

## METHODOLOGICAL ISSUES AND ADVANCES

# Pilot Trials in Health-Related Behavioral Intervention Research: Problems, Solutions, and Recommendations

Kenneth E. Freedland

Washington University School of Medicine in St. Louis

Pilot studies can help to pave the way for larger randomized controlled trials of health-related behavioral interventions. Unfortunately, there is widespread uncertainty and confusion about the kinds of studies that should or should not be called pilot trials, and about their relationship to other types of preliminary studies of behavioral interventions. The traditional conceptualization of pilot studies as “preliminary efficacy” trials has been especially problematic. This report identifies some common and problematic weaknesses in pilot trials. It also describes a strategy for preliminary research on behavioral interventions that can prevent these problems, and provides recommendations for researchers and reviewers.

*Keywords:* clinical trials as topic, peer review, pilot projects, preliminary data, research design

*Supplemental materials:* <http://dx.doi.org/10.1037/hea0000946.supp>

In ancient Greek mythology, Chaos was the primordial void that preceded the creation of the universe. In modern behavioral intervention research, Chaos is the primordial mess that precedes the creation of randomized controlled trials (RCTs). Unlike its ancient predecessor, ours is not a formless, featureless void; it teems with preliminary and pilot studies. Many of them are well-designed and informative, but too many others have problematic aims, designs, and impacts.

The chaotic nature of this realm of behavioral science causes uncertainty about the kinds of preliminary studies that should or should not be conducted, and about the preliminary findings that researchers report in grant applications when seeking funding for behavioral RCTs. It causes uncertainty among grant reviewers about the kinds of preliminary findings they should or should not want to see in RCT grant applications, and uncertainty among peer reviewers and editors about how to evaluate articles reporting preliminary intervention research or pilot trials. It leaves all of us with an early phase intervention research literature that has ample room for improvement.

This paper discusses deficiencies that diminish the scientific value of many pilot studies, and it describes alternative to traditional pilot trials that can help to improve early phase research and

foster more robust behavioral interventions. It also provides recommendations for authors, reviewers, and readers of *Health Psychology*. However, these issues are not unique to this journal or to the field of health psychology; they are widespread. Consequently, the recommendations are also intended for other journals, authors, grant applicants, and grant and article reviewers.

### Problematic Pilot Studies

#### “Preliminary Efficacy” Trials

Across many areas of intervention research, most pilot studies have been preliminary efficacy trials in which the primary outcome analysis is severely underpowered. Many of these trials yield null results (Shanyinde, Pickering, & Weatherall, 2011) and end up in the “file drawer” (Rosenthal, 1979), thereby halting work on promising and unpromising interventions alike. In contrast, pilot trials with positive results are often interpreted as providing “preliminary evidence of efficacy.” Many pilot trial reports conclude that there is a need for “future research” on the intervention, without mentioning whether anyone is actually planning to conduct this research. In some cases, the report concludes that the findings give the authors themselves a green light to proceed to a larger trial. If they ever propose a larger trial, they will probably use the efficacy effect size from the pilot study in their power analysis.


This approach has several serious drawbacks. First, only a small percentage of published pilot trials have ever led to a full-fledged RCT. This is common in many areas of health-related research. For example, a recent analysis of the orthopedic surgery literature found that only 10% of published pilot trials were ever followed by an adequately powered RCT. The authors of the pilot trials were

---

*Editor's Note.* Susan Czajkowski served as the action editor for this article.—KEF

---

This article was published Online First July 2, 2020.

Correspondence concerning this article should be addressed to  Kenneth E. Freedland, Department of Psychiatry, Washington University School of Medicine in St. Louis, 4320 Forest Park Avenue, Suite 301, St. Louis, MO 63108. E-mail: [freedlak@wustl.edu](mailto:freedlak@wustl.edu)

surveyed to identify reasons why they never conducted a larger trial. The most common response was that they believed that despite its small size and other deficiencies, their pilot trial had answered the research question, thereby obviating the need for a larger trial. Other common responses included failure to demonstrate the preliminary efficacy of the intervention, inability to meet recruitment targets, and lack of funding (Desai et al., 2018).

Comparable analyses of behavioral intervention pilot studies are not available, so to illustrate the situation in our own field, the present author searched Medline for RCTs of health behavior interventions that were published (or published online ahead of print) in 2010 and that were described by the authors as pilot studies. The year 2010 was chosen to give the authors at least eight years to publish a larger RCT of the same intervention that they tested in their pilot trial. Medline was searched again for papers that cited the pilot trial, to identify a larger trial by the same authors. As shown in Table 1 in the online supplemental materials, the search identified 11 self-described randomized controlled pilot trials of interventions for smoking, diet, physical activity, or medication adherence. Williams et al. (2014) is the design paper for an RCT for which Williams et al. (2010) was the pilot study. The RCT's status in clinicaltrials.gov is "completed," although the results have not yet been published. Several of the other pilot studies were followed by further intervention development research or pilot studies, but none of them has been followed so far by a published RCT. Thus, only 9% of this small sample of pilot studies have been followed by larger RCTs.

Second, the positive findings of small preliminary efficacy trials are more likely to be false than true (Ioannidis, 2005), and they do not guarantee that the results of larger trials will also be positive. This has been demonstrated repeatedly, even in areas of medical research in which the "small" trials tend to be larger than many of our relatively "large" trials. For example, many Phase II trials of heart failure medications have enrolled hundreds of patients and have yielded positive findings, only to be followed by much larger Phase III trials that fail to show evidence of efficacy or that even show evidence of harm (Vaduganathan, Greene, Ambrosy, Gheorghiu, & Butler, 2013). Much smaller pilot trials are even worse predictors of favorable RCT outcomes.

Third, severely underpowered preliminary efficacy trial reports often conclude with purported implications for clinical practice (e.g., Everly, Lating, Sherman, & Goncher, 2016; Redeker et al., 2015; Wells et al., 2014). Even ones that do not make any such claims may be cited in subsequent articles or books as providing evidence of the intervention's efficacy or clinical value. In addition, some meta-analyses of intervention research include underpowered pilot trials (e.g., Chamberlain et al., 2017; Conn, Ruppert, Chase, Enriquez, & Cooper, 2015; Davies, Spence, Vandelanotte, Caperchione, & Mummery, 2012; Samdal, Eide, Barth, Williams, & Meland, 2017). This can create additional opportunities for small pilot trials to be misinterpreted as providing credible evidence of efficacy and a rationale for translation to clinical practice.

In summary, preliminary efficacy pilot trials seldom lead to full-fledged, adequately powered RCTs, even when the findings are favorable. When they are followed by RCTs, the results are often less favorable than those of the pilot trials. Whether they stand alone in the literature or are eventually joined by larger RCTs, they tend to be misinterpreted as providing credible evi-

dence of efficacy and of clinical applicability. On balance, preliminary efficacy trials do much more harm than good for the cause of evidence-based behavioral medicine.

For these reasons, preliminary efficacy reports that are submitted to *Health Psychology* tend to get a chilly reception. It does not get any warmer if the cover letter includes a plea for special dispensation. The implicit message is typically something like:

This is a small, severely underpowered pilot study, but please take it seriously as if it were a full-fledged RCT. And please give us a break—it's only a pilot study, after all, so don't hold it to the same standards as a full-fledged RCT.

It is hard to abide these self-contradictory requests.

## The Power Analysis Problem

**Use of pilot results in power analyses.** For many years, most of us believed that it made sense to use the between-groups effect size from a randomized pilot trial in the power analysis to determine the target sample size for a larger RCT. The implicit logic was that the pilot trial's effect size provided a valid estimate of the efficacy effect size, and that the reason for conducting a larger trial was to more convincingly replicate or to confirm the results of the pilot trial. Ironically, the small size of the pilot trial tended to make us more rather than less confident in the results and in our chances of replicating the findings on a larger scale. The assumption was that if such a small trial could produce statistically significant results, then the intervention must be very effective.<sup>1</sup>

These erroneous beliefs were shattered by Kraemer, Mintz, Noda, Tinklenberg, and Yesavage's (2006) classic paper entitled "Caution Regarding the Use of Pilot Studies to Guide Power Calculations for Study Proposals." Kraemer and colleagues convincingly argued that the two most likely consequences of using a pilot study effect size in an RCT power analysis are that the RCT will never be conducted even if the true effect is clinically significant, or that the RCT will be too small to detect the true effect, even if it is clinically significant. Both eventualities greatly diminish the chances that the efficacy of the intervention will ever be definitively established.

These problems occur because small pilot trials provide very inaccurate effect size estimates, especially if the sample is heterogeneous in any way that could affect the outcomes (Piantadosi, 2017, pp. 33–35).<sup>2</sup> However, Kraemer et al. (2006) identified an even more serious problem, which is that pilot study effect sizes are the wrong ones to use even if they were much more accurate estimators of efficacy effect sizes.

