of study in this category is a head-to-head comparison of two implementation strategies for a specific intervention that occurs in two different sites or two groups of sites—randomly assigned when applicable. A variant of this design could be to assign different units or wards within an organization and compare the effectiveness of different implementation strategies. A useful rule of thumb is that randomization should be at the “level of implementation,” meaning the level where the full impact of the strategy is designed to occur.58 Brown and colleagues’ also discuss special cases of this family of between-site designs: factorial designs; double randomized, two-level nested designs; and site selection of implementation strategies using a decision support strategy, as opposed to an a priori randomization that forces sites to use a certain implementation strategy.

Within- and Between-Site Comparison Designs: that is, sites begin as one implementation condition and move to another. The most common example of this design is a roll-out randomized implementation trial, which is similar to the stepped-wedge59 and dynamic waitlisted design60 that has been used in effectiveness research for decades. In such designs, sites are randomized to a time to crossover from implementation as usual to the use of the implementation strategy. In this way, each site serves as its own control (within-site) and can be compared with the performance of other sites (between-site). A number of variants to this design are possible, including clustering sites to start at different time points and pairwise-enrollment roll-out designs.60 The roll-out trial is particularly palatable to community organizations, as they are assured of receiving the active implementation strategy at some point in the trial as opposed to serving solely as a control site. Last, in situations with a very small number of units to randomize, power can be increased and the experimental design can be strengthened by using randomized multiple-baseline designs,61 which come from the single-subject experimental design tradition.

The observed dominance of D&I research designs with randomization supports the benefits of such designs for addressing common threats to interpretation of study findings. However, it is well to consider the nature of threats to the integrity of a randomized trial for dissemination and implementation research before further reviewing a range of design options for D&I research.

Often, one emphasizes characteristics of well-conducted research studies or trials in order to guide researchers into conducting studies that will lead to accurate scientific inferences. While this is helpful, the authors believe this is too optimistic a perspective, since research needs more than an overly optimistic “glass half-full” attitude to make appropriate conclusions. Indeed, a study can do many things right, but fail in just one way and thereby put all its inferences at risk. The critical scientific paradigm for assigning observed differences by condition to an implementation’s effect is that the only systematic factor that differs by intervention condition is the assigned intervention. For example, an implementation trial can carry out an appropriate randomization of communities to one of two implementation conditions, use valid and reliable measures, and conduct statistical analyses that correctly take into account intraclass clustering within communities. But if community leaders in one arm of the trial are more likely to refuse to be interviewed or drop out more frequently, this effect on the quality of the inferences can never be compensated by other
good parts of the study. To reduce the potential for such imbalance, and to hold an implementation design in place, requires a strong and active partnership between communities, institutions, and researchers. Following, we list some of the factors that are known to affect the quality of inferences, with special attention to implementation and dissemination research.

Random assignment is the obvious choice for insuring that implementation condition is fairly distributed (sometimes with blocking into similar communities followed by random assignment within these blocks). While some have suggested that random assignment is not appropriate for implementation research, our review indicates that there have been a sizeable number of such randomized implementation trials conducted, and they do have an important place in this research agenda. The usual alternative is a comparative study where one or more select communities apply a specified implementation procedure while other communities, often selected afterwards, are used as comparison. The problem with this design is that it hopelessly confounds two factors: the implementation itself as well as community readiness, since only those communities that are “ready” are prepared to implement. One can never distinguish whether the differences in communities is due to one of these factors or both. An alternative design is a roll-out design where communities that express their willingness are randomized to the timing of implementation. Such a dynamic wait-listed design has been used in the comparison of two implementation strategies for an evidence-based intervention for foster care to be described later in the case study.

A second major threat to an implementation trial is a failure to use valid and reliable measures to assess implementation outcomes (also see chapter 14). Because the implementation process is inherently multilevel, it is critical to assess impact across the appropriate levels. One potential flaw can occur if one only measures the distal outcome on a target population that is served, since these may not be comparable across intervention conditions. For example, suppose an implementation strategy is designed to increase the number of youth who receive an evidence-based program. If we compare findings from those youth in communities randomized to the new implementation strategy to those using implementation as usual, it is quite possible for systematic differences to occur between those target youth who are exposed to different interventions. There is no mechanism to guarantee that those who receive the intervention are equivalent, and it may be that the expansion of service delivery brings in more or less challenging populations that are served. Analytic strategies that use propensity scores to adjust for differences at the nonrandomized level could be considered. There is ordinarily no need to adjust for covariates at the higher level where randomization occurs because randomization preserves balance.

While there are enormous benefits to randomization, the multilevel nature of D&I research creates issues for typical randomization at the individual level designs. Figure 13.2 shows a classic multilevel structure with four levels that Shortell has suggested for change in assessing performance improvement in organizations. While randomization can and is being done at levels higher than the individual, there is an issue in having sufficient power as one moves to higher levels with diminishing units to be used in a randomized design (see later section on calculation of power in multilevel designs Selected Websites and Tools). Since power is so critical to the use of randomized designs, it is reasonable to consider quasi-experimental designs without randomization for D&I research. This clearly is the thrust of Glasgow and colleagues in their 2005 article on practical designs. Another useful source for the alternative designs and the trade-offs between randomized and nonrandomized designs is the Handbook of practical program evaluation, edited by Wholey, Hatry, and Newcomer, especially the contrasting chapters on “Quasi-experimentation” by Reichardt and Mark, and the chapter on “Using randomized experiments” by Pierre. Note that Reichardt and Mark describe four prototypical quasi-experimental study designs: (1) before-after; (2) interrupted time-series; (3) nonequivalent group; and (4) regression-discontinuity designs. Their rendering of alternative designs is quite comparable to the EPOC classification of designs in implementation research and to the discussion of alternative designs by Glasgow and colleagues. Also worth noting are important recent studies that have compared the results between randomized experiments and observational studies. These have included both examples of strong divergence between observational studies and experimental studies, such as found in studies of hormone replacement therapy, as well as examples where the results were remarkably
consistent. Notable is the paper published in 2000 by Concato, Shah, and Horwitz\(^7^5\) that examined meta-analyses of randomized controlled trials and meta-analyses of either cohort or case-control studies on the same intervention. Across interventions addressing five very different medical conditions, the authors found remarkable similarity between the results between the two types of designs, which are perceived to be quite different in the hierarchy of evidence. The authors concluded that “The results of well-designed observational studies (with either a cohort or a case-control design) do not systematically overestimate the magnitude of the effects of treatment as compared to those in randomized, controlled trials on the same topic” (p. 1887). A later study published in 2008 by Cook, Shadish, and Wong\(^7^6\) came to the same conclusion when comparing the results from randomized experiments and regression-discontinuity designs. In both papers, the authors argued that the quality of the observational studies had to be high to be comparable to the results in randomized designs. This line of research informs the emerging field of implementation science by suggesting that observational studies may also be seriously considered.\(^7^7\)

