Article

American Journal of Evaluation
2017, Vol. 38(2) 246-261
© The Author(s) 2016
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1098214016651489
journals.sagepub.com/home/aje
SAGE

The Nonuse, Misuse, and Proper Use of Pilot Studies in Experimental Evaluation Research

Erik Westlund[1] and Elizabeth A. Stuart[2,3,4]

Abstract

This article discusses the nonuse, misuse, and proper use of pilot studies in experimental evaluation research. The authors first show that there is little theoretical, practical, or empirical guidance available to researchers who seek to incorporate pilot studies into experimental evaluation research designs. The authors then discuss how pilot studies can be misused, using statistical simulations to illustrate the error that can result from using effect sizes from pilot studies to decide whether to conduct a full trial or using effect sizes from pilot studies as the basis of power calculations in a full trial. Informed by the review of the literature and these simulation results, the authors conclude by proposing practical suggestions to researchers and practitioners on how to properly use pilot studies in experimental evaluation research.

Keywords

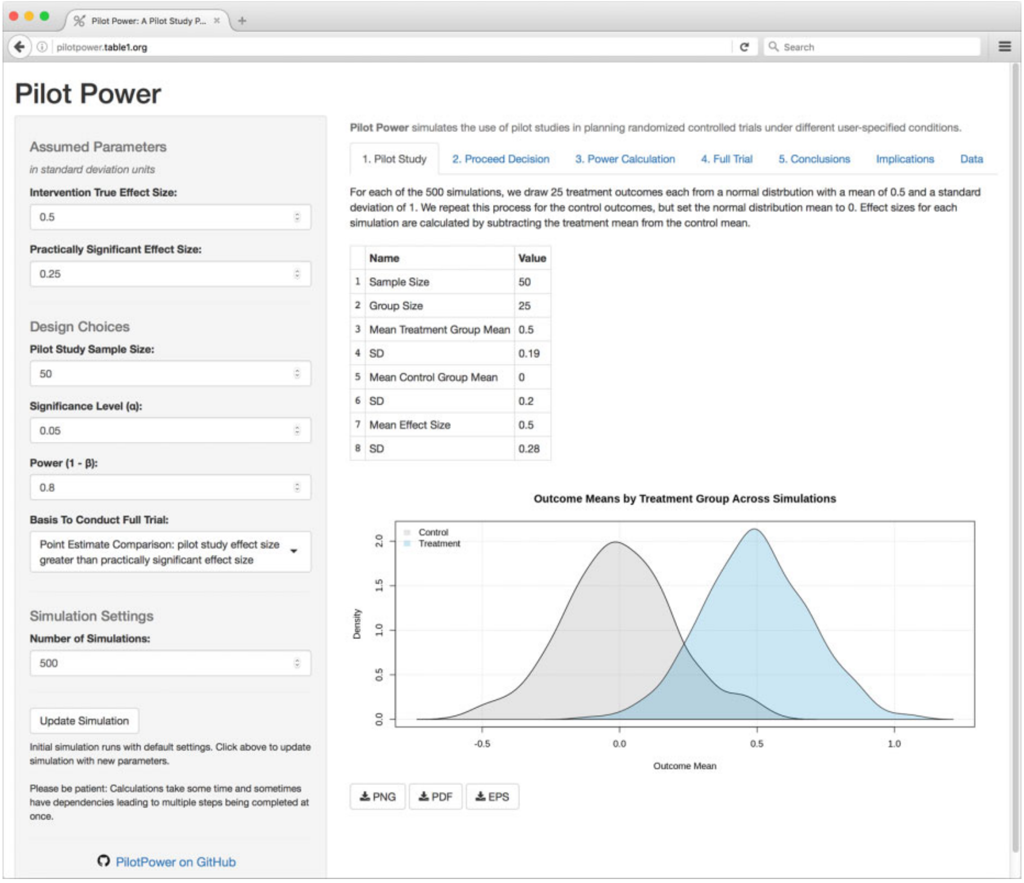
evaluation practice, RCT, research design, evaluation theory

When designing and conducting experimental research, pilot studies—small-scale studies conducted before a full trial—are considered a best practice. In evaluation research, where experiments are the gold standard of research design, pilot studies should be an integral part of the research design process. Yet, there is little explicit guidance in the literature on evaluation or research design to guide researchers who seek to incorporate pilot studies into experimental evaluation research.

This article has three aims. First, we document the dearth of guidance available to researchers conducting pilot studies as part of experimental research. Using program evaluation in education research as a case study, we also show that although pilot studies are required by grant protocols of major education research grant-making bodies, methods for their appropriate use are rarely

[1] Department of Sociology, University of Iowa, Iowa City, IA, USA
[2] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[3] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[4] Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Corresponding Author:
Erik Westlund, Department of Sociology, University of Iowa, 140 Seashore Hall West, Iowa City, IA 52242, USA.
Email: erik-westlund@uiowa.edu



Pilot Power

Assumed Parameters

in standard deviation units

Intervention True Effect Size: 0.5

Practically Significant Effect Size: 0.25

Design Choices

Pilot Study Sample Size: 50

Significance Level (α): 0.05

Power (1 - β): 0.8

Basis To Conduct Full Trial: Point Estimate Comparison: pilot study effect size greater than practically significant effect size

Simulation Settings

Number of Simulations: 500

Initial simulation runs with default settings. Click above to update simulation with new parameters.

Please be patient: Calculations take some time and sometimes have dependencies leading to multiple steps being completed at once.

Pilot Power simulates the use of pilot studies in planning randomized controlled trials under different user-specified conditions.

For each of the 500 simulations, we draw 25 treatment outcomes each from a normal distrbution with a mean of 0.5 and a standard deviation of 1. We repeat this process for the control outcomes, but set the normal distribution mean to 0. Effect sizes for each simulation are calculated by subtracting the treatment mean from the control mean.