<sup>1</sup> Paradoxically, this can work against grant applicants. Large pilot study effect sizes generate enthusiasm in some reviewers but lead others to conclude that efficacy has already been established and that a larger trial is unnecessary.

<sup>2</sup> Although small trials yield inaccurate effect size estimates, some of them can be useful for estimating other parameters that may affect the sample size requirements of a future trial, such as the variability of outcome measures or, for cluster-randomized trials, intraclass correlations (Eldridge, Costelloe, Kahan, Lancaster, & Kerry, 2016; Whitehead, Julious, Cooper, & Campbell, 2016).

According to Kraemer et al. (2006), the right effect size to use in the power analysis for an RCT is the threshold of clinical significance (TCS). This is the smallest effect that would disturb the existing state of clinical equipoise, that is, that would decrease the uncertainty among experts about the relative therapeutic value of each arm of the RCT. When investigators choose a TCS for a proposed trial, they must persuade other experts (i.e., grant reviewers) that the TCS was well chosen. The choice should be justifiable in relation to previous and planned research, the potential clinical impact of the intervention, and other considerations such as the costs, participant and clinician burdens, and potential harms. Severely underpowered preliminary efficacy trials provide little if any of the information needed to justify the TCS for a full-fledged RCT.

Preliminary efficacy data have traditionally served two purposes in RCT proposals: One is to support the argument that the pilot study effect would be worth finding again in a larger trial, and the other is that the larger trial probably will find it again. However, the publication of Kraemer et al. (2006) made it untenable to assume that preliminary efficacy is reason enough to conduct a larger RCT or that it tacitly guarantees that an RCT will confirm the efficacy of a pilot-tested intervention. It also confronted us with some difficult questions about our interventions and how to evaluate them. How small of an effect can an intervention produce and still have clinical value? How can applicants persuade grant reviewers, or even themselves, that a proposed RCT has a good chance of yielding informative results if they do not conduct a preliminary efficacy trial?

#### **Problematic alternatives.**

**Better estimates of the wrong effect size.** Because severely underpowered pilot trials yield inaccurate effect size estimates, some investigators assume that it is better to use meta-analytic effect sizes or the results of well-powered trials as the effect size inputs for RCT power analyses. Although these sources may provide more accurate estimates than can be derived from small pilot trials, they may still be estimates of the wrong effect.

As discussed above, the power analysis for an RCT should be based on a defensible TCS, not whatever effect happens to have been obtained in previous studies. Why conduct a trial to detect the same effect that has already been found in previous studies? This may be a legitimate goal in some circumstances (Borm, Bloem, Munneke, & Teerenstra, 2010), such as when a new trial is conducted to replicate previous findings or when an established intervention is being tested in a new population. In general, however, we may be perpetuating an unsatisfactory status quo when we base a TCS solely on the outcomes of previous trials. This approach can leave us with new or refined interventions that are no more effective than the old ones. It is at odds with the immediate goal of disturbing clinical equipoise and with the overarching scientific goal of making discernible progress (i.e., by achieving better outcomes than were possible before).

**Arbitrary choices.** Ever since Cohen defined small, medium, and large effect sizes (Cohen, 1962, 1988), these values have often been used as arbitrary inputs for RCT power analyses, despite Cohen's admonition about overreliance on them (Cohen, 1988, p. 12). This practice tempts us choose effect sizes that are large enough to justify feasible sample sizes even if the hypothesized effects are implausible, and to conduct trials that are too small to

detect treatment effects that are modest but still clinically relevant (Reardon, Smack, Herzhoff, & Tackett, 2019).

## **Moving Forward**

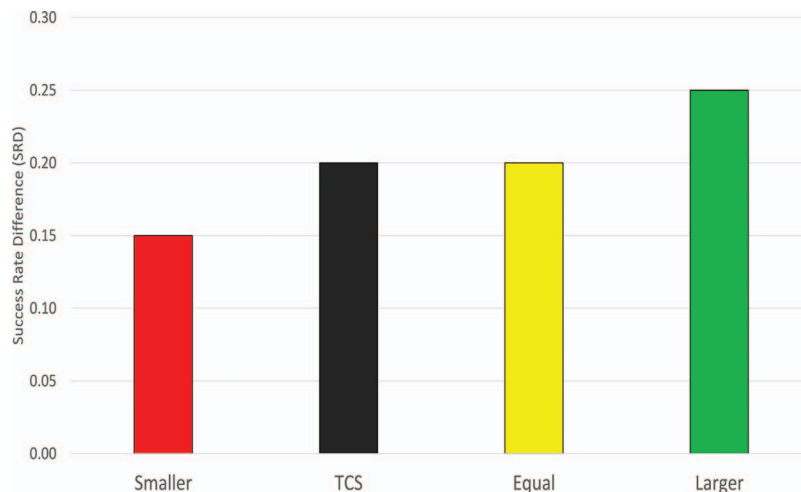
### **Two Effects Are Better Than One**

The alternative to our problematic preliminary efficacy tradition starts with the recognition that RCT proposals should address two effect sizes, not just one. The first is the TCS, that is, the effect size that should be used in the power analysis to determine the target sample size for the proposed RCT. The second is the most plausible effect size, that is, the one that seems most likely to be obtained if the RCT is conducted. The observed effect size obviously cannot be known unless and until the trial is conducted, but before the trial is conducted, a range of effect sizes may be plausible. Some may be more plausible than others or more credibly conservative.

If the reviewers of an RCT proposal have been persuaded that the TCS was well chosen, they must also be persuaded that an effect as large as or larger than the TCS is a plausible outcome of the RCT. As illustrated in Figure 1, the most plausible effect size might be smaller, larger, or equal to the TCS. If it is smaller, it will be hard to justify the trial and hard for reviewers to believe that the trial would be worth conducting. If it is at least as large as the TCS, reviewers are more likely to believe that the RCT has a reasonable chance of yielding a positive result. If the trial yields an effect as large as the TCS, it will be not only statistically significant but clinically significant as well. Furthermore, by defining a TCS, the investigator is tacitly pledging to conclude that the intervention is efficacious only if the observed effect size is as large as or larger than the TCS; that is, statistical significance is necessary but not sufficient. Thus, the "dual effect" strategy presents reviewers with a more compelling argument than "our sample will be large enough conduct a null hypothesis statistical test with adequate power." It says to them that "our target sample size is large enough to give us a reasonable chance of finding a clinically significant effect."

### **Defining the TCS**

Efforts to define clinical significance have a long history in intervention research (D. Revicki, Hays, Cella, & Sloan, 2008; D. A. Revicki et al., 2006; Engel, Beaton, & Touma, 2018; Kazdin, 1999; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999; Lemay, Tulloch, Pipe, & Reed, 2018; N. S. Jacobson, Follette, & Revenstorf, 1984; N. S. Jacobson, Roberts, Berns, & McGlinchey, 1999; N. S. Jacobson & Truax, 1991). Much of this work focuses on the clinical significance of treatment-related change within individuals (Hsu, 1999). This is an important question, but it is not the one that must be addressed when choosing the effect size for an RCT power analysis. The TCS for a clinical trial refers to the difference between the groups in the primary outcome measure, not to within-person change. The values that define clinically significant improvement within patients and clinically significant differences between groups are not interchangeable (Dworkin, 2016). Consequently, appropriate methods are needed to define clinically significant differences between groups in RCTs of behavioral interventions. Some of the most useful methods for defining TCSs for



*Figure 1.* Plausibility of a hypothesized effect size The threshold of clinical significance (TCS) is the smallest between-groups difference that would disturb clinical equipoise, that is, that would be consistent with a clinically meaningful difference between the intervention and comparator arms. The TCS bar represents the hypothesized effect size used in the power analysis for a proposed randomized controlled trial. The other bars represent a range of possible scenarios regarding whether an effect as large as the TCS is a plausible outcome of the proposed trial. The effect size is expressed on the Y-axis in success rate difference (SRD) units. Unless previous studies and/or preliminary data suggest that a difference as large or larger than the TCS is a plausible outcome, the proposed trial may be hard to justify, and reviewers may doubt its chances for success. See the online article for the color version of this figure.

clinical trials are discussed in several articles by Kraemer and colleagues (Kraemer & Kupfer, 2006; Kraemer et al., 2003, 2006; Kraemer, Neri, & Spiegel, 2020).