**RETHINKING RANDOMIZED DESIGNS**

An alternative approach to problems with use of classic RCT designs, so important in efficacy and effectiveness research stages, is to rethink how randomized designs can be adapted to meet the special needs of research across the four phases of D&I research. The next section discusses some of this rethinking, by considering nontraditional ways of using random assignment in D&I research.

Random assignment provides a method for making fair comparisons between intervention conditions, but the standard procedure for assigning individuals as well as organizations or communities often is impractical to administer, unacceptable to stakeholders, or irrelevant to the major research question.\(^7^8\) This section describes ways that nontraditional means of using random assignment can still be used to address important dissemination and implementation questions. One fundamental way of using random assignment involves the timing of delivery of an intervention. Often communities are uncomfortable withholding an intervention completely from a subpopulation, especially if the

---

**FIGURE 13.2** Four levels of change and assumptions about change.

(Adapted from: Shortell SM.\(^7^9\))

---

---
intervention is preventive in nature, holds little apparent risk of being iatrogenic, or addresses a major health problem in a community. In such a situation, researchers may suggest the use of a wait-listed design, which allows half the units to receive the intervention first, followed by the second set receiving it later. Communities are often more comfortable with this design compared with one involving a traditional control condition, and such a design can be used to address whether proximal impact on implementation targets (e.g., program adoption, fidelity, and reach into a target population) occurs. However, such standard wait-listed designs are inherently inefficient because the data from the second half of the study when the wait-listed group receives the intervention is weak or of negligible use in answering questions on implementation and dissemination.

An alternative class of designs for dissemination and implementation is a roll-out randomized design,⁴⁰ which are sometimes described as dynamic wait-listed⁴⁴ or stepped-wedge,⁷⁹,⁸⁰ although historically these have been used much more in effectiveness than implementation research. When used to study implementation, these designs randomly assign units, which may be schools⁵⁴,⁶⁵ or larger settings such as counties,⁵⁰,⁶⁶,⁶⁷ to different "start" times and/or types of implementation strategies. Starting with a small number of units assigned at the first time point, eventually all of the units receive an active implementation strategy. Because the timing is random and measurement occurs before and after implementation, one can track whether intended changes occur both across time and across condition for most outcomes and outputs on most if not all the units. This type of design has been used to test two implementation strategies for an evidence-based program in foster care.⁶⁶,⁶⁷ There are three appealing features about this roll-out design compared with traditional wait-listed designs. First, the statistical power is substantially greater for the roll-out design in answering many relevant research questions.⁵⁴,⁶⁰ Second, it is often not practical to train a large number of communities in an intervention all at once, and the roll-out feature nicely focuses on a small manageable number who can receive the training attention that they need. Third, the use of multiple times for assignment provides more robustness of the design to influence by exogenous factors, such as economic downturns, that could otherwise destroy any chance to make inferences if its timing occurred at the most critical time in the assignment to training.

The authors recommend that a roll-out design be considered when an implementation program is being introduced into a set of communities as part of a federal, statewide, or local policy change.²⁰,⁵⁴,⁶⁵ Such roll-out trials could be of high utility in implementation research (see 1st edition chapter 12). As an example of an appropriate use of such a trial, consider examining the following typical strategy for improving the quality and effectiveness of services by improving the training of mental health counselors. There are four levels that we need to pay attention to: the level of the client who receives services, the therapist who delivers the services, the supervisor whose job it is to improve service delivery within a mental health agency, and the agency itself.

Let us consider an implementation strategy that changes the supervision practices of therapists within agencies. A first question is what level of randomization should be used to provide maximum utility to understanding whether a new supervision strategy is effective relative to that now being used. We can consider randomizing at any and all four levels from the client to the agency, but many of these levels will not be very useful. A key step is to identify what the most salient "level of intervention" is being examined; here we are fundamentally interested in how the use of a new supervision program will affect outcomes downstream, and therefore the key level of intervention is at the supervisor level; this is the first level where we would expect behavior to change. An empirically validated rule is the following: whenever possible one should randomize at the level of intervention, since randomization at lower levels (e.g., therapist) would contort supervisors from using new techniques that now being used. We can consider randomizing at any and all four levels from the client to the agency, but many of these levels will not be very useful. A key step is to identify what the most salient "level of intervention" is being examined; here we are fundamentally interested in how the use of a new supervision program will affect outcomes downstream, and therefore the key level of intervention is at the supervisor level; this is the first level where we would expect behavior to change. An empirically validated rule is the following: whenever possible one should randomize at the level of intervention, since randomization at lower levels (e.g., therapist) would contort supervisors from using new techniques to train therapists in the standard condition, and randomization at higher levels (e.g., agency) will generally result in major reductions in power.⁵⁸ If we do randomize at the level of the supervisor within an agency, then the agency is considered a "blocking factor." Blocking is a well-known way of reducing variability and thereby increasing power. In a roll-out design, we would randomly determine the order and timing of training of supervisors and consequently their transmission of new behaviors to their supervisee therapists and in turn their own clients.

