1 INTRODUCTION

This book provides researchers in education and related fields with the knowledge and tools they need to design small efficacy studies. We call these ***small efficacy studies*** because they focus on determining if interventions work in ideal circumstances, as opposed to ***effectiveness studies*** (sometimes called pragmatic trials), which tend to be focused on "real-world" implementation of an intervention. This is not to say that efficacy studies do not take place in the real world—indeed, we assume that they will take place in schools and classrooms—but instead that by nature of the study being small, there will be more researcher control than under widespread adoption of the intervention.

The book is divided into four sections, each with several chapters. Each section could stand on its own, though we introduce notation and a variety of terms in Section I:

- **Section I:** Background Concepts
- **Section II:** Randomized Designs
- **Section III:** Quasi-Experimental Designs
- **Section IV:** Tools and Reporting

Throughout the book, we assume that the reader already has an intervention in mind that they would like to evaluate and that this intervention is well developed. In this section, we provide resources related to the development of interventions for those who may need these. This review is brief not because these concepts and processes are unimportant but because they are not the focus of this book. We also introduce a series of examples that we will use throughout this book to give an illustration of the range of interventions that this book could be used for evaluating, as well as the inspiration for these examples. Readers should note that while all these examples are focused on education, not all take place in schools.

A. WHAT IS AN INTERVENTION?

This book is about the evaluation of new curricula, programs, trainings, policies, and practices. Throughout the book, we refer to these as ***interventions***, a word we choose since it indicates that an existing, current practice is *intervened upon* by this new program, policy, or practice. Throughout this book, we also refer to the alternative practice—often the current practice ("business as usual")—as the ***comparison***. This has the benefit of being vague and broad enough to encapsulate a variety of conditions.

We have chosen these words instead of ***treatment*** and ***control***, which are far more common in the field of **statistics**, as they strike us as far too clinical for the types of interventions those in education are likely investigating. At the same time, we have chosen to retain use of *treatment* when discussing **statistical models,** including the average treatment effect, which is the **parameter** of focus in small efficacy studies. We do so to be consistent with the broader statistical literature, so that readers of this book can easily make connections between this book and other texts.

A.1 Interventions and Equity

While we have settled upon the word "intervention," we realize this isn't ideal, either, since the word can suggest power imbalances. Often, those doing the intervening (e.g., researchers, policymakers) are not members of the communities affected by the interventions. In education, these interventions far too often reflect the desires and norms of the dominant culture, framing current practices in the communities being intervened upon as deficient ("deficit based"). In research, sometimes these interventions have not included adequate consent and even involved coercion.

At the same time, we have seen that locally grown, community-based intervention is still possible. Teachers and schools often develop their own curricula, and communities often develop their own programs. Indeed, many of the conversations and questions that led to this book were from those with small, locally grown programs that wanted to be able to evaluate if their programs worked but with considerably smaller budgets. These locally grown policies and programs are, by our definition, also called interventions—they are changes to existing practices.

Nonetheless, we realize that the word "intervention" can be fraught, and yet, we have struggled to find an alternative to "intervention" that can be consistently used in our text. Some have suggested using a variety of words throughout the text instead of sticking with a single word, but from our viewpoint as statisticians, it is important for clarifying concepts if we have a clear and consistent vocabulary and set of symbols. Thus, we will continue to use "intervention" and "comparison" throughout, though with some trepidation.

Finally, we commend readers to pay attention to the growing movement in both the scientific and evaluation communities to consider concerns with equity throughout the research process. This is referred to as concerns with "ethics" in some places (e.g., Asiedu et al., 2021) but as "equitable evaluations" (e.g., Balu et al., 2023; Cerna et al., 2021; National Center for Educational Research, 2022) and "culturally responsive evaluation" (e.g., Hood et al., 2015; Kushnier et al., 2013) in others.