In these articles, Kraemer and colleagues (Kraemer & Kupfer, 2006; Kraemer et al., 2003, 2006, 2020) recommend a particular effect size index, the success rate difference (SRD), for between-groups comparisons. If a patient is randomly sampled from the treated population (T) and another patient is randomly sampled from the comparison population (C), the one who has the better clinical outcome is the called the “success.” The SRD is equivalent to the difference between the probability that the T patient is the success and the probability that the C patient is the success. If there is no between-groups difference whatsoever in the probability of success, then the SRD value is zero. If everyone in the T group has a better outcome than everyone in the C group,  $SRD = 1$ ; conversely, if everyone in the C group has a better outcome than everyone in the T group,  $SRD = -1$ . Thus, trial planners typically expect (or at least hope) to find an SRD that is somewhere between zero and +1.

The SRD is a remarkably versatile effect size index. It can be used with normally distributed, ordinal, and time-to-event outcomes, and conversion formulas are available for Cohen’s  $d$  and for hazard ratios. Its main disadvantage is that it is difficult to for researchers and clinicians and interpret. Fortunately, SRD is directly convertible into an effect size index that is easier to interpret, the number needed to treat (NNT).

In general, lower NNTs are preferable to higher ones. However, low-cost, low-risk interventions may have relatively high NNTs and still be useful. Evidence-based NNTs for a wide range of medical treatments have been posted on <https://www.thennt.com/>. It is instructive to compare the NNTs for various treatments,

especially ones that differ in terms of cost, risk, or burden. For example, low-dose aspirin for secondary cardiovascular prevention after a heart attack or stroke has an NNT of 50 (Baigent et al., 2009) and the Mediterranean diet has an NNT of 18 for prevention of recurrent heart attacks (de Lorgeril et al., 1998). In other words, for every 50 patients who take low-dose aspirin or for every 18 patients who adhere to the Mediterranean diet, approximately one patient will be spared from having a heart attack that would have occurred without the aspirin or the diet. Thus, relatively few patients will have better cardiovascular outcomes with than without these treatments, but since they are relatively inexpensive and safe, they still play important roles in secondary prevention.

The NNT for a relatively expensive, intensive, time-consuming, or otherwise burdensome behavioral intervention would probably have to be much lower than 50 to convince patients and other stakeholders that it is worth providing as part of clinical care. Fortunately, some of the evidence-based behavioral interventions that we currently have to offer do have much lower NNTs than that of low-dose aspirin. For example, a recent meta-analysis of 64 Internet-delivered cognitive behavior therapy (CBT) trials for major depression, panic disorder, social anxiety disorder, or generalized anxiety disorder yielded an NNT of 2.34 (Andrews et al., 2018).

As these examples illustrate, the clinical significance of between-groups differences in RCTs can be evaluated in terms of how much more likely patients are to benefit from the treatment than from the comparison condition, and what it takes to produce this benefit in terms of time, money, or other sorts of costs and risks. If other treatments for the same problem or disorder are available, the clinical significance of a novel treatment can also be judged by comparing the NNTs of the novel and existing treat-



ments, assuming that they have been tested against similar comparators.

### Establishing Plausibility

**Perceived plausibility.** When an investigator hypothesizes that the difference between groups will be equal to or greater than the TCS, the next question is whether an effect as large as that is plausible. In other words, does an effect that large seem likely to be obtained, assuming that the trial is well designed and rigorously executed? Investigators usually believe that the effect they are hypothesizing is plausible; this belief undergirds their decision to propose an RCT. However, the plausibility of the hypothesized effect is in the eye of the beholder, and the key beholders of RCT proposals are its reviewers. The applicant must persuade them that an effect as large as the TCS is a plausible outcome of the proposed RCT because if the observed effect turns out to be smaller than that, the trial is unlikely to have much of an impact. Thus, the reviewers' perception of the potential impact of the proposed trial depends on the plausibility of the hypothesized effect.

Equipoise has been defined in a variety of ways, but the classic definition asserts that clinical equipoise exists if there is substantial uncertainty within the expert community about the comparative values of the alternatives that will be tested in an RCT (Freedman, 1987). If the experts who review an RCT grant proposal are quite certain about how a trial would turn out if it were conducted, they will probably not give it a favorable score. This is true whether they strongly believe that there will be no significant difference between the groups or that the intervention will turn out to be very clearly superior to the comparator. It is also true if the objective of the study is to attempt to replicate established effects. If the outcome seems to be a foregone conclusion, the reviewers will probably think that the proposed trial is not worth conducting.

Thus, the hypothesized effect size can seem utterly implausible, entirely too plausible, or somewhere in between. Reviewers who are in perfect internal equipoise about a trial think that its chance of success (i.e., the chance that the trial will yield an effect as large as the TCS) is about 50/50. However, complete uncertainty does not engender much enthusiasm among reviewers. Those who guess that the chance of success is higher than 50/50 will be persuaded that an effect as large as the TCS is at least plausible. Those who guess that the chance is quite a bit higher than 50/50 may be more enthusiastic about the proposed trial. However, those who think that the chance of success is too high may conclude that the trial is unnecessary. **Therefore, to make a strong argument for the plausibility of the hypothesized effect size, applicants must persuade reviewers that the chance of success is better than 50/50 but nowhere near 100%.**

What does it take to move reviewers away from complete uncertainty or even skepticism about a predicted or hypothesized effect and toward the sense that it is a plausible or very plausible result?<sup>3</sup> **Preliminary studies can provide valuable evidence, but persuasive arguments also depend on information about the scientific credibility of the intervention and on what might be called "comparables," to borrow a term from real estate.**

**Scientific credibility.** Scientists who review extramural research grants for the National Institutes of Health or other major funding agencies tend to be very skeptical about interventions that

they think are premised on pseudoscience, vague theories, baseless commercial claims, or other dubious grounds. Conversely, they often see promise in interventions that are built on a solid foundation of behavioral and social science. Imagine, for example, that a review panel receives two RCT grant applications that are very similar and that propose the same TCS but that aim to test different interventions for the same problem. One would test an intervention that is based on well-established findings from behavioral economics research (e.g., Asch et al., 2015). The other would test thought field therapy (Callahan, 2001), an intervention that many experts consider to be a fringe psychotherapeutic practice (Lilienfeld, Lynn, & Lohr, 2015; Lilienfeld, Ritschel, Lynn, Cautin, & Latzman, 2014). All else being equal, most reviewers would think that the first intervention is more credible than the second and that it is more likely to be efficacious. Thus, interventions that are rooted in T1 translational research (Fort, Herr, Shaw, Gutzman, & Starren, 2017) tend to be viewed as more scientifically credible than other interventions.

Evidence that an intervention has undergone a systematic, early phase, development, refinement, and evaluation process can also enhance its scientific credibility. Several recently developed frameworks can be used to make early phase intervention research more systematic and productive. For example, the multiphase optimization strategy can be used to screen out ineffective components of complex behavioral interventions (Collins, Murphy, Nair, & Strecher, 2005); the NIH Science of Behavior Change framework can be used to identify and measure underlying mechanisms that can be targeted to change health behaviors (Nielsen et al., 2018); and a framework developed by Voils et al. (2014) can be used to optimize the dose (i.e., frequency, amount, and/or duration) of a behavioral intervention. Novel interventions that have been systematically developed and refined within these types of frameworks tend to be viewed as more scientifically credible than ones that emerge fully formed from personal or clinical experiences, untested theories of health behavior, popular culture, or traditional healing practices.

Many informative trials test interventions that are not novel or innovative, such as when the efficacy of a well-established intervention is evaluated in a different population than the one(s) in which it has typically been used. For example, many behavioral interventions for medically ill patients are borrowed from research conducted in psychiatric patient populations (e.g., Freedland, Carney, Rich, Steinmeyer, & Rubin, 2015). Evidence of systematic development and refinement may be less important when testing well-established interventions than when testing novel ones. Nevertheless, intervention development and dose-finding frameworks such as the ones discussed above can be used to adapt established interventions for use in new populations. For instance, medically ill patients may respond better to a shorter or more flexible version of an intervention than the original one that healthier individuals are able to tolerate. Thus, there are circumstances in which the credibility of a well-established intervention can be enhanced

<sup>3</sup> Ioannidis (2005) presents an analysis of factors that affect the pretrial probability that the hypothesized effect is true. The relationship of this probability to reviewer perceptions of plausibility is not well understood, but relevant research undoubtedly helps to shape these perceptions.

through the kinds of early phase studies that are usually associated with novel or innovative interventions.

**Comparables.** Progress in behavioral intervention research tends to be incremental; dramatic breakthroughs are rare. Consequently, reviewers are likely to view a predicted effect as implausible if it is dramatically larger than the ones that have been found in previous studies. However, if the proposal is to test a novel intervention or an adaptation of an established one, there may not have been any rigorous trials of the same intervention for the same problem in the same patient population. In this situation, reviewers might compare the predicted effect size to the results of trials of other treatments for the same problem, or of similar treatments for other problems or populations. This is analogous to the real estate practice of comparing the sale price of a house to comparables, that is, the market values of similar properties in the same neighborhood.