	Name	Value
1	Sample Size	50
2	Group Size	25
3	Mean Treatment Group Mean	0.5
4	SD	0.19
5	Mean Control Group Mean	0
6	SD	0.2
7	Mean Effect Size	0.5
8	SD	0.28

Figure 1. "Pilot Power" web application.

discussed or provided. As such, we show how researchers seeking guidance on how to use pilot studies in experimental evaluation research will find little guidance from either theory or practice.

Second, we discuss how pilot studies can be misused in experimental evaluation research. Using the *shiny* framework for the R programming language, we built an application called "Pilot Power" (see Figure 1). We used this application to conduct simulations of pilot studies in experimental research designs to gather insight into how the uncertainty inherent to small studies can result in their misuse (R Core Team, 2015; RStudio, Inc., 2013; Westlund & Stuart, 2016). We focus on two particular errors: using pilot studies to decide whether to conduct a full trial of an intervention and using pilot studies to determine how many cases to sample in such a full trial. We argue that researchers conducting randomized experiments should never use pilot study effect sizes alone to determine whether a full trial should be conducted. We also show how uncertainty inherent in small studies will systematically lead to underpowered full trials, if researchers used pilot study effect sizes as the basis for their power calculations. We contend that full trials should always be powered to detect a predetermined level of "practical" or "clinical" significance rather than based on the effect size observed in a small pilot study.

Third, we draw on the literature, our experience in reviewing grant protocols, and the information gleaned from these simulations to offer practical advice on how to responsibly use pilot studies in experimental research. We argue that pilot study results should be used as one data point among many, possibly including data from past observational research and solid theoretical models, to determine whether a full trial is worth conducting. We also argue the value of pilot studies for assessing feasibility of the intervention or study design. We discuss how, as a complement to pilot study results, researchers can use the results of observational studies to inform important design choices. Finally, we suggest that in some cases adaptive study designs can function as an alternative to pilot studies. We conclude with a discussion of our findings and recommendations for practice.

The Lack of Use of Pilot Studies in Experimental Evaluation Research

Little guidance is available to researchers on how to responsibly and effectively use pilot studies in experimental evaluation research. Below we first review the evaluation methods literature, showing that neither classic texts on research design and program evaluation nor peer-reviewed journal articles provide substantial discussion on how to use pilot studies in evaluation research. We use education research as a case study to show that grant-review protocols that require pilot studies to be incorporated into the research designs of funded studies offer little guidance on exactly how to conduct or use pilot study results. We also review the published literature in four major journals that publish evaluation research, showing that very few published studies of experimental evaluations report using pilot studies as part of their research design. This demonstrates the lack of available exemplars of how to effectively and responsibly use pilot studies in experimental evaluation research.

Guidance in the Evaluation Methods Literature

Standard authorities on research design and program evaluation fail to provide specific guidance on the use of pilot studies. For example, two major textbooks on program evaluation do not have an entry for pilot studies in their topic indexes (Rossi, Lipsey, & Freeman, 2004; Shadish, Cook, & Leviton, 1991). Likewise, a foremost textbook on research design for studies attempting to make causal inferences has no index reference for pilot studies, does not define them in its glossary, and does not discuss them thoroughly in the text (Shadish, Cook, & Campbell, 2002).

Scholarly journals offer some insight absent from textbooks, but very little of this insight is directed at social science researchers or program evaluators. We searched JSTOR and Google Scholar for articles on pilot studies using the following search terms: “pilot stud*,” “pilot stud*” AND “experiments,” “pilot stud*” AND “experimental research,” “pilot stud*” AND “evaluation,” “pilot stud*” AND “evaluation research,” “pilot stud*” AND “program evaluation,” “pilot stud*” AND “social science research,” “pilot stud*” AND “social science,” “pilot stud*” AND “education research,” and “pilot stud*” AND “education.” This search returned a number of studies, but theses studies were limited to fields such as psychology and medicine. None of the results offered practical advice directed specifically at experimental evaluation researchers, although a number of existing studies provide relevant insights.

Several studies provide discussions of the basics of conducting pilot studies in different contexts, such as nursing (Hertzog, 2008) or clinical medical research (Lancaster, Dodd, & Williamson, 2004; Thabane et al., 2010). A number of articles discuss statistical techniques, offering general overviews (Schoenfeld, 1980), guidance on specific practices such as scale development (Johanson & Brooks, 2010), or on specific statistical techniques such as survival analysis (Liu, Dahlberg, & Crowley, 1993). Several articles discuss the use of internal pilot studies and adaptive designs as alternatives to standard pilot studies; these approaches may provide cost-effective ways to adjust sample sizes and study procedures in light of information learned in the earlier stages of a clinical trial (Baldi, Gouchon, Di Giulio, Buja, & Gregori, 2011; Wittes & Brittain, 1990; Zucker, Wittes, Schabenberger, & Brittain, 1999). Others suggest rules for stopping trials early in light of evidence of a treatment’s danger, inefficacy, or efficacy (Whitehead, 2004). A number of studies show how pilot studies, when used for power analysis, produce unreliable effect size estimates. For example, Vickers (2003) discusses the importance of carefully considering the standard deviation (*SD*) used in power calculations. He analyzed 98 experiments with results in major medical journals and found that the *SD* used to conduct power analyses was typically smaller than that found in full trials, resulting in underpowered full trials. Kraemer, Mintz, Noda, Tinklenberg, and Yesavage (2006) ran simulations of pilot studies to demonstrate that making design choices based upon pilot study effect sizes often has two unfortunate results: Worthwhile studies are prematurely aborted, and studies that are conducted are often underpowered.

Education Research as a Case Study

Assessing how pilot studies are used in experimental evaluation research in education provides further insight into the problems faced by evaluation researchers seeking insight on how to use pilot studies. Below, we document this in two ways. First, we review the grant protocols of the one of the primary funding bodies of experimental research in education, the Institute of Education Sciences (IES). We then review the published literature in three major journals publishing peer-reviewed educational research, showing that there are very few exemplars available for researchers to follow when incorporating pilot studies into the designs of experimental evaluation research.