To assess impact in this trial design, we would likely want to measure behaviors of the supervising process (supervisor–therapist interactions), as
well as of the therapeutic session itself (therapist–client), and perhaps at the level of the client target behavior as well. All three of these measures would normally be assessed over time. Thus the design in this study would typically involve multiple observation/coding times for supervisors before and after they were trained to deliver a different type of supervision, multiple observation/coding times for their respective therapists, and multiple observation times for different clients across time. Three levels of analysis could be done, with the first examining changes in supervisor behavior across time before and after training; the second involving therapist-client behavior, also coded in terms of how therapist behavior with clients related to the timing of the supervisory training they received, and third the behavior of the clients themselves. To connect these three analyses, we would conduct mediation analyses, which are presented later in this paper.

The sample size for such a trial may be nested at several levels. For example, sample size determination may include the number of agencies, number of supervisors in each agency, the distribution of the number of therapists who receive supervision within these agencies, and finally the number of clients served by each therapist. Characteristics of timing include when supervisors in each agency receive training, the number of observation points in each supervision, the number of observation points for each therapist in interacting with a client, and the baseline and follow-up times for assessing client behavior. To make sure that there is sufficient power in this design to answer all three questions of impact on supervisor behavior, on therapist behavior, and on client behavior, as well as on mediational pathways, we would need to carry out a sophisticated study of how statistical power relates to these sample sizes and timing. While there are programs that allow one to compute power in multilevel designs, to date the calculations for such a trial may be nested at several levels. For example, sample size determination may include the number of agencies, number of supervisors in each agency, the distribution of the number of therapists who receive supervision within these agencies, and finally the number of clients served by each therapist. Characteristics of timing include when supervisors in each agency receive training, the number of observation points in each supervision, the number of observation points for each therapist in interacting with a client, and the baseline and follow-up times for assessing client behavior. To make sure that there is sufficient power in this design to answer all three questions of impact on supervisor behavior, on therapist behavior, and on client behavior, as well as on mediational pathways, we would need to carry out a sophisticated study of how statistical power relates to these sample sizes and timing. While there are programs that allow one to compute power in multilevel designs, to date the calculations for dynamic wait-listed designs are generally only available to do by simulation (see later section on power calculations Selected Websites and Tools).

RETHINKING STAGES OF RESEARCH WITH HYBRID DESIGNS
An innovative approach to implementation research design was published in 2012 by Geoffrey Curran and colleagues titled “Effectiveness-Implementation Hybrid Designs: Combining Elements of Clinical Effectiveness and Implementation Research to Enhance Public Health Impact.” The authors were all involved with the Veteran's Administration QUERI program in an early federal agency approach to the development and use of implementation science approaches to improving the quality of care in the VA health system. As the title of this section indicates, the 2012 paper is grounded in a rethinking of the stages of research in the “traditional research pipeline,” where the steps from efficacy to effectiveness to implementation were seen as discrete stages, each of which was expected to be “completed” before the next stage was begun. Adhering to the pipeline meant that the implementation research stage was seldom entered in until the work and outcomes of the efficacy and effectiveness stages were finished. The traditional pipeline also suggested that the level of evidence issue had been concluded with proof of a strong evidence base “now ready for implementation” and provided no guidance for what to do with interventions with less than stellar evidence but which might have already shown robustness in fit for real-world service systems.

The concept of hybrid designs, which blends the two stages of effectiveness and implementation, was originally proposed as a way to increase the speed of moving research findings into routine adoption. The authors argued that one did not need to wait for the “perfect” effectiveness data before moving to implementation research and that it was possible to “backfill” effectiveness data while testing implementation strategies. Further, the hybrid concept shed light on the critical question of how clinical/prevention outcomes might relate to level of adoption and rate of fidelity—which can only be known when we have data from “both sides,” or simultaneously from effectiveness and implementation types of data. The notion to blend effectiveness and implementation research also encourages consideration of building for “implementability” as early as you can in the development of interventions (e.g., considering service delivery issues such as geographic context, staffing, technology, and “dose”), seeking end-user input and partnership before initiating effectiveness trials, and using implementation frameworks in design of components (e.g., CFIR framework). The sample size for such a trial may be nested at several levels. For example, sample size determination may include the number of agencies, number of supervisors in each agency, the distribution of the number of therapists who receive supervision within these
Design and Analysis in Dissemination and Implementation Research

Types of Hybrids

- **Hybrid Type 1**: test clinical/prevention intervention, observe/gather information on implementation
- **Hybrid Type 2**: test clinical/prevention intervention, test/study implementation strategy
- **Hybrid Type 3**: test implementation strategies, observe/gather information on clinical/prevention outcomes

**FIGURE 13.3** Types of hybrids.
(Adapted from: Curran GM. Research Questions and Design Considerations: Effectiveness to Implementation. Presented at the University of Wisconsin, Dissemination and Implementation Short Course Program, Madison, WI, October 2016.)

agencies, and finally the number of clients served by each therapist.

The language of hybrid designs requires some consistency in terminology. This section uses the term “intervention” to refer to the clinical/prevention practice and the term “strategy” to refer to the implementation-support activities/tools. Because both are interventions, usually the term “effectiveness” is used when referring to clinical/prevention outcomes and the term “impact” when describing implementation outcomes.