Given the manifold demands of the grant review process, reviewers may have to rely on their existing fund of knowledge about relevant studies or meta-analyses rather than conducting a literature review of their own if the applicant does not do it for them. Thus, applicants should consider presenting some comparables from the literature to help to build a case for the plausibility of the effect they are hoping to obtain in their proposed RCT.

**Preliminary data.** If preliminary efficacy trials are ill-advised, can other kinds of preliminary studies be used to support the plausibility of a predicted effect? One possible solution is to conduct a small randomized trial but without conducting a statistical test of the between-groups difference and without claiming that the results show preliminary evidence of efficacy. The confidence interval around the between-groups difference will probably be very wide and it may be highly skew. If the TCS is somewhere within this interval, the predicted effect might seem plausible, but it might not. As an example, consider a research team that has defined the TCS for a proposed trial as  $SRD = .20$ . Their small preliminary trial produces an effect size of  $SRD = .25$ , which suggests that an effect as large as the TCS would be a plausible outcome of the larger trial they hope to conduct. However, the confidence interval around the preliminary study effect ranges from  $.10$  to  $.75$ . Some reviewers might see the  $.25$  as encouraging, but others might see the  $.10$  as a reason to doubt the plausibility of the predicted effect. Thus, investigators who adopt this strategy are gambling that the reviewers will find the central tendency more persuasive than the lower confidence limit. This could be a risky strategy.

The NIH-funded Obesity-Related Behavioral Intervention Trials (ORBIT) Consortium recently developed a translational research model for developing and testing behavioral interventions for chronic diseases (Czajkowski et al., 2015). The ORBIT model recommends “proof-of-concept” studies as an efficient, cost-effective way to determine whether an intervention merits more rigorous and costly testing in an adequately powered RCT. Proof-of-concept studies do not focus on preliminary efficacy (i.e., between-groups effects). Instead, they evaluate whether the intervention can produce clinically significant improvement within individuals.

One way to do this is via a small, uncontrolled or “open label” trial of the intervention. In some areas of research, another approach would be to conduct a series of single-subject ( $N = 1$ ) studies (Kazdin, 2011; Ridenour & Tueller, 2019; Shaffer,

Kronish, Falzon, Cheung, & Davidson, 2018). Either way, the investigator first defines a clinically significant level of improvement (e.g., 50% decrease in a depression scale score relative to baseline) or a clinically significant target value (e.g., 80% adherence to a medication), and then determines whether treated participants reach this level or target. If many or most of the participants reach the desired level or target, this provides some encouraging support for a Phase II trial of the intervention. It is more encouraging if this degree of improvement is uncommon in clinical practice or in historical controls.

Another way is to use exploratory data from trials of mechanistic targets or of surrogate or intermediate outcomes that are on the hypothesized path to the primary outcome of an anticipated trial. For example, intentions to engage in physical activity are considered to be a modifiable mechanistic target in exercise intervention research (Rhodes & Dickau, 2012; Rhodes & Rebar, 2017). A research team might start by conducting a Phase I or early Phase II trial of physical activity promotion program that is designed to strengthen exercise intentions. The primary outcome would be a measure of intentions, and the amount of activity might be an exploratory outcome. This trial would have adequate power for the primary outcome but not for the exploratory outcome. Nevertheless, the exploratory data could be used to determine whether increases in intentions are followed by increases in exercise, and if so, to assess the magnitude of the change in exercise behavior. Favorable findings could be used in the RCT proposal as preliminary evidence that the intention-focused intervention yields enough change in physical activity to justify a larger trial with exercise as the primary outcome and intentions as an intermediate (secondary) outcome. This type of preliminary trial is an RCT in its own right, not a pilot study, and it should be large enough to test its primary hypothesis with adequate power (Whitehead, Sully, & Campbell, 2014).

## Establishing Feasibility

**Feasibility studies.** If a trial is not feasible to conduct, it cannot yield informative results. Thus, the plausibility of a clinically significant finding is predicated on the trial’s feasibility. What does it take to persuade reviewers that a proposed trial is feasible? For that matter, what does it take for investigators to persuade themselves that a trial they would like to propose is going to be feasible?

In the faulty logic of the preliminary efficacy tradition, the completion of a small, underpowered preliminary efficacy trial suggests ipso facto that a much larger trial is feasible. This is like claiming that running a mile proves that one is ready to run a marathon. A more valid way to assess the feasibility of a proposed trial is to conduct a feasibility study.

A feasibility study may not be necessary if a research team’s recent projects have already shown that their next project is feasible, or if the anticipated RCT is an inexpensive, low-risk study such as one in which volunteers will be recruited from an undergraduate subject pool for a brief intervention. In contrast, if the anticipated RCT will entail larger investments of time and funding, if its feasibility is uncertain, or if preliminary data are needed to inform the choice of an outcome measure or other design decisions (Blatch-Jones, Pek, Kirkpatrick, & Ashton-Key, 2018; Lancaster,

Dodd, & Williamson, 2004), then a feasibility study may be necessary.

Until recently, there was no consensus about what feasibility studies should be designed to accomplish or what they should be called (Arain, Campbell, Cooper, & Lancaster, 2010). That began to change when an expert panel formed to develop a feasibility study extension to the Consolidated Standards of Report Trials (CONSORT; Eldridge, Chan, et al., 2016). The panel produced consensus definitions and a checklist (Thabane et al., 2016). Based on Delphi surveys and consensus meetings, they agreed that “. . . all pilot studies are feasibility studies but that some feasibility studies are not pilot studies” (Eldridge, Lancaster, et al., 2016, p. 11). They also defined three types of feasibility studies: (a) *Randomized pilot studies* are ones in which an anticipated RCT is conducted on a smaller scale to determine whether it can be done. (b) *Nonrandomized pilot studies* are feasibility studies that do not involve randomization of participants, but that do expose participants to the intervention. Some nonrandomized pilot studies include comparator arms but others do not. (c) *Feasibility studies that are not pilot studies* investigate certain aspects of RCT feasibility but do not implement the intervention. A data-mining study of a hospital’s electronic medical records to determine the size and characteristics of the pool of potential participants in an anticipated trial is an example of a feasibility study that is not a pilot study.

Thus, a randomized pilot study is a feasibility study, but one with a distinctive feature: It has the same design as the anticipated RCT, including random assignment of participants to the same intervention and comparator arms. There may be peripheral differences between the pilot and main trial such as different lengths of follow-up, but the randomized pilot study is essentially a smaller version of the anticipated trial with respect to core design features.

The designs of randomized pilot studies and preliminary efficacy trials are indistinguishable. Unlike preliminary efficacy trials, however, randomized pilot studies as defined in the CONSORT extension do not include statistical tests of efficacy hypotheses, and they do address specific feasibility questions such as whether it will be possible to recruit enough participants for the anticipated RCT or whether protocol adherence will be adequate. These questions are addressed by setting feasibility criteria and evaluating whether they can be met.

This is possible only if the investigators have a formative plan for the efficacy trial. In particular, they must have chosen its design and comparator, and they should have a rough estimate of its target sample size. This is necessary because randomized pilot studies should answer concrete, practical questions about the feasibility of a specific, anticipated RCT. For example, if the investigators have a rough estimate of the target sample size that will be needed for the efficacy trial, and if they know the approximate duration its enrollment phase, they can determine the enrollment rate (e.g., the number of enrollments per month) that they will probably have to achieve. This rate becomes the enrollment feasibility criterion in the pilot study. The pilot study will probably have a much shorter enrollment phase than the anticipated trial. Nevertheless, if the enrollment rate in the pilot study meets or exceeds the enrollment rate that will be required for the anticipated trial, this suggests that the anticipated trial is feasible, at least in terms of enrollment. Thus, it is the achievement of the enrollment

rate criterion that suggests feasibility, not the total sample size of the pilot study or the mere fact that a pilot study was conducted.

**The fallibility of feasibility.** Evidence of feasibility does not guarantee that the anticipated trial will unfold as planned; unexpected challenges and setbacks are commonplace in clinical trials. It is not unusual, for example, for recruitment to be more difficult in the second half of a trial than in the first half. Thus, the achievement of an enrollment rate criterion in a brief pilot study can lead to overoptimistic expectations about the feasibility of an anticipated trial. Although the achieved enrollment rate is a descriptive statistic, a confidence interval around the rate can be calculated to give the reviewers and the investigators themselves a sense of how much worse (or better) the enrollment rate might be in the anticipated trial than in the pilot study. A recent review found only modest bias in external pilot trials as predictors of RCT randomization rates and attrition, but the confidence intervals were very wide (Cooper, Whitehead, Pottrill, Julious, & Walters, 2018).