Pilot studies in grant writing: An example. Education research is one of the largest fields in which experiments are routinely used as tools of program evaluation. Within this field, the IES—the official research wing of the U.S. Department of Education—is the primary institutional proponent and funder of experimental research. We use the IES as an example to illustrate the types of research design guidelines and protocols that researchers may encounter when seeking funding for and conducting experimental research.

The IES grant protocols consider pilot studies in two ways. First, they require researchers who are seeking funding for the development of educational interventions to incorporate pilot studies into their proposed research programs. Second, they encourage researchers writing grant applications for studies to evaluate existing interventions to use pilot studies as evidence of an intervention's promise (U.S. Department of Education, 2016).

Grants for the development of educational interventions. Besides requiring that a pilot study be conducted and that this pilot study be conducted in a true-to-life educational environment, the protocols for those seeking grants to develop interventions offer no set standards or guidelines for designing a pilot study. The guidelines set no limits on the number of cases required in the pilot study, noting that pilot studies may be fully powered trials or they may be underpowered studies "that provide unbiased effect size estimates of practical consequence which can stand as evidence of promise while not statistically significant" (U.S. Department of Education, 2016, p. 54). There are also no specific guidelines for determining whether the conducted pilot study provides evidence that the intervention being developed is promising. The protocols do, however, require grant seekers to articulate how they plan to use evidence from a pilot study to assess such promise.

Grants for studying already existing interventions. The IES encourages grant writers seeking funding to study existing interventions to provide pilot study results as evidence of the promise of the intervention to be studied. However, as above, no specific guidelines are provided to assess whether the cited pilot studies do indeed provide evidence of the intervention's promise. Again, grant writers must use their own discretion.

In each of the above cases, it is left to the researcher to choose an appropriate way to design and interpret pilot study results. We do not object to this per se. However, we emphasize that it is understandable that many researchers may be unsure exactly how to design or interpret pilot study results in light of the lack of guidelines for doing so that exists in the literature, as we documented in the prior section. For this reason, it is unsurprising, as we show below, that few published evaluations report the use of pilot studies.

Table 1. Summary of the Reported Use of Pilot Studies in Empirical Education Research Evaluation Published in Four Peer-Reviewed Journals Since 2000–2015.

	American Educational Research Journal (JREE)	*Education Evaluation and Policy Analysis (EEPA)*	*Journal of Research on Educational Effectiveness*[a]	*Studies in Educational Evaluation*	Total
Total empirical evaluation studies referencing pilot studies	0	0	3	0	3
Purpose of Pilot Study		Number of Studies Meeting Below Criteria			
Assess feasibility	0	0	0	0	0
Intervention development	0	0	2	0	2
Instrument/measurement development and validation	0	0	1	0	1
Sample/site selection	0	0	0	0	0
Develop preliminary findings	0	0	0	0	0

[a] The *JREE*'s first publication was 2008, and so has a shorter search window than *EEPA* or the *American Journal of Evaluation*.

Pilot studies in published empirical education research. Because the methodological literature on program evaluation in the social sciences is largely quiet on how to use pilot studies, researchers might seek guidance on how to use pilot studies when designing experimental research by reviewing past published work. As we show below, however, very few published studies of empirical research in peer-reviewed journals document pilot studies as part of their research design.

To demonstrate this, we searched the archives of four prominent journals that routinely or solely publish educational evaluation research: *American Educational Research Journal*, *Education Evaluation and Policy Analysis*, *Journal of Research on Educational Effectiveness* (*JREE*), and *Studies in Educational Evaluation*. We searched the full text for the phrase "pilot stud*" on all articles published since the year 2000. We then reviewed all hits and excluded all studies that do not report the results of an evaluation of a program or educational intervention. We excluded nonempirical articles (e.g., theoretical articles, literature reviews, documentations of instrument development). This process resulted in only three articles, all of which were in the *JREE*. For each of these articles, we determined the purpose of the pilot study as it pertains to the particular evaluation. The results of this review are presented in Table 1.

All three studies published in *JREE* used experimental designs. Of these three studies, two used pilot studies to help improve the intervention being studied. The third used pilot studies to validate measurement instruments. None of these three articles went into detail on the mechanics of the pilot study, nor did they provide extensive comment on how the pilot studies informed their final trial design choices.

This review shows how researchers who seek guidance from the published literature on how to use pilot studies to inform research design choices in experimental research will find few insights from the published empirical literature. Although it would be inappropriate to conclude that pilot studies are rarely being conducted because there will certainly be a publication bias where studies with unpromising pilot studies are not reported on in the academic literature, we can determine that searching the published literature for exemplars on how to use pilot studies in empirical research leads to few examples.

The Potential Misuse of Pilot Studies: Evidence from Simulations

Having discussed the relative nonuse of pilot studies in experimental research, it is useful to discuss the ways in which they might be misused. In this section, we focus on two potential misuses: (1) using pilot study results *in isolation* to determine whether to conduct a full trial and (2) using effect size estimates from pilot studies for power calculations in full trials. We use mathematical simulations to illustrate how each of the above practices can result in researchers making poor research design choices.