The 2012 paper by Curran and colleagues proposed three types of hybrid designs, which are shown in Figure 13.3. Hybrid Type 1 tests intervention effectiveness while gathering information on implementation issues, and Hybrid Type 3 tests implementation strategies while gathering information on clinical/prevention outcomes. Hybrid Type 2 tests the clinical/prevention intervention while also studying an implementation strategy or strategies. This typology suggests that the emphasis on effectiveness outcomes is greatest in the first type and the emphasis on implementation strategy outcomes is greatest in the third type.

Another way to clarify the distinction between the types is presented in Table 13.1, which indicates what is the primary aim and secondary aim for each type. Note here that Curran et al.’s 2012 paper is included in the suggested readings for this chapter, and the reader is recommended to consider not only the text of the article but also the detailed tables that lay out the key terms and definitions, the characteristics of the three “ideal” types, and very detailed characteristics and key challenges for each type. In addition, the article provides published examples of empirical studies that used a specific type of hybrid design. However, it is important to consider that these are “ideal” types and specific proportions of emphasis on effectiveness and implementation issues may vary considerably among studies that indicate they used a specific type.

An important consideration for the three types of hybrid designs is the locus of randomization and power estimation, which in turn impacts overall study cost. Because the primary aim for Type 1 is clinical/prevention effectiveness, randomization would be at the individual patient/client level with appropriate power calculations. Consequently, the burden of study cost would be primarily driven by the effectiveness part of the study, while the implementation-focused activities would deploy less expensive processes such as surveys and qualitative interviews. The reverse would likely be done for a Hybrid Type 3 design, where randomization will take place at the level of organization or units within service organizations (e.g., providers or clinics). This typically means that adequate power is more difficult to achieve. The burden of cost and the adequacy of power for implementation impact may be
increased if secondary data sources are available for the effectiveness outcomes (e.g., administrative data routinely collected or electronic health records).

Recently, Curran and colleagues have been conducting a structured review of published studies and have come up with more than 80 papers that report using hybrid designs. While a more detailed summary of the manuscripts and an accompanying set of revised recommendations for hybrid designs will be forthcoming, a brief set of findings and recommendations from the review are offered here. Concerning Type 1, currently published examples in the literature range from conventional effectiveness studies with limited exploratory measurement of barriers/facilitators to implementation (more common) to more intensive parallel process evaluations guided by implementation frameworks explaining effectiveness findings and elucidating implementation factors (less common). Concerning Type 2, the majority of published examples thus far embed a patient-level effectiveness trial nested with a pilot study of an implementation strategy (non-randomized). Very few “dual-randomized” designs are being used. Concerning Type 3, currently published examples indicate an array of targets for randomization (e.g., providers, clinicians, systems) with clinical/prevention outcomes almost exclusively being collected/observed through medical record review or administrative data (nonprimary data collection at this level).

### Table 13.1 Hybrid Types and Study Aims

<table>
<thead>
<tr>
<th>Study Characteristic</th>
<th>Hybrid Type I</th>
<th>Hybrid Type II</th>
<th>Hybrid Type III</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary Aim:</strong></td>
<td>Determine effectiveness of an intervention</td>
<td>Determine effectiveness of an intervention</td>
<td>Determine impact of an implementation strategy</td>
</tr>
<tr>
<td><strong>Secondary Aim:</strong></td>
<td>Better understand context for implementation</td>
<td>Co-Primary* Aim: Determine feasibility and/or (potential) impact of an implementation strategy</td>
<td>Secondary Aim: Assess clinical outcomes associated with implementation</td>
</tr>
<tr>
<td><strong>Primary Aim:</strong></td>
<td><strong>Primary Aim:</strong> Determine effectiveness of an intervention</td>
<td>Co-Primary* Aim: Determine feasibility and/or (potential) impact of an implementation strategy</td>
<td></td>
</tr>
<tr>
<td><strong>Secondary Aim:</strong></td>
<td>*or &quot;secondary&quot;...</td>
<td>or &quot;secondary&quot;...</td>
<td></td>
</tr>
</tbody>
</table>

#### SPECIAL ISSUES IN DESIGN AND ANALYSIS OF D&I RESEARCH

**Power and Sample Size Calculations**

The simpler tables that exist for calculating statistical power and sample size are typically not appropriate for implementation studies because of the multilevel or clustering and longitudinal nature of the data. Two online tools are often useful for these calculations, the Optimal Design system available through the W.T. Grant Foundation, and the RMASS program developed by the Center for Health Statistics at the University of Chicago. Websites for both are listed at the end of this chapter.

**Randomization at Single versus Multiple Levels in D&I Research**

Designs that involve multiple levels of random assignment or allocation are often useful in increasing the precision of inferences. For example, to examine sustainability of a classroom-based intervention, the authors randomized first grade teachers and classrooms within schools to the timing of training. This was the primary unit of intervention for this study. To compare early versus later training in this classroom-based intervention design, we would essentially rely on the average differences within schools for classrooms.
that had early versus late training. However, we knew that because schools ordinarily assign students with like ability in the same classrooms, called ability tracking, classrooms within schools would typically not be well matched. Therefore, a design that tried to compare outcomes for the one or two early-trained teachers in each school to those who were trained later would have to introduce a large heterogeneity in classrooms unless children were matched into similar classrooms within schools. This in fact was done by random assignment of all children to classroom as they were enrolled in the school,86 and this design greatly increased the statistical power over a design that did not have balanced classrooms.85,87

Such designs that use random assignment of groups at one level and random allocation of individuals into these groups can provide substantial improvement in power.

An alternative way to use randomization is to randomize units at two levels to different types of interventions. For example, a system-level intervention can be tested by random assignment of agencies to different intervention strategies, say to business as usual versus a new system for monitoring fidelity and providing corrective.