Confidence intervals are inferential statistics, but inferences based on feasibility data are unlike ones based on other kinds of studies. In most other studies, data are collected on a sample of participants drawn from a larger population, to support generalizable inferences about the whole population. In contrast, feasibility studies are conducted to support inferences about the researchers themselves and about the environment in which their anticipated trial will be conducted. They address questions such as whether the research team will be able to recruit enough patients from the hospitals or clinics where they plan to enroll participants for their trial. Feasibility studies seldom support inferences about other research teams working with other patients in other environments. It would be unusual to conclude, for example, that because we are able to recruit enough participants at our center for our study, other researchers working at other centers will therefore be able to recruit enough participants at their centers for their studies.

There is nothing wrong with that, except that it means that the results of some feasibility studies may not be of much interest to anyone other than the reviewers of the RCT proposal and the investigators themselves. This can make it difficult to publish the results.<sup>4</sup> A feasibility study may be more publishable if it includes some elements that would be of interest to a broader audience, such as data bearing on how to overcome common recruitment or retention challenges. However, it should not include a severely underpowered statistical test of the efficacy of the intervention.

**Feasibility studies versus other preliminary studies.** If a feasibility study is needed, it is usually the *last* preliminary study before the RCT is proposed. If other kinds of preliminary studies such as dose-finding or proof-of-concept studies are needed, they usually precede the feasibility study. In fact, some funding opportunities discourage investigators from interposing additional studies between a feasibility study and submission of an RCT proposal. For example, the National Heart, Lung, and Blood Institute’s Clinical Trial Pilot Studies announcement (PAR-18-463) states that completion of the pilot study’s specific aims should provide results that are both necessary and sufficient to make a final decision about the subsequent trial.

<sup>4</sup> The difficulty of publishing typical feasibility studies in journals such as *Health Psychology* is one of the reasons why the *Pilot and Feasibility Studies* journal was launched in 2015.



Many investigators are reluctant to limit the aims of their final preliminary study to questions about feasibility, and there are disincentives to do so. First, as noted above, it can be difficult to publish reports that focus solely on practical questions about the feasibility of a trial. Second, some reviewers of exploratory/developmental (R21) or planning (R34) grant applications may not be very enthusiastic about proposals that are limited to feasibility aims. However, difficult questions can arise about the compatibility of the feasibility aims and the other aims of these applications.

For example, an applicant might propose to conduct a study that would serve both as a proof-of-concept test of the intervention and as a feasibility pilot study. Some reviewers may view the efficiency of this combination as a strength, but others may see it as a weakness. They would see it that way if they believe that it would be premature to investigate the feasibility of an RCT before the investigator has established that the intervention is promising enough to justify testing its efficacy in an RCT. However, this objection may be less likely to arise if the investigator has already produced some promising proof-of-concept data and plans to use the proposed feasibility pilot study as an opportunity to collect some additional proof-of-concept data.

Also, some of the “feasibility” or “feasibility pilot” study reports that are submitted for publication are not about feasibility as defined by the CONSORT extension. In other words, they do not examine the feasibility of a specific, anticipated (i.e., planned or proposed) RCT. Instead, they investigate questions about the feasibility of the intervention itself (e.g., Barry et al., 2019). These studies tend to address questions such as whether therapists or counselors can deliver the intervention as intended and whether the patients find it to be acceptable and are willing to complete the program. These are certainly important questions, and the feasibility of a future RCT may very well depend on affirmative answers to them. However, if a study focuses primarily on questions about the intervention per se rather than on questions about the feasibility of an anticipated trial, it would be better to frame it as a Phase I intervention development or refinement study rather than as a feasibility or feasibility pilot study.

### Advantages and Disadvantages

The traditional approach to pilot testing offered a simple and appealing solution for preliminary trial data. It ostensibly provided “preliminary evidence of efficacy,” an effect size to plug into our RCT power analysis, evidence that our proposed RCT is feasible, and potentially publishable results, all in one convenient study. But as has been said before, “. . . there is always an easy solution to every human problem—neat, plausible, and wrong” (Mencken, 1949, p. 443).

The alternative recommended in this report is more complicated. It requires us to consider some difficult questions about the aims, probative value, limitations, and publishability of our preliminary studies; about information that must be gleaned from sources other than small pilot trials; and about what is at stake when we propose RCTs of health-related behavioral interventions. In some cases, the alternative approach may entail a series of different kinds of preliminary studies with different purposes and designs. In other cases, it may not require any “preliminary” studies at all, such as when one full-fledged trial sets the stage for the next full-fledged

trial. Thus, unlike the traditional approach, the alternative is not a simple, “one size fits all” solution.

However, the alternative also has some clear advantages. Investigators can tailor this approach to the unique needs of their research program and use it to answer a variety of questions that tend to come up when behavioral RCT proposals are reviewed. It can help to eliminate the problems that preliminary efficacy trials create, and it encourages researchers to take full advantage of the translational models and optimization frameworks that have been developed to improve behavioral intervention research. It can also help to increase our adherence to evolving methodological standards for pilot studies and for RCTs. In short, we have much more to gain than to lose from ending our traditional reliance on preliminary efficacy trials and adopting more contemporary translational research models and intervention optimization frameworks.

### Recommendations

*Clinical inertia* is a well-known phenomenon in medical practice; it occurs when physicians recognize treatable problems in their patients but fail to initiate or intensify appropriate treatments (Phillips et al., 2001). Many of us have experienced what might be called *methodological inertia* regarding preliminary research on behavioral interventions. Since the publication of Kraemer et al.’s classic cautionary paper (Kraemer et al., 2006), it has been clear that something was fundamentally wrong with our traditional ideas about pilot studies, effect sizes, and power analyses for efficacy trials. Since Thabane, Eldridge, Leon, and others started providing expert guidance on feasibility and pilot studies (e.g., Eldridge, Chan, et al., 2016; Leon, Davis, & Kraemer, 2011; Thabane et al., 2010), we have known what these studies should look like and what their aims should or should not be. And since the emergence of translational models, optimization frameworks, and dose-finding methods for behavioral intervention research (e.g., Collins, Murphy, & Strecher, 2007; Czajkowski et al., 2015; Michie, van Stralen, & West, 2011; Nielsen et al., 2018; Onken, Carroll, Shoham, Cuthbert, & Riddle, 2014; Voils et al., 2014), it has been clear that a variety of different kinds of studies can help to lay the groundwork for efficacy trials, yet many of us still see traditional pilot trials as the be-all and end-all of preliminary intervention research.

Methodological inertia means that we are slow and reluctant to change longstanding beliefs and research practices; it does not mean that we are incapable of doing so. We can change our thinking, at least eventually, when new empirical findings challenge established facts. We can do the same when methodological advances challenge established research paradigms. The following recommendations are offered in that spirit.

### Recommendations for Researchers

- Do not conduct preliminary efficacy trials that are severely underpowered by design.
- Use suitable translational research models and/or intervention optimization frameworks to guide the preliminary study or studies that lead up to an RCT proposal.
- Contextualize the aims and foreseeable limitations of preliminary studies within these models when you discuss them in RCT proposals or report them in articles.



- Reserve terms such as “randomized pilot study” or “randomized pilot trial” for studies that meet the CONSORT definition of randomized pilot and feasibility trials.
- Use terms that are consistent with applicable translational models or optimization frameworks for other kinds of preliminary work (e.g., proof-of-concept studies).
- Carefully define and justify a TCS for an anticipated RCT and use it in the power analysis for the RCT proposal.
- Cite and/or generate evidence to support the scientific credibility of the intervention, the plausibility of finding an effect as large as the TCS, and the feasibility of the RCT.
- Do not propose or conduct RCTs that are severely underpowered by design; this includes preliminary trials (e.g., trials with mechanistic or intermediate outcomes) that are designed to pave the way for more definitive trials with clinically important outcomes.

### Recommendations for Grant Reviewers

- Do not expect or encourage applicants to conduct preliminary efficacy trials or to use preliminary efficacy data as effect size inputs for power analyses in RCT proposals.
- Do not look to preliminary efficacy data for tacit assurance that a proposed RCT will confirm the efficacy of the intervention.
- Consider instead whether the application provides persuasive reasons to move from skepticism or complete uncertainty about the feasibility and potential value of a proposed RCT toward a judgment that the RCT seems both feasible and worth conducting.
- Expect applicants to define and justify a TCS or the smallest between-groups difference that would be worth finding, given the objectives of the research, and to include this value in the power analysis and statistical analysis plan.
- Do not expect this value to be based primarily (if at all) on a randomized pilot trial.
- Recognize that recent developments in translational research models, optimization frameworks, and reporting guidelines have opened the door to a variety of different kinds of preliminary research on behavioral interventions, and that they are replacing some of the ways that researchers have traditionally laid the groundwork for their RCT proposals.