Simulation Scenarios

To assess how pilot studies can be misused when deciding whether to conduct a full trial or when conducting power calculations for a full trial, we consider four specific scenarios that experimental evaluation researchers might encounter. All effect sizes are in *SD* units. We base the effect sizes in our simulations on conventional criteria of "small," "medium," and "large" effect sizes as established by Cohen (1988). Following the What Works Clearinghouse (2011), we define "practical significance"—the effect size at which an intervention has clinical or practical importance—as 0.25 *SD*s.[1] In practice, of course, what constitutes a practically important effect size is context dependent.[2]

> **Scenario 1.** *No effect.* An ineffective intervention where the true effect size of the program is 0 *SD*s.
>
> **Scenario 2.** *Small effect without practical significance*. In this scenario, we define the true effect size of the intervention to be 0.20 *SD*s, conventionally considered a small effect. This scenario is meant to capture situations where an intervention is moderately effective and might be of theoretical importance, but the magnitude of the effect size is too small to be considered of practical significance and justify widespread deployment.
>
> **Scenario 3.** *Medium effect with practical significance.* In this scenario, we define the true effect size of the intervention to be 0.50 *SD*s, conventionally considered a medium-sized effect. In education research, this scenario is meant to capture situations where an intervention is both effective and widely considered of substantive practical importance (What Works Clearinghouse, 2011).
>
> **Scenario 4.** *Very effective intervention with practical significance.* In this scenario, we define the true effect size of the intervention to be 0.80 *SD*s, conventionally considered a large-sized effect. In virtually all real-world situations, this would be considered an effect size of considerable practical importance.

To run our simulations, we developed an open source, publicly available web application using the R programming language and the *shiny* web framework (R Core Team, 2015; RStudio, Inc., 2013; Westlund & Stuart, 2016).[3] The logic of this program follows that laid out by Kraemer et al. (2006) in their articulation of the problem with using pilot studies in psychological research. Our program allows researchers to set various assumptions about the true effect size of an intervention. It also allows them to choose decision rules for interpreting pilot study results.

Using Pilot Studies to Determine Whether to Conduct a Full Trial

One potential and alluring use of pilot studies is to determine whether an intervention has sufficient promise of effectiveness to justify conducting a full trial. A sufficiently large effect size observed in a pilot study could be seen as providing researchers with a quantifiable data point to justify proceeding to a full trial. The prospect of using statistically nonsignificant, but unbiased estimates as

Table 2. Percentage of Simulations Where a Full Trial is Conducted Given Different Decision Rules to Proceed to a Full Trial.

	True Effect Size			
Proceed Rule	Scenario 1 0.00	Scenario 2 0.20	Scenario 3 0.50	Scenario 4 0.80
If effect size point estimate is greater than 0	50	76	96	100
If effect size point estimate is greater than practically significant effect size	19	43	81	97
If treatment mean not equal to control mean (two-tailed *t*-test)	5	11	41	79

guidance for whether to proceed with a larger trial is an interesting one, but one that we argue is potentially dangerous due to the inherent uncertainty built into underpowered trials. There are two dangers:

1. Overestimating the effectiveness of an intervention and investing resources in a costly trial of an ineffective intervention.
2. Underestimating the effectiveness of an intervention and failing to conduct a full trial of an actually effective intervention.

Possible decision rules for determining whether to conduct a full trial. If we were to use pilot trial results to inform whether or not we should conduct a full trial, we are faced with two decisions. First, researchers must decide what effect size will count as evidence of intervention effectiveness—a level of practical or clinical significance. This level will be dependent upon the nature of the research (as noted, we set this value to 0.25 *SD*s in our simulations). Second, researchers must decide what statistical test to apply, if any at all. Plausible approaches to making this decision include:

1. Proceed if pilot effect size point estimate is greater than a predetermined practically significant effect size.
2. Proceed if the pilot study effect size is significantly different from 0 at some prespecified significant level (e.g., $\alpha = .05$).

For each of the four scenarios, we simulate 100,000 pilot studies by drawing 100,000 samples of 50 cases. In each draw, 25 control cases are drawn from a normal distribution where the mean is set to 0 (i.e., control cases) and 25 treatment cases are drawn from a normal distribution where the mean is set to the scenario's predetermined effect size (i.e., treatment cases). Effect sizes are calculated by subtracting the observed treatment mean from the observed control mean.

We consider a correct decision to proceed to trial whenever the true effect size is greater than or equal to the practically significant effect size of 0.25 *SD* units. Table 2 shows how often across the 100,000 pilot study simulations researchers would choose to conduct a full trial given certain true effect sizes and different decision rules.

In Scenarios 1 and 2, where the true effect size is smaller than the practically significant effect size of 0.25, we consider it a mistake to proceed to a full trial. If we conducted a full trial only when the pilot study treatment mean was significantly different from the control group mean using a two-tailed *t*-test, we would mistakenly proceed to trial in 5% of cases in Scenario 1 (this follows from the chosen significance level, α, of .05) and 11% of cases in Scenario 2. If we chose to proceed to trial when the pilot study effect size point estimate was greater than practical significance, we would mistakenly proceed to a full trial in 19% of cases in Scenario 1 and 43% of cases in Scenario 2.

In Scenarios 3 and 4, the true effect size is greater than or equal to the practically significant effect size of 0.25. In these scenarios, we consider it a mistake *not* to proceed to a full trial. If we chose to proceed to trial only when the observed pilot study effect size point estimate is greater than practical significance, we would conduct full trials in 81% of cases in Scenario 3 and 97% of cases in Scenario 4; thus making an incorrect decision 19% and 3% of the time for Scenarios 3 and 4, respectively. Using a two-tailed t-test to compare treatment and control groups and proceeding to a full trial only when the resulting p value is less than .05, we would proceed to trial in Scenario 3 only 41% of the time and in Scenario 4 only 79% of the time. Consequently, in Scenarios 3 and 4 we would prematurely abandon a full trial of an effective intervention 59% and 21% of the time, respectively.

Conclusions regarding use of pilot studies to determine when to conduct a full trial. It is clear that it is improper to apply null hypothesis significance testing to pilot studies in determining whether or not to continue with a full trial. The typical sample size of a pilot study (in our simulations, 50 cases) is often too small to detect true effects when they exist, resulting in premature abandonment of trials. However, even when using unbiased point estimates to make decisions, we still make incorrect decisions at a relatively high rate even when applying the laxer rule of conducting a full trial only when the pilot study effect size is greater than the predetermined level of practical significance. For this reason, we contend that using only a point estimate to base a decision is in most cases a mistake. We discuss this further below.

