Boruch, R., Allen-Platt, C., & Gerstner, C.-C. (2019). To randomize or not to randomize? That is the question. In C. A. Christie & M. C. Alkin (Eds.), *Theorists' Models in Action: A Second Look*. *New Directions for Evaluation*, *163*, 73–82.

6

To Randomize or Not to Randomize? That is the Question

Robert Boruch, Claire Allen-Platt, Clara-Christina Gerstner

Abstract

In this chapter, Robert Boruch, Claire Allen-Platt, and Clara-Christina Gerstner explore whether an evaluation that employs a randomized controlled trial is appropriate for the Women Affirming Motherhood program. To make this determination, they focus on the program's clientele and pipeline, how programming is conducted, and what outcomes and effects should be explored. The authors conclude that an exploratory study is necessary before a summative evaluation can be pursued.

Introduction

Women Affirming Motherhood (WAM) recently received substantial support from the Empire Foundation for WAM's work in assisting pregnant young women. As part of the grant, Empire required WAM to elicit bids for an independent evaluation of WAM's performance. In what follows, the focus is on evaluating WAM's effects. In particular, the concern is whether and how a randomized controlled trial (RCT) might be deployed so as to produce a fair estimate of WAM's effects.

The approach to addressing the concern is interrogatory, as the title suggests. The topic's handling here is more Socratic than it is Shakespearian, however.

Assumptions About WAM and Interest in Its Effects

Program managers, foundation people, and a prospective trialist must depend on assumptions about the state of play. This article depends on Dillman, Scott, and Kinarsky (in this issue) characterization of the program. In particular, WAM's targets are declared to be expectant mothers in low-income areas. WAM's leadership includes able people who developed what they believe to be a promising program. With foundation funding, the assumption is that the WAM staff will continue to service the mothers. If this information is not dependable, the trialist may withdraw from engagement.

The prospective trialist may assume, or must assume, that the potential users of evidence about WAM's effects are serious in this interest. Identifying the potential users is important to support this assumption. For instance, they may include WAM's leadership and staff, the Empire Foundation, and other stakeholders including WAM's target population. For the trial's design, an advisory panel that includes such people is usually helpful to the design process. More about this anon.

If there is no real prospect of fair evidence being used by any of these stakeholders, or if there will be no serious input from stakeholders in the trial's design, there is not much point to a trialist going further.

Interrogatory Approach to Deciding About RCTs, and Their Design

In evaluation studies generally, and in deciding about whether to engage in RCTs in particular, three questions are fundamental.

Q1. What is the nature, severity of the problem and how do we know?
Q2. Has the program that is designed to address the problem been deployed, and how do we know?
Q3. What is the program's effect and how do we know?

These questions, put in other vernacular, underlie evaluation policies of some foundations, multi-national organizations, and federal agencies in the United States and in textbooks on evaluation. See for instance Boruch, Chao, and Lee (2016) and especially the references therein.

Addressing the last question, about effect, depends heavily on evidential answers to the first two questions. Absent good evidence on these two, designing good RCTs, or any impact study is likely to be unsatisfactory at best. It may be futile at worst. Each question is considered in the WAM context in what follows.

Q1. What Is the Nature and Magnitude of the Problem and How Do We Know?

Dillman, Scott, and Kinarsky (in this issue) reported that WAM's take-up rate at the end of the first-year was "a little more than a hundred expectant

mothers were served, with demand for service increasing regularly . . . " (p. 21). The report further avers that " . . . the number of young women seeking WAM's services was steadily growing . . . " in Year 2 (p. 21). This is promising for funders of course.

For a trialist, however, the information is vague. Demand is said to be "Increasing regularly . . . . and steadily growing" But the trialist must ask: by how much exactly? Steadily 10%? 2%? What? Determining how much is important partly because designing a good RCT depends on the trialist's understanding of the number of young women in WAM's pipeline. Pipeline studies are required in advance of many RCTs in medical, criminological, and educational sectors. In reporting on final results of a trial, the CONSORT Standards, for instance, require such information to assure transparency of RCTs results (CONSORT, 2017).