### Recommendations for Article Reviewers and Journal Editors

- Take grossly inadequate sample sizes and other methodological weaknesses seriously, even when authors ask you to overlook them by labeling their reports as being based on “pilot studies” or “preliminary research.”
- Do not ask or allow the authors of bona fide randomized pilot trials, as defined by the CONSORT extension, to report severely underpowered tests of efficacy hypotheses.
- Be duly critical of preliminary efficacy trials.
- Judge preliminary studies of behavioral interventions in terms of whether they help to lay the groundwork for an

anticipated, full-fledged, efficacy or effectiveness trial that (at least provisionally) seems to be worth conducting, not in terms of the potential clinical impact or public health significance of the preliminary work itself.

### Summary and Conclusion

Although preliminary studies of behavioral interventions may seem easier to design, conduct, and review than full-fledged RCTs, they are not necessarily any easier. Confusion, uncertainty, and misguided thinking about preliminary studies are common problems in our field, and the preliminary efficacy tradition has held us back for many years. Fortunately, recent methodological developments have started to move our field in a better direction. Adoption of the recommendations provided in this report can help to accelerate this positive trend.

### References

- Ames, S. C., Werch, C. E., Ames, G. E., Lange, L. J., Schroeder, D. R., Hanson, A. C., & Patten, C. A. (2010). Integrated smoking cessation and binge drinking intervention for young adults: A pilot investigation. *Annals of Behavioral Medicine, 40*, 343–349. <http://dx.doi.org/10.1007/s12160-010-9222-4>
- Andrews, G., Basu, A., Cuijpers, P., Craske, M. G., McEvoy, P., English, C. L., & Newby, J. M. (2018). Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: An updated meta-analysis. *Journal of Anxiety Disorders, 55*, 70–78. <http://dx.doi.org/10.1016/j.janxdis.2018.01.001>
- Arain, M., Campbell, M. J., Cooper, C. L., & Lancaster, G. A. (2010). What is a pilot or feasibility study? A review of current practice and editorial policy. *BMC Medical Research Methodology, 10*, 67. <http://dx.doi.org/10.1186/1471-2288-10-67>
- Asch, D. A., Troxel, A. B., Stewart, W. F., Sequist, T. D., Jones, J. B., Hirsch, A. G., . . . Volpp, K. G. (2015). Effect of financial incentives to physicians, patients, or both on lipid levels. *Journal of the American Medical Association, 314*, 1926–1935. <http://dx.doi.org/10.1001/jama.2015.14850>
- Baigent, C., Blackwell, L., Collins, R., Emberson, J., Godwin, J., Peto, R., . . . the Antithrombotic Trialists’ (ATT) Collaboration. (2009). Aspirin in the primary and secondary prevention of vascular disease: Collaborative meta-analysis of individual participant data from randomised trials. *Lancet, 373*, 1849–1860. [http://dx.doi.org/10.1016/S0140-6736\(09\)60503-1](http://dx.doi.org/10.1016/S0140-6736(09)60503-1)
- Barry, D. T., Beitel, M., Cutter, C. J., Fiellin, D. A., Kerns, R. D., Moore, B. A., . . . Schottenfeld, R. S. (2019). An evaluation of the feasibility, acceptability, and preliminary efficacy of cognitive-behavioral therapy for opioid use disorder and chronic pain. *Drug and Alcohol Dependence, 194*, 460–467. <http://dx.doi.org/10.1016/j.drugalcdep.2018.10.015>
- Blatch-Jones, A. J., Pek, W., Kirkpatrick, E., & Ashton-Key, M. (2018). Role of feasibility and pilot studies in randomised controlled trials: A cross-sectional study. *British Medical Journal Open, 8*, e022233. <http://dx.doi.org/10.1136/bmjopen-2018-022233>
- Borm, G. F., Bloem, B. R., Munneke, M., & Teerenstra, S. (2010). A simple method for calculating power based on a prior trial. *Journal of Clinical Epidemiology, 63*, 992–997. <http://dx.doi.org/10.1016/j.jclinepi.2009.10.011>
- Callahan, R. J. (2001). The impact of thought field therapy on heart rate variability. *Journal of Clinical Psychology, 57*, 1153–1170. <http://dx.doi.org/10.1002/jclp.1082>
- Carpenter, M. J., & Gray, K. M. (2010). A pilot randomized study of smokeless tobacco use among smokers not interested in quitting: Changes in smoking behavior and readiness to quit. *Nicotine & Tobacco Research, 12*, 136–143. <http://dx.doi.org/10.1093/ntr/ntp186>

- Chamberlain, C., O'Mara-Eves, A., Porter, J., Coleman, T., Perlen, S. M., Thomas, J., & McKenzie, J. E. (2017). Psychosocial interventions for supporting women to stop smoking in pregnancy. *Cochrane Database of Systematic Reviews*, 2, CD001055. <http://dx.doi.org/10.1002/14651858.CD001055.pub5>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65, 145–153. <http://dx.doi.org/10.1037/h0045186>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Collins, L. M., Murphy, S. A., Nair, V. N., & Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine*, 30, 65–73. [http://dx.doi.org/10.1207/s15324796abm3001\\_8](http://dx.doi.org/10.1207/s15324796abm3001_8)
- Collins, L. M., Murphy, S. A., & Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): New methods for more potent eHealth interventions. *American Journal of Preventive Medicine*, 32 (5, Suppl.), S112–S118. <http://dx.doi.org/10.1016/j.amepre.2007.01.022>
- Conn, V. S., Ruppert, T. M., Chase, J. A., Enriquez, M., & Cooper, P. S. (2015). Interventions to improve medication adherence in hypertensive patients: systematic review and meta-analysis. *Current Hypertension Reports*, 17, 94. <http://dx.doi.org/10.1007/s11906-015-0606-5>
- Cooper, C. L., Whitehead, A., Pottrill, E., Julious, S. A., & Walters, S. J. (2018). Are pilot trials useful for predicting randomisation and attrition rates in definitive studies: A review of publicly funded trials. *Clinical Trials*, 15, 189–196. <http://dx.doi.org/10.1177/1740774517752113>
- Czajkowski, S. M., Powell, L. H., Adler, N., Naar-King, S., Reynolds, K. D., Hunter, C. M., . . . Charlson, M. E. (2015). From ideas to efficacy: The ORBIT model for developing behavioral treatments for chronic diseases. *Health Psychology*, 34, 971–982. <http://dx.doi.org/10.1037/hea0000161>
- Davies, C. A., Spence, J. C., Vandelandotte, C., Caperchione, C. M., & Mummery, W. K. (2012). Meta-analysis of internet-delivered interventions to increase physical activity levels. *The International Journal of Behavioral Nutrition and Physical Activity*, 9, 52. <http://dx.doi.org/10.1186/1479-5868-9-52>
- de Lorgeril, M., Salen, P., Martin, J. L., Monjaud, I., Boucher, P., & Marmelle, N. (1998). Mediterranean dietary pattern in a randomized trial: Prolonged survival and possible reduced cancer rate. *Archives of Internal Medicine*, 158, 1181–1187. <http://dx.doi.org/10.1001/archinte.158.11.1181>
- Desai, B., Desai, V., Shah, S., Srinath, A., Saleh, A., Simunovic, N., . . . Bhandari, M. (2018). Pilot randomized controlled trials in the orthopaedic surgery literature: A systematic review. *BMC Musculoskeletal Disorders*, 19, 412. <http://dx.doi.org/10.1186/s12891-018-2337-7>
- Dworkin, R. H. (2016). Two very different types of clinical importance. *Contemporary Clinical Trials*, 46, 11. <http://dx.doi.org/10.1016/j.cct.2015.11.007>
- Ebbert, J. O., Edmonds, A., Luo, X., Jensen, J., & Hatsukami, D. K. (2010). Smokeless tobacco reduction with the nicotine lozenge and behavioral intervention. *Nicotine & Tobacco Research*, 12, 823–827. <http://dx.doi.org/10.1093/ntr/ntq088>
- Eldridge, S. M., Chan, C. L., Campbell, M. J., Bond, C. M., Hopewell, S., Thabane, L., . . . the PAFS consensus group. (2016). CONSORT 2010 statement: Extension to randomised pilot and feasibility trials. *Pilot and Feasibility Studies*, 2, 64. <http://dx.doi.org/10.1186/s40814-016-0105-8>
- Eldridge, S. M., Costelloe, C. E., Kahan, B. C., Lancaster, G. A., & Kerry, S. M. (2016). How big should the pilot study for my cluster randomised trial be? *Statistical Methods in Medical Research*, 25, 1039–1056. <http://dx.doi.org/10.1177/0962282015588242>
- Eldridge, S. M., Lancaster, G. A., Campbell, M. J., Thabane, L., Hopewell, S., Coleman, C. L., & Bond, C. M. (2016). Defining feasibility and pilot studies in preparation for randomised controlled trials: Development of a conceptual framework. *PLoS ONE*, 11, e0150205. <http://dx.doi.org/10.1371/journal.pone.0150205>
- Engel, L., Beaton, D. E., & Touma, Z. (2018). Minimal clinically important difference. *Rheumatic Diseases Clinics of North America*, 44, 177–188. <http://dx.doi.org/10.1016/j.rdc.2018.01.011>
- Everly, G. S., Jr., Lating, J. M., Sherman, M. F., & Goncher, I. (2016). The potential efficacy of psychological first aid on self-reported anxiety and mood. *Journal of Nervous and Mental Disease*, 204, 233–235. <http://dx.doi.org/10.1097/NMD.0000000000000429>
- Fort, D. G., Herr, T. M., Shaw, P. L., Gutzman, K. E., & Starren, J. B. (2017). Mapping the evolving definitions of translational research. *Journal of Clinical and Translational Science*, 1, 60–66. <http://dx.doi.org/10.1017/cts.2016.10>
- Freedland, K. E., Carney, R. M., Rich, M. W., Steinhilber, B. C., & Rubin, E. H. (2015). Cognitive behavior therapy for depression and self-care in heart failure patients: A randomized clinical trial. *Journal of the American Medical Association Internal Medicine*, 175, 1773–1782. <http://dx.doi.org/10.1001/jamainternmed.2015.5220>
- Freedman, B. (1987). Equipoise and the ethics of clinical research. *The New England Journal of Medicine*, 317, 141–145. <http://dx.doi.org/10.1056/NEJM198707163170304>
- Hayes, S., Uszynski, M. K., Motl, R. W., Gallagher, S., Larkin, A., Newell, J., . . . Coote, S. (2017). Randomised controlled pilot trial of an exercise plus behaviour change intervention in people with multiple sclerosis: The Step it Up study. *British Medical Journal Open*, 7(10), e016336. <http://dx.doi.org/10.1136/bmjopen-2017-016336>
- Hennrikus, D., Pirie, P., Hellerstedt, W., Lando, H. A., Steele, J., & Dunn, C. (2010). Increasing support for smoking cessation during pregnancy and postpartum: Results of a randomized controlled pilot study. *Preventive Medicine: An International Journal Devoted to Practice and Theory*, 50, 134–137. <http://dx.doi.org/10.1016/j.ypmed.2010.01.003>
- Hsu, L. M. (1999). A comparison of three methods of identifying reliable and clinically significant client changes: Commentary on Hageman and Arrindell. *Behaviour Research and Therapy*, 37, 1195–1202.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2, e124. <http://dx.doi.org/10.1371/journal.pmed.0020124>
- Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, 15, 336–352. [http://dx.doi.org/10.1016/S0005-7894\(84\)80002-7](http://dx.doi.org/10.1016/S0005-7894(84)80002-7)
- Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, 67, 300–307. <http://dx.doi.org/10.1037/0022-006X.67.3.300>
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12–19. <http://dx.doi.org/10.1037/0022-006X.59.1.12>
- Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology*, 67, 332–339. <http://dx.doi.org/10.1037/0022-006X.67.3.332>
- Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.
- Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, 67, 285–299. <http://dx.doi.org/10.1037/0022-006X.67.3.285>
- Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, 59, 990–996. <http://dx.doi.org/10.1016/j.biopsych.2005.09.014>
- Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power