Using Pilot Studies for the Basis of Power Calculations

If we decide to conduct a full trial, we must decide on an effect size on which to base power calculations used to determine the sample size of the full trial. An alluring prospect is to use the effect size observed in a pilot study to proceed. If, for example, we observe a very large effect size, in the resource-constrained environment in which many researchers operate we may consider it justifiable to sample fewer subjects than we would in the absence of an unbiased estimate of the intervention's effect size (and, in fact, the pilot studies with large effect size estimates are those that are most likely to proceed to a full trial, as discussed above). However, the danger is that those large effect sizes are likely overestimates of the true effects, as detailed below, which means that we are likely to, in fact, enroll too few subjects in the full trial to detect an actual existing effect size (i.e., commit a Type 2 error).

Choosing a sample size. Using the same simulation setup as above, we conduct full trial sample size power calculations in two ways. First, we calculate the sample size needed to detect the observed pilot study effect size. For example, if we observed an effect size of 0.40, we would sample 200 cases in our full trial: the required number of cases to detect a statistically significant effect size of 0.40 using a two-tailed t-test with a significance level (α) of .05, 80% of the time (i.e., power or $1 - \beta = .80$). Second, we calculate the sample size required to detect the practically significant effect size. In each case, this value is 506, or the number of cases required to detect the practically significant effect size of 0.25 under the same assumption ($\alpha = .05$, $1 - \beta = .80$).

Table 3 provides a summary of these calculations for each of the four scenarios. Scenario 3, where the true effect size is 0.50, the mean sample size needed to detect the observed pilot study effect size is 137. The minimum sample size calculated is 12, corresponding to an observed pilot study effect size of 1.20 (0.70 larger than the true effect size). In Scenario 4, the mean sample size needed to detect the observed pilot study effect size is 73. The minimum sample size calculated is 10, corresponding to an observed pilot study effect size of 1.32. Of course, both of these correspond

Table 3. Characteristics of Sample Size Power Calculations to Detect Observed Pilot Study Effect Size, When the Decision is Made to Proceed to Trial.[a]

	True Effect Size			
Descriptive statistics of sample size needed to detect observed pilot study effect size across all simulations	Scenario 1 0.00	Scenario 2 0.20	Scenario 3 0.50	Scenario 4 0.80
Mean	247	208	137	73
Minimum	24	18	12	10
Maximum	506	506	506	504
Standard deviation	123	122	105	67
Sample size required to detect practically significant effect size	506	506	506	506
Difference between sample size required to detect practically significant effect size and average sample size required to detect observed practically significant effect size (cases "saved")	259	298	369	433

[a]These figures are based upon proceeding to a full trial, whenever the observed pilot study is larger than the practically significant effect size.

to full trials that would be severely underpowered to detect the true effect, a result of the pilot study in fact providing an overly optimistic effect size estimate.

The maximum sample size yielded by the power calculations in both scenarios is 506, which corresponds to the practically significant effect size, which we set to 0.25. This value never exceeds 506 because we proceed to a full trial only when the observed pilot study effect size is greater than or equal to the practically significant effect size. However, even if we relaxed our rule for proceeding to trial, it would still make sense only to enroll enough cases to detect a practically significant result in order to conserve resources.

If we subtract the average sample size calculated using the observed pilot study effect size from that required to observe the practically significant effect size, we can quantify how many cases are "saved" by using the observed pilot study effect size to calculate power, which yields smaller sample size estimates, instead of the practically significant effect size, which will always yield a larger sample size estimate under our scenario of when to proceed to a trial. In Scenario 3, we "save" on average 369 cases; in Scenario 4, we save an average of 433 cases. These saved cases, representative of research dollars required to run larger studies, illustrate the allure of using pilot study effect sizes to conduct power calculations. We place saved in quotations because, as detailed below, these "savings" are illusory.

Concluding whether an intervention is effective. For each simulated full trial, we now conclude whether an intervention is effective using a standard null hypothesis statistical test. When we do find a statistically significant effect, we assess how often the observed effect size point estimate overestimates the true effect size. A summary of these results is provided in Table 4.

Power loss. The first row of Table 4 shows how often we detect a treatment effect (i.e., reject the null hypothesis) after conducting a full trial (in this case, when the pilot study effect size was larger than the practically significant effect size) given different simulation parameters and null hypotheses.

In Scenarios 1 and 2, the true effect sizes (0.0 and 0.20, respectively) are smaller than the practically significant effect size of 0.25. We would consider it to be a mistake to proceed to a full

Table 4. Simulation Results Across Scenarios (Percentages).[a]

True effect size							
Scenario 1		Scenario 2		Scenario 3		Scenario 4	
0.00		0.20		0.50		0.80	
Power Full Trial to Detect Pilot Study Effect Size	Power Full Trial to Detect Practically Significant Effect Size	Power Full Trial to Detect Pilot Study Effect Size	Power Full Trial to Detect Practically Significant Effect Size	Power Full Trial to Detect Pilot Study Effect Size	Power Full Trial to Detect Practically Significant Effect Size	Power Full Trial to Detect Pilot Study Effect Size	Power Full Trial to Detect Practically Significant Effect Size
Over all simulations where full trial was conducted (i.e., pilot study effect size was greater than the practically significant effect size), percentage of simulations where null hypothesis was rejected (i.e., effect detected)							
5	5	29	61	68	100	77	100
Over all simulations, including those where no full trial was conducted, percentage of cases where null hypothesis was rejected in the full trial ("achieved power")							
1	1	13	26	56	81	75	97
Over all simulations where null hypothesis was rejected (i.e., effect detected), percentage of cases where true effect size was overestimated in the full trial (%)							
51	49	97	50	68	50	63	50

[a] In each case, the null hypothesis is that the effect size is 0. Results are based on two-tail *t*-tests.

trial for these scenarios. Moreover, in Scenario 1, if we did proceed to a full trial, any detected statistically significant effects would be Type 1 errors because the true effect size is 0.0. In Scenario 2, even if we fully powered a full trial to detect the practically significant effect size (instead of using the observed pilot study effect size), any detected statistically significant effects would be overestimates of the true effect size because the true effect size is smaller than the practically significant effect size with which we powered the trial. Thus, it does not make sense to discuss power loss for these scenarios.