A.2 Intervention Core Components

In the pharmaceutical world, interventions are very clearly defined (e.g., they are very clearly defined dosages of specific drugs). In education and related disciplines, this is not often the case. Certainly, some interventions are packaged neatly (e.g., guaranteed income programs, science kits containing laboratory supplies), but many others are more fluid (e.g., teaching practices, mentoring relationships). Regardless, it is important that the intervention being studied is clearly defined. This means determining boundaries around what constitutes the intervention and what does not. Sometimes an intervention is not ready for an evaluation because these clear definitions of practices and components still need to be worked out. Other times, ensuring that the intervention can be implemented in the desired contexts needs further study. These alternatives are discussed in Chapter 3.

One tool that can be helpful for researchers to elicit features of the proposed intervention is to develop a logic model. This model includes a series of boxes and arrows, including elements related to the problem statement, assumptions, strategies and activities, resources, short- and long-term outcomes, impacts, and outputs (e.g., Shakman & Rodriguez, 2015). Developing this logic (or "intervention") model involves identifying direct components and support components (requirements of the existing system for the intervention to be implemented). For each component, researchers should plan for its content, quantity, mode, and quality (Weiss et al., 2014). Section II of Hill et al. (2023) provides an overview of each of these, as well as tools and examples.

A.3 Business as Usual

Interventions are meant to change a current system or set of practices or policies. Since the effect of an intervention is always relative to this current practice, it is essential that researchers conducting small efficacy studies understand the system and practices they are trying to intervene upon. As Thomas and Klopfenstein (2020) note, this interest in the comparison condition and current, regular practices in schools and communities is far from common in intervention research. Bryk et al. (2015)—and the broader field of "improvement science"—show that it makes little sense to jump to solutions without adequately understanding the problem being solved. These concerns are particularly important since what is an intervention now may become common practice later; that is, what we compare an intervention to is itself changing over time and place (e.g., Lemons et al., 2014).

For those designing and testing interventions, this means paying careful attention to how schools (or the institutions under study) are structured, which courses are taught, who teaches them, which curricula and products are commonly used in classrooms, and what student backgrounds and experiences are common in the environments under study. We will return to these concerns as they relate to **external validity** in Chapter 4. It can be helpful to turn to data on these practices, whether from state data systems, surveys, or, even better, classroom observations. This is a place in which qualitative research can play an important role in improving an intervention and thus an **efficacy study**. In the implementation guide by Hill et al. (2023), this is also investigated more formally by identifying the core components of an intervention and then measuring these components within comparison ("business-as-usual") schools. Researchers are often shocked to find that many of the practices of their intervention are being implemented in schools already, though maybe under different names or forms. In our experience, it is far better to know this before beginning an efficacy study than to discover this in the field.

B. EXAMPLES

Throughout this book, we will provide examples regarding trade-offs between different research designs. Our hope is that the examples will make these decisions and trade-offs more concrete. To that end, we focus throughout on five case studies, which we introduce here. These case studies are based on real interventions—often conglomerations of several—that have been developed for children, often in schools. We don't think that these will cover *every* possible situation, but they address many of the intervention types that we have run into in our own consultations with researchers conducting small efficacy studies.

We include here a range of ages—from Pre-K to high school. We have also included a range of programs—from full curricula to EdTech games to after-school programs. While most of this book is focused on schools, many interventions focus on informal learning environments, so we include two examples outside of schools—one a museum (Study 4) and one a cultural center (Study 5). We have tried to vary the students of focus throughout, too, and include one example with a **population** that is important but likely small (Native American students). Most of our examples focus on mathematics and science interventions, though the same ideas could apply to other intervention types. Our goal is to indicate the range of possible intervention types and how different types of studies of these could be conceived. Each example is found in several but not all chapters.