A client-based intervention can be tested against standard practice by randomly assigning clients to these conditions within each agency. Such a two-level intervention design is known as a “split plot,” and is commonly used in industrial experiments88 but is suitable for implementation trials as well. It can be used to compare overall effect of the implementation strategy, overall effects of the client-based intervention, and their interaction. An example of such a design for implementation studies is given by Chambers.89

Pilot Studies and Estimating Effect Size

In developing a new implementation strategy, it is often very sensible to conduct a small-scale study to examine feasibility and assess sources of variation to determine the size one would need to have a fully powered study. While this is a useful approach, some caution is advised, especially with the estimation of effect sizes from a pilot study. As pointed out by Kraemer,90 the precision of an effect size estimate from a pilot study is much lower than that in a fully powered study, and consequently, considerable uncertainty in power is introduced when this estimated effect size from the pilot study is entered into power calculation programs. It is possible to use these pilot studies to get some useful information about the magnitude of different sources of variance,20 but the intended magnitude of the effect that one is aiming to achieve should best be determined by clinical relevance, rather than the estimated value obtained from the pilot study.

Analyzing for Mechanisms Through Mediation

A major task for D&I method development is to determine new mediation analysis methods for implementation modeling. There will be many challenging causality questions, particularly around how to best handle protocol deviations in implementation trials. One of these challenges will be to deal with the possibility of differential attrition in intervention agents in the intervention arms. For example, teachers who are successfully trained in an EBP such as the Good Behavior Game that promotes improved classroom behavior may be less frustrated with teaching and less likely to leave the profession.91 Such attrition differences can become manifest in a longitudinal comparison of teachers who have been trained in this intervention versus those not trained,21 and may be one explanation of persistent intervention effects.

This work also needs to incorporate innovative developments in causal inference for mediation models, that is, principal stratification (PS).92–104 Traditional PS causal modeling characterizes subgroups that are responders to an intervention as well as nonresponders92,95,97 and can assess the magnitude and extent of different impact that these subgroups experience. PS techniques are also closely connected to latent variable modeling and mediation analysis,105–110 and these latent variable approaches can be used for evaluating intervention impact and examining rates of differential response across intervention conditions.111 There is, however, a need to extend these methods to take into account repeated measures over time and multilevel modeling. As one example, Brown and colleagues have proposed developing two-level PS to account for different strata of intervention agents (e.g., teachers’) response to training. Like traditional PS, teachers could be classified into always successful managers of their classroom whether or not they received intervention training, always unsuccessful even if given training, and responders to the training. Also like PS, in implementation trials we cannot know any teacher’s classification exactly, since we can only observe their response to the training.
or no training condition. However, causal inferences for outcomes of youth, who are exposed to these teachers, are indeed possible in randomized trials. The authors have identified a class of such two-level PS models using the Georgia Gatekeeper Trial and are developing and testing several approaches in implementation research.

Finally, work needs to consider application of existing methods for multistage follow-up to evaluate the effect of prevention programs on severe levels of drug abuse, an area that has not received sufficient attention in the literature in many evaluations. These applications for new trials will involve epidemiologic assessment of the strength of antecedent risk factors, that is, multiple drug use in adolescence, on later severe abuse. While the method is straightforward, it will require complex calculations to take into account unknown levels of differential attrition and any direct effects from program to drug abuse that do not involve drug use.

**TWO SYSTEM SCIENCE APPROACHES IN D&I: SYSTEM ENGINEERING AND INTELLIGENT DATA ANALYSIS**

The term “system science” refers to a transdisciplinary approach to understanding how interactions between elementary units produce complex patterns and, using the NIH/OBSSR definition, take into account the “complexity, dynamic nature, and emergent phenomena.” System science methods typically include social network analysis, agent-based modeling, and systems dynamics (see chapter 10) as well as other tools such as decision analysis and systems engineering. These system science methods are critical to moving implementation research forward as a science for the following reasons. First, implementation is inherently interactional, across multilevels within systems as well as between systems, and only when different systems function together can we expect implementation to succeed. The system science methods discussed directly deal with interactions, in contrast to, say, traditional statistical modeling involving standard regression modeling, which assumes complete independence, or multilevel growth modeling, which allows correlation across persons and time but does not explicitly model these processes. Second, implementation process data, which are essential in research to monitor progress, are heavily dependent on interactions between actors, as exemplified in the development of community-researcher partnerships.

Third, implementation process data also are essential for communities, organizations, and service systems themselves to provide monitoring and feedback for quality improvement. Most of today’s research level implementation process data are very expensive for these systems to collect. Consequently, we need to develop cost-effective ways to assemble quality implementation process data, conduct analyses on these data, and integrate this into a monitoring and feedback system. The authors fully anticipate that such systems can be built by automating these steps as much as possible. Systems science methods will need to be used for all these purposes. Described in the following are system science methods not included in chapter 10, and brief illustrations of how they can be used in implementation.

**Systems Engineering**

Systems engineering refers to the processes of identifying and manipulating the properties of a system as a whole rather than by its component parts. Systems engineering is both a discipline and process to guide the development, implementation, and evaluation of complex systems. A system is an aggregation of components structurally organized to accomplish a set of goals or objectives. All systems have the following characteristics: a structure, interacting components, inputs and outputs, goals and objectives. Systems are dynamic in that each system component has an effect on the other components.

System optimization requires design consideration of all components of a system, in contrast to a more traditional reductionist approach focusing on individual components. Attempts to design a system without considering the dynamic physical and social environments where the system operates will degrade system performance. For example, if a school-based prevention program competes for time against instruction, output would be low. Task analysis is one systems engineering tool used to characterize the implementation process and necessary resource and skill requirements. It was successfully applied to describe 15 complex intervention programs in the Resources for Alzheimer’s Caregiver Health program conducted by Czaja and colleagues. Another system engineering tool is the Analytic Hierarchy Process. This can be used to capture decision-making within systems. By focusing on the decision process and establishing priorities,
the Analytic Hierarchy Process can identify critical attributes of an intervention and areas where an intervention might be modified or adapted. It is not well known that strategies for implementation vary dramatically. We have been struck by the vast different approaches that have been used in different implementation strategies. To compare these alternatives, a full characterization of different strategies is required, and that requires the use of a standard procedure for eliciting intended implementation strategies as well as identifying where inefficiencies and other problems exist. The use of task analysis and the analytic hierarchical process and related techniques provides the ability to develop an ordering of priorities in decision-making to distinguish different implementation models in theory and practice.