In Scenarios 3 and 4, the true effect sizes (0.50 and 0.80, respectively) exceed the practical significance level of 0.25. When we do power calculations for a full trial using the observed pilot study effect size from each simulation, we reject the null hypothesis 68% of the time in Scenario 3 (true effect size of 0.50) and 77% of the time in Scenario 4 (true effect size of 0.80). Given our power level of 0.80, this indicates a 12% point loss of power in Scenario 3 and a 3% point loss of power in Scenario 4. This shows how using pilot study effect sizes for power calculations results in underpowered studies. We would like to emphasize that these power losses are amplified, when true effect sizes are smaller. For example, if we rerun our simulations with a true effect size of 0.30—a level that is slightly over the practical significance level of 0.25 but smaller than the true effect sizes in Scenarios 3 and 4—we detect effects in full trials using a two-tailed *t*-test only 48% of the time, representing a power loss of 32% points.

"Achieved power" loss. The above figures account for only the times where we proceed to a full trial. Using simulations allows us also to consider power lost across the whole research process from pilot study to full trial. Here, we define achieved power as rejecting the null hypothesis in the full trial given that a true effect exists, across all of the above-described stages of our simulation from deciding whether to conduct a trial at all to actually conducting a full trial. In other words, achieved power is lost both by not proceeding to trial and by failing to reject the null hypothesis in the full trial. As above, we continue to trial only in simulations where the observed pilot study effect size

point estimate is greater than the practically significant effect size of 0.25 (readers could consider other decision rules using the online software).

The second row of Table 4 shows achieved power lost. In Scenario 3, when we power the full trial using the observed pilot study effect size, we conclude that there are statistically significant differences between the treatment and control groups (using a two-tailed *t*-test) only 56% of the time across all simulations. In Scenario 4, we detect an effect 75% of the time. This represents achieved power loss of 24% and 5% points, respectively, for Scenarios 3 and 4. Again, we would like to emphasize that power loss is amplified when true effect sizes are lower. Running the simulation with a true effect size of 0.30 leads to us detecting a true effect in only 28% of all simulations, representing achieved power loss of 52% points.

Overestimating effect sizes. The third row of Table 4 shows how often we over estimate the true effect size in our full trials. When using pilot study effect sizes as the basis for full trial power calculations, we overestimate the true effect size in 68% of cases in Scenario 3 and 63% of cases in Scenario 4. In a fully powered trial, due to random chance we would expect to overestimate the true effect size in one-half of cases. This high rate of overestimation is a consequence of powering the full trial using the observed pilot study effect size: the full trials are systematically powered to detect effect sizes *larger* than the true effect size, and so cases where an effect is detected in the full trial also overestimate the true effect size.

Conclusions about the misuse of pilot studies for informing research design choices. Simulating possible uses of pilot studies to inform research design choices leads us to three conclusions. First, using unbiased effect sizes from underpowered pilot studies to decide whether to conduct a full trial is a considerably error-prone and volatile practice. Even applying lenient rules for proceeding to a full trial leads to prematurely abandoning effective interventions in a high number of cases. Conversely, applying these lenient rules to an ineffective intervention results in mistakenly conducting full trials in a substantial number of cases. Both of these results are a consequence of the uncertainty inherent to conducting studies with small sample sizes.

Second, using observed pilot study effect sizes as the basis for power calculations in full trials results in significant loss in statistical power in comparison to basing power calculations upon a predetermined practically significant effect size.

Third, using observed pilot study effect sizes as the basis for proceeding to trial and then using the same effect size as a basis for power calculations for those full trials results in systematically overestimating the effect size in cases where statistically significant effect sizes are detected in those full trials. The closer in magnitude the true effect size is to the practically significant effect size, the more likely the true effect size will be overestimated in the full trial.

These conclusions are a logical consequence of the power analysis calculations taught in any "Statistics 101" class. Yet, we find that a reminder of these lessons is important and that the simulations make what we all know in theory from power formulas more concrete and the results more stark.[4]

Properly Using Pilot Studies in Experimental Evaluation Research

Despite the cautions we have offered so far, we believe without question that pilot studies are useful and have a place in the development of interventions and evaluations of their effectiveness. In light of these conclusions, but also in recognition of real-world research needs, we offer the following suggestions and thoughts for researchers and practitioners. We have organized our discussion by potential uses of pilot studies.

Using Pilot Studies as Justification for a Full Trial

Because of the uncertainty inherent to small sample pilot studies, researchers conducting randomized experiments should never use pilot study effect sizes alone to determine whether a full trial should be conducted. Instead, pilot study results should be used as one data point among many to determine whether a full trial is worth conducting. Other factors to consider include qualitative impressions from researchers and pilot study participants on how well the intervention worked in the field, the fidelity of the intervention's deployment in the field vis-à-vis the intervention's theory of change, and past experimental and nonexperimental research on the plausibility of such an intervention being effective. If the results of a pilot study reflect promise in every regard but the observed effect size, researchers may still be justified in conducting a full trial, especially if, for example, there is strong nonexperimental evidence of the intervention's effectiveness. And, conversely, in cases where a small pilot study showed a large effect size but the theory of action is unclear, more effort should be put into determining how realistic that effect size is and whether the intervention is in fact promising.