- calculations for study proposals. *Archives of General Psychiatry*, 63, 484–489. <http://dx.doi.org/10.1001/archpsyc.63.5.484>
- Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child & Adolescent Psychiatry*, 42, 1524–1529. <http://dx.doi.org/10.1097/00004583-200312000-00022>
- Kraemer, H. C., Neri, E., & Spiegel, D. (2020). Wrangling with p-values versus effect sizes to improve medical decision-making: A tutorial. *International Journal of Eating Disorders*, 53, 302–308. <http://dx.doi.org/10.1002/eat.23216>
- Lancaster, G. A., Dodd, S., & Williamson, P. R. (2004). Design and analysis of pilot studies: Recommendations for good practice. *Journal of Evaluation in Clinical Practice*, 10, 307–312. <http://dx.doi.org/10.1111/j.2002.384.doc.x>
- Lemay, K. R., Tulloch, H. E., Pipe, A. L., & Reed, J. L. (2018). Establishing the minimal clinically important difference for the Hospital Anxiety and Depression Scale in patients with cardiovascular disease. *Journal of Cardiopulmonary Rehabilitation and Prevention*, 39, E6–E11
- Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The role and interpretation of pilot studies in clinical research. *Journal of Psychiatric Research*, 45, 626–629. <http://dx.doi.org/10.1016/j.jpsychires.2010.10.008>
- Lilienfeld, S. O., Lynn, S. J., & Lohr, J. M. (2015). *Science and pseudo-science in clinical psychology* (2nd ed.). New York, NY: Guilford Press.
- Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Lutzman, R. D. (2014). Why ineffective psychotherapies appear to work. *Perspectives on Psychological Science*, 9, 355–387. <http://dx.doi.org/10.1177/1745691614535216>
- Logan, K. J., Woodside, J. V., Young, I. S., McKinley, M. C., Perkins-Porras, L., & McKeown, P. P. (2010). Adoption and maintenance of a Mediterranean diet in patients with coronary heart disease from a Northern European population: A pilot randomised trial of different methods of delivering Mediterranean diet advice. *Journal of Human Nutrition and Dietetics*, 23, 30–37. <http://dx.doi.org/10.1111/j.1365-277X.2009.00989.x>
- McDonald, D., O'Brien, J., Farr, E., & Haaga, D. A. (2010). Pilot study of inducing smoking cessation attempts by activating a sense of looming vulnerability. *Addictive Behaviors*, 35, 599–606. <http://dx.doi.org/10.1016/j.addbeh.2010.02.008>
- McEvoy, C. T., Moore, S. E., Appleton, K. M., Cupples, M. E., Erwin, C. M., Hunter, S. J., . . . Woodside, J. V. (2018). Trial to encourage adoption and maintenance of a Mediterranean diet (TEAM-MED): Protocol for a randomised feasibility trial of a peer support intervention for dietary behaviour change in adults at high cardiovascular disease risk. *International Journal of Environmental Research and Public Health*, 15, 1130. <http://dx.doi.org/10.3390/ijerph15061130>
- McEvoy, C. T., Moore, S. E., Appleton, K. M., Cupples, M. E., Erwin, C., Kee, F., . . . Woodside, J. V. (2018). Development of a peer support intervention to encourage dietary behaviour change towards a Mediterranean diet in adults at high cardiovascular risk. *BMC Public Health*, 18, 1194. <http://dx.doi.org/10.1186/s12889-018-6108-z>
- Mencken, H. L. (1949). *A Mencken chrestomathy* (1st ed.). New York, NY: A. A. Knopf.
- Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science*, 6, 42. <http://dx.doi.org/10.1186/1748-5908-6-42>
- Motl, R. W., Dlugonski, D., Wójcicki, T. R., McAuley, E., & Mohr, D. C. (2011). Internet intervention for increasing physical activity in persons with multiple sclerosis. *Multiple Sclerosis*, 17, 116–128. <http://dx.doi.org/10.1177/1352458510383148>
- Motl, R. W., Hubbard, E. A., Bollaert, R. E., Adamson, B. C., Kinnett-Hopkins, D., Balto, J. M., . . . McAuley, E. (2017). Randomized controlled trial of an e-learning designed behavioral intervention for increasing physical activity behavior in multiple sclerosis. *Multiple Sclerosis Journal—Experimental, Translational and Clinical*. Advance online publication. <http://dx.doi.org/10.1177/2055217317734886>
- Nielsen, L., Riddle, M., King, J. W., Aklin, W. M., Chen, W., Clark, D., . . . the NIH Science of Behavior Change Implementation Team. (2018). The NIH Science of Behavior Change Program: Transforming the science through a focus on mechanisms of change. *Behaviour Research and Therapy*, 101, 3–11. <http://dx.doi.org/10.1016/j.brat.2017.07.002>
- Onken, L. S., Carroll, K. M., Shoham, V., Cuthbert, B. N., & Riddle, M. (2014). Reenvisioning clinical science: Unifying the discipline to improve the public health. *Clinical Psychological Science*, 2, 22–34. <http://dx.doi.org/10.1177/2167702613497932>
- Patten, C. A., Koller, K. R., Flanagan, C. A., Hiratsuka, V. Y., Hughes, C. A., Wolfe, A. W., . . . Thomas, T. K. (2019). Biomarker feedback intervention for smoking cessation among Alaska Native pregnant women: Randomized pilot study. *Patient Education and Counseling*, 102, 528–535. <http://dx.doi.org/10.1016/j.pec.2018.10.009>
- Patten, C. A., Lando, H. A., Desnoyers, C. A., Barrows, Y., Klejka, J., Decker, P. A., . . . Burhansstipanov, L. (2019). The Healthy Pregnancies Project: Study protocol and baseline characteristics for a cluster-randomized controlled trial of a community intervention to reduce tobacco use among Alaska Native pregnant women. *Contemporary Clinical Trials*, 78, 116–125. <http://dx.doi.org/10.1016/j.cct.2019.01.012>
- Patten, C. A., Windsor, R. A., Renner, C. C., Enoch, C., Hochreiter, A., Nevak, C., . . . Brockman, T. (2010). Feasibility of a tobacco cessation intervention for pregnant Alaska Native women. *Nicotine & Tobacco Research*, 12, 79–87. <http://dx.doi.org/10.1093/ntr/ntp180>
- Phillips, L. S., Branch, W. T., Jr., Cook, C. B., Doyle, J. P., El-Kebbi, I. M., Gallina, D. L., . . . Barnes, C. S. (2001). Clinical inertia. *Annals of Internal Medicine*, 135, 825–834. <http://dx.doi.org/10.7326/0003-4819-135-9-200111060-00012>
- Piantadosi, S. (2017). *Clinical trials: A methodologic perspective* (3rd ed.). Hoboken, NJ: Wiley.
- Reardon, K. W., Smack, A. J., Herzhoff, K., & Tackett, J. L. (2019). An N-pact factor for clinical psychological research. *Journal of Abnormal Psychology*, 128, 493–499. <http://dx.doi.org/10.1037/abn0000435>
- Redeker, N. S., Jeon, S., Andrews, L., Cline, J., Jacoby, D., & Mohsenin, V. (2015). Feasibility and efficacy of a self-management intervention for insomnia in stable heart failure. *Journal of Clinical Sleep Medicine*, 11, 1109–1119. <http://dx.doi.org/10.5664/jcs.m.5082>
- Revicki, D. A., Cella, D., Hays, R. D., Sloan, J. A., Lenderking, W. R., & Aaronson, N. K. (2006). Responsiveness and minimal important differences for patient reported outcomes. *Health and Quality of Life Outcomes*, 4, 70. <http://dx.doi.org/10.1186/1477-7525-4-70>
- Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, 61, 102–109. <http://dx.doi.org/10.1016/j.jclinepi.2007.03.012>
- Rhodes, R. E., & Dickau, L. (2012). Experimental evidence for the intention-behavior relationship in the physical activity domain: A meta-analysis. *Health Psychology*, 31, 724–727. <http://dx.doi.org/10.1037/a0027290>
- Rhodes, R. E., & Rebar, A. L. (2017). Conceptualizing and defining the intention construct for future physical activity research. *Exercise and Sport Sciences Reviews*, 45, 209–216. <http://dx.doi.org/10.1249/JES.0000000000000127>
- Ridenour, T., & Tueller, S. (2019). Recent advances and illustrations of modern within-subjects clinical trials methods. *Annals of Behavioral Medicine*, 53, S172. <http://dx.doi.org/10.1093/abm/kaz007>
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638–641. <http://dx.doi.org/10.1037/0033-2909.86.3.638>