**Intelligent Data Analysis**

Intelligent data analysis refers to advanced computational methods to automate the generation of meaning from text, video, or audio signals. These techniques can be used to reduce large amounts of process data on implementation that come in digitized form. Most implementation process data that are typically collected in agencies, such as number of people attending meetings or self-ratings of fidelity, are only crude indicators of the implementation process. However, there are other sources of information on the implementation process that are rarely codified but can be converted to analyzable data. Notably, these include audio and videotaped training as well as program delivery sessions. Such information is highly useful for supervision, but often just a small portion of this information is ever used. Automated signal processing and feature extraction of videotapes are possible using intelligent data analysis, and the outcomes of such methods would help to identify specific ways to improve facilitator fidelity.

As one illustration of the use of intelligent data analysis, consider the availability of contact logs, process notes, emails, and other communications to monitor implementation. Automated generation of meaning from text is one important tool to convert information that generally requires time-consuming human judgment. If we were able to automate the process of transferring such information into meaningful data on the implementation process, it would result in a major savings in costs for agencies and service providers. The transfer of text to meaningful terms involves intelligent data analysis. Similar to psychometrics, computer science research has used latent variables via the Hidden Markov Model (HMM) to derive meaning from text information. In HMM, there is a time-dependent process governed by state changes and its associated state-dependent distribution that produces observations at each step (e.g., in speech a syllable is a state and the sound is the observation). In HMM the goal is to estimate the number of states, the state transition matrix, and the state-dependent distributions of observations. A more advanced Latent Semantic Indexing method models implicit meanings behind texts and has been used extensively in research, including our own, to classify and categorize documents (i.e., automatic groupings). A general assumption is that the documents come from a set of unknown categories and the words and phrases appearing in the documents are produced by category-dependent distributions. Given a set of observed data, these methods construct matrices that explain document generation and word/phrase generation using iterative optimization and/or matrix decomposition. Note that completely automated systems for converting text to meaning are not likely to succeed by themselves. However, intelligent data analysis provides not only a best classification based on similarity to correctly classified text information, but also a probability assessment of this correct classification. Thus we can discriminate between text that is clearly classified and text that may be incorrectly classified. By screening these texts, we can concentrate the high-cost human interaction on those messages that have uncertain classification, greatly limiting the cost involved in producing valid implementation process data.

**CASE STUDY: COMMUNITY DEVELOPMENT TEAMS TO SCALE-UP TREATMENT FOSTER CARE OREGON (TFCO) IN CALIFORNIA**

**Background**

Each year, 87,000 children and adolescents are placed in group, residential, and institutional care settings in the United States (http://www.childwelfare.gov/pubs/factsheets/foster/), with over 15,000 in California. While there is some quasiexperimental evidence to support positive short-term effects for highly structured and individualized group care models such as Teaching Family Homes, the majority of studies have linked placement in group and residential care
with increased odds of an array of negative outcomes such as increased odds of association with delinquent peers, delinquency, isolation from family, and lowered odds of reunification. Group home placements are also expensive, consuming 43% of the substitute care dollars in California in 2001. Treatment Foster Care Oregon (TFCO) was developed as an alternative to group and residential care for children and adolescents with severe emotional, behavioral, and mental health problems being placed through juvenile justice and child welfare systems. In TFCO, the child/teen is placed singly in a highly trained and supported community foster home where they and their family (biological, adoptive) receive intensive clinical services. A series of randomized trials have shown significantly better outcomes for participants in TFCO versus group care, leading TFCO to be designated as a top tier evidence-based model by multiple scientific advisory boards and organizations. TFCO has been implemented in over 90 agencies in the United States and Europe since 2001. However, these agencies likely do not reflect typical publicly funded service systems; they are early innovators. In fact, it is estimated that 90% of public systems do not implement evidence-based practice models. It is these non-early adopting agencies that are the focus of this case study.

Context

The California Institute of Mental Health (now California Institute for Behavioral Health Solutions; CIBHS) originated the Community Development Team (CDT) model to increase the number of California counties that successfully adopt, implement, and sustain evidence-based practices. CDTs operate through several well-specified mechanisms including multicounty team meetings, expert local consultation, peer-to-peer exchanges, and regular multicounty conference calls. Forty California counties that were non-early adopters of TFCO were invited to participate in an NIMH-funded randomized trial to scale-up TFCO in California, and all agreed to be part of this trial. After matching counties on demographic characteristics (e.g., size, % minority, poverty, previous use of mental health funding) into three cohorts of 12 to 14 counties, a two-step randomization process was used. Counties were first randomized to condition (CDT or Individualized or standard ‘as usual’ implementation). Then they were randomized again to one of three timeframes for implementation start date that spanned across 3 years. Randomization to start date allowed the management of issues related to implementation capacity, as training all counties at once was impossible. The random assignment of counties to cohort also increased protection against the influence of exogenous factors.

Measurement and Analytic Framework

A primary research question relates to whether participation in the CDT condition improves the adoption, implementation, fidelity/adherence, and sustainability of TFCO. Secondarily, contextual and organizational factors are hypothesized to mediate the association between experimental condition and implementation outcomes. For example, as a result of participating in the CDTs, counties are expected to make better progress due to more positive attitudes toward the TFCO model and supportive organizational climates. However, regardless of experimental condition, those counties with higher scores on the hypothesized mediators are expected to achieve better implementation outcomes.