CASE STUDY 1: ELEMENTARY SCHOOL PROJECT-BASED LEARNING PROGRAM

Over the past 5 years, Dr. Wanda Smith and colleagues have been working closely with Detroit public schools to develop an innovative, hands-on science program for elementary school teachers and their students. During this time, Dr. Smith has worked closely with teachers in three elementary schools to develop and refine a professional development program and science units focused on environmental science and biology with children in Grades 3 to 5. These units are hands-on and project based and focus not only on scientific knowledge but also on centering students as scientists and observers in their communities. Over time, the team has developed five science units, associated activities, and teaching guidelines. Teachers are deeply engaged in this work, and observations in the classrooms suggest that students are, indeed, making connections between science and their lives. Student knowledge of science seems to have increased, as does student motivation. This example is inspired by Barron et al. (1998).

	Case Study 1
Population	Elementary school teachers and students in Grades 3 to 5 in elementary schools in Detroit serving high percentages of low socioeconomic status, majority Black students
Intervention	Professional development program (teachers) + science units
Comparison	In these schools, science is otherwise taught using textbooks.
Outcome	Motivation, science test

CASE STUDY 2: MIDDLE SCHOOL EDTECH MATHEMATICS GAME

Over the past decade, a research team at WXY Research Inc. has been interested in the possibilities of interactive educational technologies for improving student understanding of middle school math concepts, particularly for schools and students in rural areas. This interest was spurred by a partnership with the Rural Schools Collective, an organization connecting rural schools with professional development and curriculum resources. The team has been focused on the development of a building app ("BuildFrac") that teaches kids fractions, decimals, and percentages through the process of building a city. The app involves both lessons and games and implements state-of-the-art cognitive science findings to help kids develop a deep understanding of these concepts. The program has been refined in partnership with several schools and is now beginning to be used in these schools more broadly. Preliminary findings suggest that over time, students do improve in their understanding of these concepts, particularly when students use the app more than an hour a week and when it is paired with in-class supports. This example is inspired by Roschelle et al. (2000).

	Case Study 2
Population	Middle school students in rural middle schools
Intervention	"BuildFrac" app (available on iPads)
Comparison	As a supplemental program, this is unclear.
Outcome	Understanding of fractions, decimals, and percents

CASE STUDY 3: HIGH SCHOOL AFTER-SCHOOL ROBOTICS PROGRAM

Dr. Carlos—a computer scientist—and his students have been running an after-school outreach program focused on robotics at a nearby high school in Atlanta. This high school predominately serves Black students—who are historically underrepresented in computer science—and the students come from both middle-class and low socioeconomic status families. To be admitted to the program, which is free, students had to apply and be recommended by a teacher. Once admitted, students in the program meet once a week for 2 hours for 16 weeks. During the lessons, they learn to program and build robots to complete tasks, and at the end of the program, they take part in a mini-"competition." The research team has been excited to see how the students have responded to the program, including how this has affected their interest in computer science in general, as well as in studying a STEM-related field in college. Teachers have reported that students in the program are more engaged in their science classes. This example is inspired by Barker and Ansorge (2007).

	Case Study 3
Population	Black high school students interested in robotics in an Atlanta public school
Intervention	After-school robotics program, 16 weeks × 2 hours
Comparison	Unclear
Outcome	Belonging (in STEM), science grades in school, interest in applying to college

CASE STUDY 4: MUSEUM SCIENCE SPATIAL PROGRAM FOR PRESCHOOLERS

A STEM museum in Florida has had a longstanding partnership with two researchers at a nearby university. Drs. Andrews and Zheng are interested in how informal experiences in museums may encourage families to engage with their children in ways that inspire and support STEM learning at home. This interest has led them to work with the museum staff to design a series of exhibits that foster interactions (while at the museum), while also providing materials to take home, with the hope of continued engagement. Understanding if this program actually improves these connections is important, given the continued cost of these take-home materials. This study focuses on a particular exhibit that focuses on fostering the development of spatial skills through a series of building activities. This example is inspired by Marcus et al. (2017).