We nevertheless understand that due to the limited resources to which researchers and funders have access, information from pilot studies will inevitably be used to determine whether or not a full trial is worth conducting. We find this entirely sensible, as the pilot study will provide the most concrete evidence of the intervention's feasibility and potential effectiveness. However, we would argue that the effect size obtained in a pilot study is only one such piece of evidence—and, as our simulations show, an unreliable one. We also highlight that this may be an area where strong nonexperimental study designs may help yield evidence on an intervention's effectiveness. There are likely cases where a well-done nonexperimental study will yield more accurate results than a very small pilot study (see Imai, King, and Stuart [2008] for a framework for thinking about these trade-offs).

Deciding whether to conduct a full trial poses researchers and funders with a tough choice that must be made with incomplete information about the intervention's true effectiveness. Although we cannot offer any formula for determining when a full trial is justified, we are comfortable in saying that if "everything is there but the effect size" after a pilot study's conclusion, researchers are likely justified in conducting a full trial. On the other hand, we would suggest that if "nothing is there but the effect size," researchers are justified in choosing not to conduct a full trial.

For example, consider a scenario where "everything is there but the effect size." We may observe a small effect size that does not even come close to approaching statistical or practical significance. However, we may also observe that everything else about the intervention suggests it is effective: The program has a strong and evidence-based theory of change, the program was implemented with fidelity, and participants expressed that they thought it was effective and could explain exactly how it helped them. In such a scenario, we would argue that researchers could justifiability press forward with a full trial.

On the other hand, consider a scenario where "nothing is there but the effect size." In this scenario, a pilot study may yield a large, statistically significant effect size, suggesting it is an effective intervention. However, we may also observe that recruitment was poorly done, that randomization resulted in treatment and control group that are unbalanced on baseline covariates, that participants expressed a strong dislike for the intervention, and that participants even reported not implementing the intervention with any fidelity. In such a scenario, even if we observe a large effect size, it would *not* be advisable to conduct a full trial with the intervention as it is. It is likely that the observed effect is a result of factors other than the intervention. All the information available suggests that the effect size is not a reliable indicator of the program's effectiveness and should be discontinued.

Our point here is that deciding whether to conduct a full trial based upon a pilot study requires making tough and possibly expensive choices without knowing with certainty whether an intervention is actually effective. The process is inherently risky. Periodic failure, even given a good decision, is inevitable. However, a careful consideration not only of quantifiable indicators but also of numerous qualitative indicators can be used to make justifiably good decisions. Through practice and conversation with other people conducting pilot studies, we can improve how well we assess the merits of pilot studies, but we should not expect to develop a single correct way to assess the merits of pilot studies.

Using pilot studies to inform power calculations. To avoid Type 2 errors and overestimating the true effectiveness of interventions, full trials should always be powered to detect the practically significant effect size. While it may be alluring to save resources by sampling fewer cases in light of a large observed pilot study effect size, this always results in conducting an underpowered (vis-à-vis the predetermined level of practical significance) full trial.

Consequently, before conducting a pilot study, and especially before designing a full trial, researchers should agree on an effect size they consider of practical significance and assure that resources are available to sample enough cases to achieve the statistical power necessary to detect the chosen practically significant effect size. Consequently, coming to a consensus on what effect size is of practical importance should be done deliberately. When deciding on what constitutes a practically significant effect size, we encourage researchers to review available standards in their field, to review past published research, and to consult with expert researchers and practitioners.

Finally, we want to further emphasize that using pilot study effect sizes to power trials is problematic not only because it results in underpowered trials. It is also problematic because when interventions in such underpowered trials are found effective, the observed effect sizes systematically overestimate the intervention's true effectiveness. Put succinctly, using pilot study effect sizes to power full trials leads to incorrectly judging effective interventions to be ineffectual, and judging effective interventions to be more effective than they are.

Researchers are routinely warned about the dangers of small sample sizes and know on some level to avoid reliance on them, but we find that all too often, when encountering a particular study, researchers forget those broad cautions. Although the results in this article are a direct result of standard power calculations that many evaluation researchers have learned, we believe it is valuable to see through simulation how those calculations impact practice. It is often difficult to understand exactly what those formulas imply about study design and the implications. We hope the evidence and arguments in this article help convince researchers to take seriously the implications of the improper use of pilot studies in conducting experimental research and program evaluations.

Using pilot studies to assess the feasibility of a full trial and improve its implementation. Pilot studies can be effective tools to help determine the feasibility of carrying out a larger trial of an intervention. For example, pilot studies allow researchers to judge the adequacy of recruitment and consent procedures, to evaluate the acceptability of randomization procedures, to gauge the quality of measurement instruments, and to assess whether compliance with the intervention in the field will be sufficient to achieve an intervention's goals (Lancaster et al., 2004). To achieve these potential benefits, researchers should systematically measure intervention implementation fidelity throughout the pilot study. Moreover, researchers should assess practitioners' experience trying to implement the intervention through field observation and interviews. In addition to helping researchers decide whether a full trial is worth conducting, the above data collection efforts can be used to improve the quality of the full trial should one be conducted.

Using adaptive and group sequential designs instead of pilot studies to improve intervention and evaluation quality. Researchers doing experimental evaluation research in contexts where a separate pilot study might be infeasible due to lack of time or resources should consider adaptive and group sequential designs. These designs may be feasible for programs for which it is feasible to collect outcome data early in a trial. For example, internal pilot studies are a type of adaptive design proposed as a way to forgo ordinary, or external, pilot studies with separate recruitment and treatment periods (Wittes & Brittain, 1990). In group sequential designs, differences between treatments groups are assessed at multiple points, at which point researchers can choose to either reject the null hypothesis, stop the trial early, or continue it (Baldi et al., 2011; Whitehead, 2004).