One of the main implementation outcomes to be examined is how long it takes for a county to place their first child in a TFCO home, an event that comes after the decision by multiple social service systems in the county to adopt TFCO, the selection and training of agency workers to support TFCO, and the selection and training of a foster parent in TFCO. Survival analysis techniques were used, including Cox regression modeling, to compare the time it takes for placement for CDT and standard setting counties. Survival analysis is well suited for these data, as cohorts will vary in the amount of time that they have been involved in the study, thereby creating an outcome measure that is right-censored. Survival analyses will use the entire time period available for each cohort (4.5 years for Cohort 1; 2.5 years for Cohort 2; and 1.5 years for Cohort 3). By modeling how the hazard rate depends on intervention status and other covariates, a formal test can be conducted to assess the CDT intervention impact. An unusual feature in this study is that the outcomes for counties in the CDT groups could be correlated because they work together in a peer-to-peer setting. In contrast, those counties assigned to the standard implementation condition are expected to have outcomes that are independent. To account for how this clustering effect in the CDT group affects the
standard error and testing of the intervention, the Generalized Estimating Equation sandwich-type variance estimator will be used to adjust for non-independence in Cox regression modeling, using techniques similar to those employed previously in schizophrenia studies where family factors caused clustering.138–140 These methods correct for nonindependence; test statistics can be based on exact tests where the distribution under the null is simulated and critical values are thereby obtained.

In addition to the time to first placement, implementation progress is measured using a Stages of Implementation Completion (SIC) scale developed for this trial that includes both time-based and quality indicators of completion of each of eight implementation stages: (1) Engagement, (2) Feasibility, (3) Readiness planning, (4) Staff hired/trained, (5) Fidelity monitoring system in place, (6) Services and consultation begins, (7) Adherence and competence tracked and feedback, and (8) Certification/licensure. Activities are specified at each stage and measured using dates of accomplishment and quality ratings. Progression through each stage involves unique (although sometimes overlapping) groups of constituents. For example, leaders of child welfare, mental health, and juvenile justice systems first explore the possibility of implementation (stage 1), and access the feasibility/fit of the model for their local circumstances (stage 2). If the determination to proceed is made, system leaders are likely to step back from the process and involve others who will be directly involved in the active implementation and in the planning (stage 3). During active implementation when staff are hired and trained (stage 4) and fidelity monitoring is set up (stage 5), intervention and agency staff are the primary agents involved, not the system leadership. Therefore the SIC is populated by data from a variety of agents at the various stages. Quality of participation ratings are made within several of the SIC stages.15,66 Interestingly, this real-world implementation instrument reflects, quite closely, the theoretical model developed by Aarons et al.1 and adds considerable support for the importance of different factors at various stages of the implementation process.

In addition to the SIC time to completion and quality of completion ratings, qualitative measures aimed at adding to the understanding of what factors influence decision makers to adopt evidence-based models are being examined. Palinkas found that the social networks of system leaders played a key role in their decision to adopt during the exploration stage. Network size and density (the number of reported links divided by the maximum number of possible links) was not associated with the size of the county, but was significantly associated with stage of participation; individuals in counties that were considering participation in the Cal-40 study had larger and more dense networks than individuals in counties that had already made a decision to participate or not participate.141 For those who had agreed to participate or were considering participation, information about TFCO and the Cal-40 study was obtained from presentations given by CIBHS representatives at state or regional meetings, direct contact by CIBHS with county agency directors, direct contact by other agency directors within the county, or staff within the agency. In addition, how leaders interpret and make use of evidence is being examined, including what sources of information they find credible, where they obtain information, and what types of evidence they see as relevant. The findings from the qualitative measures will enhance and extend the SIC outcomes by providing insights into specific mechanisms that drive decisions to adopt/not adopt.

Design

This implementation trial began by inviting all California county service agencies that had not previously adopted TFCO to participate. Of the 40 so-called non-early adopting California counties invited to participate, none declined and 39 consented to at least consider implementing TFCO. No counties objected to the randomization to condition, but several noted that the timeframe to which they were assigned did not fit their circumstances.142 Counties with more children placed in care and those that had more positive organizational climates consented to participate more quickly.18 The study was extended to include 11 additional Ohio counties that are not early adopters of TFCO. Recruitment, enrollment, and randomization methods used in Ohio mirrored those used in California. The design used a roll-out head-to-head comparison of two implementation strategies, a standard approach that engaged individual counties, and a Community Development Team learning collaborative consisting of six to eight counties working together. Both arms delivered TFCO. Counties were assigned randomly to which implementation strategy and which year they began implementing.
TFCO. The SIC was used to assess the duration, speed, quality, and quantity of implementation. Because of these multiple dimensions, a composite score was chosen, derived from principal components analysis, as our primary outcome. Analyses used mixed models with random effects to take into account cohort effects as well as clustering of counties inherent in the Community Development Team approach, and readers are referred to the primary outcome publication for further details.16

**SUMMARY**

A wide variety of dissemination and implementation designs are now being used to evaluate and improve health systems and outcomes. This chapter discusses randomized and nonrandomized designs for the traditional translational research continuum or pipeline, which builds on existing efficacy and effectiveness trials to examine how one or more evidence-based clinical/prevention interventions are adopted, scaled up, and sustained in community or service delivery systems. The chapter also considers other designs, including hybrid designs that combine effectiveness and implementation research, and designs that use simulation modeling. A case example of a recent large scale implementation study is presented as an example of measurement and design considerations in dissemination and implementation research. The chapter provides suggested readings and websites useful for design decisions.

**ACKNOWLEDGEMENTS**

Support for this chapter was provided by the following grant funded by the National Institute of Mental Health and the National Institute on Drug Abuse: P30 MH074678, P30 DA027828, DA019984, MH040859, DA024370, P30 MH068579, and MH080916.