	Case Study 4
Population	Children ages 3 to 5 who visit the STEM museum in Florida
Intervention	Construction exhibit, focuses on spatial skills
Comparison	Unclear
Outcome	Unclear

CASE STUDY 5: NATIVE AMERICAN CULTURAL SCIENCE PROGRAM FOR FAMILIES

Native Americans are underrepresented in STEM fields. Over the past several years, Dr. Lee has developed a partnership with an American Indian Center in Washington focused on creating connections between indigenous and Western science. Together, the team designs a curriculum that is relational and place based, focusing on the local ecosystem. This curriculum included hands-on experiences, apprenticeships with the researchers, and the development of science units for nearby schools. This example is inspired by Bang et al. (2010).

	Case Study 5
Population	Community members who take part in activities at this American Indian Center in Washington state
Intervention	A Saturday Science program that meets 6 weeks × 2 hours a week, including field trips and hands-on activities
Comparison	Unclear
Outcome	Belonging, interest in science as a career

KEY TERMS

Efficacy study
External validity (generalizability)
Parameter
Population
Quasi-experiment
Statistic
Statistical model (structural, stochastic parts of)

SECTION I

BACKGROUND CONCEPTS

2 INTRODUCTION TO SECTION I: IS A SMALL EFFICACY STUDY RIGHT FOR YOU?

The first section of this book provides a common language regarding different types of research designs (Chapter 3), validity (Chapter 4), statistical models (Chapter 5), measurement (Chapter 6), and questions of **effect size** (Chapter 7). Each chapter provides an overview of concerns that researchers designing such studies will face and offers insight into how to address these concerns. Overall, these chapters will convey that a single study—and a small one at that—cannot do everything well. For this reason, it is important that the researchers know going into the study what trade-offs they are willing to make and how the data collected will enable them to answer the questions they care about.

In this chapter, we provide an overview of these topics. To do so, we begin by elucidating the logic of an efficacy study, how to operationalize this, and specifically how the constraint of a "small" sample affects this operationalization. We conclude the chapter with a discussion of alternative research designs—other than efficacy—that researchers conducting small studies might consider. These designs answer other questions—questions that are equally valid and important but differ from those asked in an efficacy study, nonetheless.

A. WHAT IS THE LOGIC OF AN EFFICACY STUDY?

The purpose of an efficacy study is to determine if an intervention works under some (typically ideal) conditions. Such a design is often referred to as a "trial" because it is meant to be a test—ideally, a rigorous, falsifiable test—of the scientific hypothesis that the intervention works. The fact that this is falsifiable means that at the end of the study, the intervention will "pass" or "fail" the test. Passing should indicate that the intervention "works" and that it should continue to be studied (or, in some cases, adopted). Failing should indicate the opposite—that this intervention is highly unlikely to work and that in its current form, research should not continue. This harsh distinction—between passing and failing—is exactly why the study needs to be carefully designed.

It is helpful here to contrast a strong versus a weak study design. In the latter, this distinction between "passing" and "failing" is muddled. Suppose a study ends and finds that the intervention effect estimate is small and not statistically significant. But instead of concluding that it fails, in a weak design, the researcher is left with more caveats and questions. Perhaps the intervention did not pass because the outcome was not measured well, the sample size was too small, or the intervention was not implemented well. Certainly, this doesn't mean that nothing was learned from the study. But much of this learning could have been had with fewer resources—time and money—than via an efficacy study.

Importantly, the costs of such a weak research design are not simply for the research study itself. An efficacy study is often just one study in a larger possible progression of research in an area. For example, a successful efficacy study—showing positive effects—is often required before replication or effectiveness studies can be funded. These replication and effectiveness studies focus on how an intervention would fare under less researcher control and in broader use in schools—the goal for much intervention research in education. A weakly designed efficacy study thus has ethical implications, as it can halt the ability to continue research in an area, leading possibly impactful interventions to never make it to the students and schools that could benefit from them.