- Ruppar, T. M. (2010). Randomized pilot study of a behavioral feedback intervention to improve medication adherence in older adults with hypertension. *Journal of Cardiovascular Nursing, 25*, 470–479. <http://dx.doi.org/10.1097/JCN.0b013e3181d5f9c5>
- Russell, C., Conn, V., Ashbaugh, C., Madsen, R., Wakefield, M., Webb, A., . . . Peace, L. (2011). Taking immunosuppressive medications effectively (TIMELink): A pilot randomized controlled trial in adult kidney transplant recipients. *Clinical Transplantation, 25*, 864–870. <http://dx.doi.org/10.1111/j.1399-0012.2010.01358.x>
- Russell, C. L., Moore, S., Hathaway, D., Cheng, A. L., Chen, G., & Goggin, K. (2016). MAGIC Study: Aims, design and methods using SystemCHANGE to improve immunosuppressive medication adherence in adult kidney transplant recipients. *BMC Nephrology, 17*, 84. <http://dx.doi.org/10.1186/s12882-016-0285-8>
- Samdal, G. B., Eide, G. E., Barth, T., Williams, G., & Meland, E. (2017). Effective behaviour change techniques for physical activity and healthy eating in overweight and obese adults; systematic review and meta-regression analyses. *The International Journal of Behavioral Nutrition and Physical Activity, 14*, 42. <http://dx.doi.org/10.1186/s12966-017-0494-y>
- Shaffer, J. A., Kronish, I. M., Falzon, L., Cheung, Y. K., & Davidson, K. W. (2018). N-of-1 randomized intervention trials in health psychology: A systematic review and methodology critique. *Annals of Behavioral Medicine, 52*, 731–742. <http://dx.doi.org/10.1093/abm/kax026>
- Shanyinde, M., Pickering, R. M., & Weatherall, M. (2011). Questions asked and answered in pilot and feasibility randomized controlled trials. *BMC Medical Research Methodology, 11*, 117. <http://dx.doi.org/10.1186/1471-2288-11-117>
- Thabane, L., Hopewell, S., Lancaster, G. A., Bond, C. M., Coleman, C. L., Campbell, M. J., & Eldridge, S. M. (2016). Methods and processes for development of a CONSORT extension for reporting pilot randomized controlled trials. *Pilot and Feasibility Studies, 2*, 25. <http://dx.doi.org/10.1186/s40814-016-0065-z>
- Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., . . . Goldsmith, C. H. (2010). A tutorial on pilot studies: The what, why and how. *BMC Medical Research Methodology, 10*, 1. <http://dx.doi.org/10.1186/1471-2288-10-1>
- Vaduganathan, M., Greene, S. J., Ambrosy, A. P., Gheorghiu, M., & Butler, J. (2013). The disconnect between Phase II and Phase III trials of drugs for heart failure. *Nature Reviews Cardiology, 10*, 85–97. <http://dx.doi.org/10.1038/nrcardio.2012.181>
- Voils, C. I., King, H. A., Maciejewski, M. L., Allen, K. D., Yancy, W. S., Jr., & Shaffer, J. A. (2014). Approaches for informing optimal dose of behavioral interventions. *Annals of Behavioral Medicine, 48*, 392–401. <http://dx.doi.org/10.1007/s12160-014-9618-7>
- Wells, R. E., Burch, R., Paulsen, R. H., Wayne, P. M., Houle, T. T., & Loder, E. (2014). Meditation for migraines: A pilot randomized controlled trial. *Headache: The Journal of Head and Face Pain, 54*, 1484–1495. <http://dx.doi.org/10.1111/head.12420>
- Whitehead, A. L., Julious, S. A., Cooper, C. L., & Campbell, M. J. (2016). Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical Methods in Medical Research, 25*, 1057–1073. <http://dx.doi.org/10.1177/0962280215588241>
- Whitehead, A. L., Sully, B. G., & Campbell, M. J. (2014). Pilot and feasibility studies: Is there a difference from each other and from a randomised controlled trial? *Contemporary Clinical Trials, 38*, 130–133. <http://dx.doi.org/10.1016/j.cct.2014.04.001>
- Williams, D. M., Ussher, M., Dunsiger, S., Miranda, R., Jr., Gwaltney, C. J., Monti, P. M., & Emerson, J. (2014). Overcoming limitations in previous research on exercise as a smoking cessation treatment: Rationale and design of the “Quit for Health” trial. *Contemporary Clinical Trials, 37*, 33–42. <http://dx.doi.org/10.1016/j.cct.2013.11.005>
- Williams, D. M., Whiteley, J. A., Dunsiger, S., Jennings, E. G., Albrecht, A. E., Ussher, M. H., . . . Marcus, B. H. (2010). Moderate intensity exercise as an adjunct to standard smoking cessation treatment for women: A pilot study. *Psychology of Addictive Behaviors, 24*, 349–354. <http://dx.doi.org/10.1037/a0018332>

Received December 11, 2019

Revision received March 20, 2020

Accepted April 30, 2020 ■

### E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <https://my.apa.org/portal/alerts/> and you will be notified by e-mail when issues of interest to you become available!