These approaches may not be feasible in research designs where the time and cost of adding new cases is prohibitively impractical or expensive (e.g., when training an entire school in a new curriculum intended to be deployed over an entire semester or school year), or where outcomes are measured after a long period of time. However, when feasible, the appropriate use of adaptive and group sequential designs can achieve many of the goals of pilot studies, such as improving the statistical power of evaluations, ensuring that more students receive helpful treatments, and potentially improve the quality of the treatment.

Conclusion

Throughout this article, we first documented how pilot studies are rarely used in experimental evaluation research. We discussed reasons why this might be, with a focus on evaluation in educational settings. We then used statistical simulations to document potential misuses of pilot studies, such as using them in isolation to determine whether to conduct a full trial or using effect sizes from pilot studies as the basis of power calculations. Finally, we discussed potential uses of pilot studies and offered advice on how pilot studies can be used more effectively in experimental evaluation research. We hope these discussions provide clarity to readers as to why pilot studies are so rarely conducted in experimental evaluation research and how they can be effectively used to improve the quality of interventions themselves and the research of their effectiveness.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in-part by the Johns Hopkins University Interdisciplinary Predoctoral Training Program, funded by the U.S. Department of Education Institute of Education Sciences (IES).

Notes

1. In general, our chosen practically significant effect size falls within the range of typical effect sizes used in power calculations in education research (Spybrook & Raudenbush, 2009). Likewise, it is close to the average effect size observed in experimental evaluations. For example, Slavin and Smith (2009) did a systematic review of mathematics programs and found an average effect size of 0.24 for those mathematics programs evaluated using randomized experiments.
2. Although the convention established by Cohen (1988) deems 0.20 a small-sized effect, 0.50 a medium-sized effect, and 0.80 a large-sized effect, we would suggest that 0.50 is itself large, especially for fully powered evaluations. For example, in Slavin and Smith's (2009) review of mathematics interventions, they found an average effect size of only 0.09 for trials with over 2,000 subjects. Effect sizes of over 0.50 were found only

in trials with sample sizes of less than 50 subjects, suggesting that these effect sizes may be anomalous or difficult to achieve when scaled up. See Hill, Bloom, Black, and Lipsey (2008) for an anslysis of different effect sizes in different educational contexts.

3. Source code and a link to a web-based version of our simulation program is available at https://github.com/Table1/PilotPower.
4. It is important to note that many evaluation studies are multilevel. For example, in an education content, students may be clustered in classrooms, which are clustered in schools. In these multilevel contexts, power calculations must be based upon an ''operational effect size'' that accounts for the design effects introduced by cross-school variation (Hedges & Rhoads, 2010). For simplicity, we have considered a single level setting, but the overall conclusions would be the same if we accounted for a multilevel structure. Fundamentally, using pilot studies as the basis for subsequent power calculations will cause just as much, or more, trouble in multilevel settings. Although outside the scope of the current article, future work should further investigate the exact consequences in multilevel settings, for example, by varying the intraclass correlation at the different levels and seeing how that affects power.

References

Baldi, I., Gouchon, S. M., Di Giulio, P., Buja, A., & Gregori, D. (2011). Group sequential and adaptive designs: A novel, promising tool for nursing research. *Journal of Advanced Nursing*, *67*, 1824–1833. doi:http://doi.org/10.1111/j.1365-2648.2011.05651.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). London, England: Routledge.

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research*. Retrieved from http://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf

Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health*, *31*, 180–191. doi:http://doi.org/10.1002/nur.20247

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*, 481–502. doi:http://doi.org/10.1111/j.1467-985X.2007.00527.x

Johanson, G. A., & Brooks, G. P. (2010). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement*, *70*, 394–400. doi:http://doi.org/10.1177/0013164409355692

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. doi:http://doi.org/10.1001/archpsyc.63.5.484

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. doi:http://doi.org/0.1111/j.2002.384.doc.x

Liu, P. Y., Dahlberg, S., & Crowley, J. (1993). Selection designs for pilot studies based on survival. *Biometrics*, 391–398. doi:http://doi.org/10.2307/2532552

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.

RStudio, Inc. (2013). *Shiny: Web application framework for R. R package version 0.8.0.99*. Retrieved from http://www.rstudio.com/shiny/

Schoenfeld, D. (1980). Statistical considerations for pilot studies. *International Journal of Radiation Oncology* Biology* Physics*, *6*, 371–374. doi:http://doi.org/10.1016/0360-3016(80)90153-4

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*, 500–506. doi:http://doi.org/10.3102/0162373709352369

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the institute of education sciences. *Educational Evaluation and Policy Analysis*, *31*, 298–318. doi:http://doi.org/10.3102/0162373709339524

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., & Rios, L., ... Goldsmith, C. (2010). A tutorial on pilot studies: The what, why and how. *BMC Medical Research Methodology*, *10*, 1. doi:http://doi.org/10.1186/1471-2288-10-1

U.S. Department of Education. (2016). *Request for applications: Special education research grants* (CFDA Number: 84.324A). Retrieved from http://ies.ed.gov/funding/doc/2017_84324A.docx.

Vickers, A. J. (2003). Underpowering in randomized trials reporting a sample size calculation. *Journal of Clinical Epidemiology*, *56*, 717–720. doi:http://doi.org/10.1016/S0895-4395(03)00141-0

Westlund, E., & Stuart, E. A. (2016). *PilotPower*. Retrieved from https://github.com/Table1/PilotPower

What Works Clearinghouse. (2011). *WWC procedures and standards handbook: What works clearinghouse* (Procedures and Standards Handbook No. 2.1). Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19#

Whitehead, J. (2004). Stopping clinical trials by design. *Nature Reviews. Drug Discovery*, *3*, 973–977. doi:http://doi.org/10.1038/nrd1553

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, *9*, 65–72. doi:http://doi.org/10.1002/sim.4780090113

Zucker, D. M., Wittes, J. T., Schabenberger, O., & Brittain, E. (1999). Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine*, *18*, 3493–3509.