For this reason, it is imperative to approach an efficacy study with an understanding of what is possible to learn and what is not. Throughout, it is helpful to keep the two possible outcomes of the study in mind: passing and failing. A design should be strong in the sense that if the intervention does *not* pass the test, you can be confident it means that this intervention is not quite right, and work on it (in its current form) should not be continued. Conversely, if it *does* pass the test, the evidence should solidly indicate that the intervention—and not some other process or program—shows promise and that research should continue.

B. HOW CAN THIS TEST BE OPERATIONALIZED?

Rigorously testing a theory—including an intervention—requires operationalization. The general question, "Does this intervention work?" can be operationalized an infinite number of ways. For this reason, it is important to be specific. In general, researchers have found it useful to operationalize this question using the PICOS framework:

- **Population:** Who is the intervention intended for? In some cases, the intervention might be intended for a broad audience (e.g., all elementary school students). In others, this population might be very specific. Are there some with whom the intervention is more closely aligned?
- **Intervention:** What *exactly* is the intervention? What are the components? What constitutes it being implemented "well" versus not? Are some components essential?
- **Comparison:** What would happen in the absence of this intervention? When the intervention was developed, was it intended to replace an existing approach? Be an add-on? Are there some comparisons for whom the intervention likely wouldn't work?
- **Outcomes:** When we say the intervention "works," we mean that it causes a change for at least one important outcome. Which are the most important? How can this be measured well?
- **Setting:** How will the intervention be delivered? In schools, after-school programs, in libraries? Is there one of these that is where the effects are expected to be strongest?

For more information about the PICOS framework see Higgins and Greene (2011). PICOS is not the only framework available, of course. Another framework is that of

MUTOS, wherein a study is specified in relation to the **Methods**, **Units** (akin to population), **Treatment** (akin to intervention), **Outcomes**, and **Setting** (Aloe & Becker, 2008). And yet another is the **Who** (akin to population), **What** (akin to intervention), **When** (regarding time), **Where** (akin to setting), and **How** (akin to methods) framework of Reichardt (2011). While the exact words differ across these, notice that they share a focus on clearly defining the specificity of a study.

If there are many possible PICOS that could map onto a broader question, how should a researcher choose one over another? Our advice here is to be careful and mindful of the rigorous test at the end: Is there a PICOS in which if the intervention cannot succeed, you would conceive of it as evidence that research should not move forward? Think of this as a *minimal sufficiency* requirement—if the intervention operationalized in this way cannot pass the test, then there is no reason to believe that any other version of this could be passed as well.

C. WHAT ABOUT SMALL EFFICACY STUDIES?

The focus of this book is on *small* efficacy studies—those that can be conducted with roughly 10 or fewer schools. This small sample constraint comes with a cost, however. As future chapters will indicate, for rigorous tests to be conducted with small samples, the PICOS studied need to be narrowed. In general, small samples require there to be less variation in the measures used, the population studied, and the intervention's implementation. Additionally, small samples require that the expected effect of the intervention (if it "works") is moderate to large. This means that the intervention needs to be quite different from its comparison and that the outcome measure needs to be aligned well with the intervention (e.g., proximal measure). Thus, there are several, real trade-offs at play. Here we consider two of these (for those not familiar, in Chapter 4, these validity concepts are discussed in more depth):

C.1 Statistical Conclusion Versus External Validity

As we will discuss in Chapter 5, to increase design sensitivity (including statistical power), residual variation needs to be reduced. In many designs, this involves testing the intervention in a homogeneous sample of schools—that is, schools that, without the intervention, would have similar average outcome scores. Such a focus on a homogeneous sample, however, means that the results of the study will be relevant to only a small portion of the population for whom the intervention may ultimately be useful. In Chapter 4, this concern with external validity of the study will be discussed.