**SUGGESTED READINGS AND WEBSITES**

Readings


This article describes four recommended changes in the use of health care evidence that would speed along health care and practice improvements: (1) use a range of scientific methodologies, considering both mechanisms and contexts; (2) reconsider thresholds for action on evidence, making incremental changes; (3) reconsider concepts of trust and bias; and (4) engage both academics and patient caregivers with respect.


This article provides an overview of state-of-the-science experimental and quasiexperimental research designs for D&I. Three broad categories are discussed: (1) within-site designs; (2) between-site designs; and (3) within- and between-site comparison designs.


This article describes a modification to the traditional wait-listed design, in which random assignment occurs multiple times during a trial, thus enabling subjects to receive the intervention. Still, these designs can only be used to assess short-term impact, and there is no control group left as a comparison.


Suggests adaptations to the traditional randomized trial. In this article, “adaptive design” refers to a trial in which characteristics of the study, such as assignment to an intervention or control group, change during the data collection process.


This paper presents the concept of hybrid designs that are especially appropriate for implementation research and discuss the three basic types of such designs.


This early work from Flay provides a useful overview of the concepts of efficacy and effectiveness. While the science of translational research has evolved considerably over the past three decades, many of the principles from Flay’s article remain highly relevant.


Building on the influential work of Tunis et al. on practical clinical trials (PCTs), this article provides examples of conducting PCTs with enhanced external validity, without sacrificing internal validity. The authors suggest that in order to reduce the gap between academia and real-world practice, it
Design and Analysis in Dissemination and Implementation Research

is necessary to increase the relevance of PCTs for appropriate audiences.


This paper discusses the need for the development of methodological approaches to the study of implementation processes and tests of implementation strategies through a structured review of nine studies. The authors present limitations of randomized designs and potential design alternatives to consider.


In this chapter, Pierre considers the recent push to use randomized experiments for obtaining evidence. He offers recommendations for using and designing randomized experiments, as well as how they should be executed and evaluated.


In this chapter, Reichardt & Mark consider four quasiexperimental designs: before-after, interrupted time series, nonequivalent group, and regression-discontinuity. They describe the strengths and weaknesses of each design, as well as threats to validity.


The authors consider alternatives to the randomized controlled trial that also allow for drawing causal inferences. They describe the strengths and weaknesses of each design, including threats to validity and the strategies that can be used to diminish those threats.

Selected Websites and Tools

http://wtgrantfoundation.org/resource/optimal-design-with-empirical-information-od

**Optimal Design** is a software package, developed by Stephen Raudenbush and colleagues, which helps researchers determine sample size, statistical power, and optimal allocation of resources for multilevel and longitudinal studies. This includes group-randomized trials, also called setting-level experiments. Version 2.0 was released in summer 2009. The software, a description of the updates from the previous version, and a manual containing software documentation are available for download.

http://www.rmass.org

The RMASS program computes sample size for three-level mixed-effects linear regression models for the analysis of clustered longitudinal data. Three-level designs are used in many areas, but in particular, multcenter randomized longitudinal clinical trials in medical or health-related research. In this case, level 1 represents measurement occasion, level 2 represents subject, and level 3 represents center. The model allows for random effects of the time trends at both the subject level and the center level. The sample size determinations in this program are based on the requirements for a test of treatment by time interaction(s) for designs based on either subject-level or cluster-level randomization. The approach is general with respect to sampling proportions and number of groups, and it allows for differential attrition rates over time.

http://cepim.northwestern.edu

The Center for Prevention Implementation Methodology for Drug Abuse and HIV (Ce-PIM) website provides a range of publications and presentations on implementation principles, measures, designs, and analyses

http://epoc.cochrane.org/

The Cochrane Effective Practice and Organisation of Care (EPOC) Group is a Review Group of The Cochrane Collaboration—an international network of people helping health care providers, policymakers, patients, their advocates and carers make well-informed decisions about human health care by preparing and publishing systematic reviews. The research focus of the EPOC Group are interventions designed to improve the delivery, practice, and organization of health care services. The EPOC editorial base is located in Ottawa, Canada with satellite centers in Norway, Australia, and England.

https://cyberseminar.cancercontrolplanet.org/implementationscience/

The Implementation Science Webinar Series is sponsored by the National Cancer Institute (NCI) Division of Cancer Control & Population Sciences. In 2013 the Implementation Science Team started a webinar series focused on advanced dissemination and implementation research topics, including design and analysis issues. Each session includes approximately 40 minutes for presentation(s) by leaders in the field as well as 20 minutes for engaged discussion and Q&A. The website lists topics and session titles for upcoming sessions or archived sessions.

**REFERENCES**


3. National Research Council and Institute of Medicine; Committee on Prevention of Mental Disorders and Substance Abuse Among Children, Youth, and Young Adults: Research Advances and Promising Interventions; Board on Children, Youth, and Families; Division of Behavioral and Social Sciences and Education, Mary Ellen O’Connell, Thomas Boat; Warner KE, eds. Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities. Washington, DC: National Academy Press; 2009.


56. Effective Practice and Organisation of Care (EPOC). EPOC Resources for review authors. Oslo: Norwegian Knowledge Centre for the Health Services; 2015: http://epoc.cochrane.org/epoc-specific-resources-review-authors.


84. Curran GM, Mittman BS, Landes S, Pyne J. Effectiveness-Implementation Hybrid Designs: Clarification, Refinements, and Additional Guidance Based on a Systematic Review and Reports from the Field. 8th Annual Conference on the Science of Dissemination and Implementation in Health; December 2015, December 2015; Washington, DC.


Hofmann T. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval; 1999; Berkeley, California, United States.


Design and Analysis in Dissemination and Implementation Research

General. Rockville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Injury Prevention and Control; Substance Abuse and Mental Health Services Administration, Center for Mental Health Services; and National Institutes of Health, National Institute of Mental Health; 2001.