C.2 Statistical Conclusion Versus Construct Validity

The need for a large effect size requires the intervention to be implemented well and for the measure to be closely aligned. The requirement that it is implemented well means that it needs to be well developed and that it is possible to measure and distinguish good versus poor implementation. The requirement for strong implementation certainly involves more work, as well as additional measurement. Additionally, the requirement that the outcome is measured well and aligned with the intervention produces a similar tension. At one extreme is the most aligned measure—one that simply asks participants if they received the intervention (e.g., "What color was the workbook used?"). While leading to a large effect, this does not have strong construct validity (see Chapter 4)—it does not get to the core goal of the intervention. That is, is the purpose of the intervention to have received an intervention—or is it to have changed some underlying educational outcome (e.g., increase knowledge about [topic])? While this is an extreme case, decisions regarding proximal versus distal measures here involve exactly this trade-off; Chapters 3, 5, and 6 will address these.

In both cases, these trade-offs boil down to concerns with the degree to which the particular—this study—addresses the larger, broader question of interest. This question is not technical or explicitly statistical—it is instead philosophical and scientific. In contrast, this book focuses largely on how to design a study so that the effect estimated is not just correlational but causal, and so that a test of this effect is statistically strong. We note this here because the broader questions are essential to science, but they are not the focus of most of this book.

D. WHAT QUESTIONS CAN'T A SMALL EFFICACY STUDY ANSWER WELL?

As we have argued, when done well, a small efficacy study can provide strong evidence that a specific version of an intervention does or does not on average change a specific outcome in a specific population. Notice that this phrase includes a variety of caveats. Here we turn these into questions to clarify what such a small efficacy study does *not* do well. Here we focus on two.

D.1 Which Version of the Intervention Is Better?

Sometimes, researchers early in the development of an intervention want to know: Does Version A or B work better (or does A + B work even better)? The problem here is twofold. First, remember that in a small study, we need a large effect. This means that if we want a strong test of A versus B, we need to expect that their outcomes will, on average, be very different from one another. If these are two versions of the same underlying intervention, this need for a large effect can be difficult to impossible to achieve. Second, we could instead say, "Well, I just want to know if A works and if B works." But notice that this essentially means running *two* small studies—which together are no longer small.

D.2 Which Subgroup Does the Intervention Work Better For?

Another question researchers developing an intervention often have is if the intervention results in stronger effects for one subgroup of students (or classrooms or schools) compared to another. But this question is about understanding heterogeneity—the degree to which treatment effects vary—which is at odds with the need to reduce variation (increasing homogeneity) to improve statistical power. This results in the same two problems as in the previous question. That is, to estimate two subgroups (A and B) well, the study would need to be *twice* as large. Additionally, to test differences between the subgroup, average effects in A versus B would require *substantial* variation in effects.

For both questions, the problem is that comparisons of average treatment effects—across versions of an intervention or subgroups—result in sampling variances (squared standard errors) that are *twice* as large. In a small sample, this means that tests of these differences are nearly guaranteed—from the beginning—to not be statistically significant (see Chapter 4). But if they are guaranteed to not be statistically significant, what is the reason to conduct the study? For this reason, if sample sizes must be limited, we do not encourage the efficacy study design for this purpose.

E. WHAT ARE OTHER OPTIONS?

For those developing interventions, there is certainly a circularity here. To have a small sample, the efficacy study must be conducted under conditions that will produce a strong effect—but these conditions themselves cannot be well tested in the field. These questions—What (homogeneous) population is likely to see the largest effect of this intervention? What (single) version of the intervention is likely to have the largest effect?—at this stage need to be answered by *theory*. At their core, these questions have to do with two features:

- **Implementation**. It may seem too obvious to state, but for an intervention to work well, it needs to be implemented well. But implementation is complicated—it requires that it is *possible* for it to be implemented well (i.e., that someone trying hard could do it) and that it is *desirable* to be implemented well (i.e., that someone will want to do it). There are plenty of possible interventions that may in theory work but not meet one or both criteria. For an efficacy study to provide a strong test, these concerns with implementation need to be addressed upfront. Put another way, an intervention that has not yet been studied to determine if it can be implemented well is not ready for such a strong test. This means that an alternative to conducting an efficacy study is to conduct an implementation study (see below).
- **Current practice**. For an intervention to have a large effect, it needs to be qualitatively different from what students, classrooms, and schools would be doing otherwise. Again, this may seem somewhat obvious since an average treatment effect is a direct comparison between two groups: those receiving the intervention and those who are not. But in education, those who are not receiving the intervention are *doing something*. That is, children in schools are being taught math, reading, science, social studies, and a variety of other topics each day. Even outside school, children and adults have a variety of opportunities to learn—if they are not at your museum exhibit, for example, they might be at another. Or they might be reading a book. For an efficacy study to provide a strong test, the population that is studied needs to have a current practice that is sufficiently different from the intervention. Put another way, an intervention that is the same as the current practice is not ready for such a strong test. This means that before such a test can be conducted, researchers need to know their population well enough to understand both what is common practice and what practices might be found in any population they are considering for their study. Here there is no substitute for partnerships with schools and for descriptive research.

These concerns with implementation and current practice suggest a couple of other study designs that researchers may want to consider instead of conducting an efficacy study.

Notably, while efficacy studies largely focus on the collection of quantitative data and on the testing of hypotheses, these alternative study designs tend to focus less on hypothesis testing and more on intervention refinement and hypothesis generation. To this end, these alternative designs nearly always also include the collection of qualitative data using interviews, focus groups, and participant observation. These alternatives include the following:

E.1 Feasibility Study

Before an efficacy study can be conducted, researchers need to know if such a study is even possible. This is the purpose of a feasibility study. Bowen et al. (2009) provide a list of several areas addressed by such studies. These include the *acceptability* of the intervention, the *demand* for the intervention, the likelihood that the intervention can be *implemented*, the *practicality* of the intervention given resources and constraints, *adaptations* to the intervention that are appropriate, and the ability for the intervention to be *integrated* into existing systems. Notice that feasibility studies largely focus on the feasibility of the intervention for a particular population and setting more so than on exact elements of the study design.

E.2 Implementation Study

An implementation study focuses less on the outcome from an intervention and more on its ability to be implemented in the range of populations and settings in which an intervention could be used (e.g., with different types of schools, in different types of communities, or students with different backgrounds). The goal of implementation studies is to develop, iterate on, and revise the intervention to be robust to these different conditions. In such a study, the outcome would be the ability to implement the study well and may include understanding adaptations that work well and those that do not. There is a large literature on implementation methods that we encourage the reader to turn to. For an introduction to this literature, see Fixsen et al. (2005), Goodson et al. (2019), Meyers and Brandt (2015), Nelson et al. (2012), or Peters et al. (2013).

E.3 Pilot Study

A pilot study sits at the intersection of efficacy and feasibility studies. Like an efficacy study, it uses all the same procedures, from recruiting schools to implementing a study design (e.g., randomization of students) to measuring outcomes. However, the sample size used in a pilot study is small, and the purpose—like a feasibility study—is *not* to test a hypothesis but to determine if all the study design procedures can be implemented. In a pilot study, the goal is to identify possible problems (and address these) so that a later efficacy study can indeed be a strong test of the intervention.

F. MOVING FORWARD

In this introduction, we've provided an overview of a variety of questions and concerns that you will face when designing a small efficacy study. At this stage, some of the concepts and vocabulary provided may not be familiar to you. The remainder of Section I is focused on filling any gaps and providing a strong foundation with general problems of research design. Readers familiar with these concepts may desire to skip these chapters; if so, we urge those readers to first turn to Chapter 5, which provides the general statistical notation that will be used throughout this book.

KEY TERM

Effect size