More to the point of designing an RCT so as to estimate WAM's effects, small samples, such as 100–200, often require special attention from the trialist. The mobility of the young women is important. If they disappear fast from the WAM pipeline, WAM's effect will be difficult to detect or understand. A good trialist might then suggest that WAM's evaluation then be considered an exploratory effort, rather than a confirmatory or summative effort.

In particular, the trialist may suggest that (a) the small sample study be considered as exploratory and (b) the probability of outcomes be the focus rather than formal tests of statistical hypotheses. Randomization tests, for instance, give a probability of a difference in outcomes, as opposed to inviting, or being seduced into, formal decisions based on a $p = .05$ or some other artificial threshold. The fourth edition of the Edgington and Orghena (2007) book is a dependable resource on this account.

A fundamental aspect of the WAM scenario concerns WAM's "client base." No information is provided about the base, apart from the fact that the young ladies are pregnant, at high risk, and living in a poor area of a mid-western city. It is in the interest of the trialist to learn about who refers the women to WAM and the characteristics of the women who are referred? More importantly, who are these young people? Are they school-aged? Employed? On welfare? Obese? In a functional relationship? Are they transient? Understanding what the client base here is important in understanding whether there will be ceiling effects on outcome measures; for instance, many women who enter the program may already be knowledgeable about handling the challenges of pregnancies. Identifying important subgroups in advance is important in designing an RCT and anticipating analysis issues.

In the absence of dependable information on the young ladies in the WAM pipeline, it is difficult, or impossible, to design a good RCT. No decent statistical power analysis is possible, for instance. No sensible interpretation of the RCT's results, including generalizability, is possible absent dependable information on the WAM client base and referral system.

If there is good initial evidence on these matters, the trialist may then proceed further. Otherwise, he or she may pull the plug, for good reason.

Q2. Is the Program Deployed and How Do We Know?

Over the past 40 years, big foundations have given big dollars to programs that have not been deployed or have been deployed poorly. Nowadays, the best foundations require evidence about this. They further ask how the money was spent.

Evaluators who specialize in program implementation or formative assessments are typically more informed about program operations and practice than a specialist in RCTs can be. Specialists in formative evaluations, for instance, can assist in understanding WAM's deployment, and how WAM's services cut across agencies. Any given service component may demand attention to indicators of service delivery and receipt. Undergirding the deployment is a theory of change, that is, a set of ideas about what should happen in WAM's service delivery and about what should happen to young women as a consequence of its delivery.

In designing a randomized trial, the prospective trialist would be foolish to ignore WAM's theory of change, and the evidence that might be offered by implementation specialists about WAM's deployment. Absent such evidence, or the promise of it, the trialist may desist from further work. The trialist may then proceed if there is a promise of evidence on WAM's deployment in addition to evidence on WAM's client base.

Q3. What is WAM's effect? How Do We Know?

In the case at hand, designing a randomized controlled trial depends on evidential answers to Question 1, on WAMs' pipeline and client base, and Question 2, on WAM's deployment. Let us assume that the questions have been addressed, or will be addressed as part of a trial.

This gets to the hard work by the trialist: reviewing the relevant literature, identifying the main outcome variables, structuring the random allocation and explanations of it, and staging a trial so as to benefit from inevitable missteps or mistakes.

Reviewing the Relevant Literature

Industrious trialists will inquire of colleagues who have done related work so as to learn from them. For the WAM case, a trialist will try to learn from the published literature who has measured what in randomized field experiments that have direct attention to young pregnant women in high poverty areas, how, when, and why the trial was done, and what the outcomes were.

In the health care sector, for instance, the Cochrane Collaboration regularly produces reviews of evidence on related programs. The Cochrane

Library covers evidence on effectiveness of interventions designed to avert or handle domestic violence against pregnant women, the antenatal issues that affect women in low-income areas, and the handling of prenatal and postnatal health and behavior issues among the women. See http://cochranecollaboration.org.

Outcome Measures

Deciding which outcome variables should be measured is no easy matter in a multicomponent program like WAM. WAM's stakeholders and Foundation people are important in this, as is the theory of change.

The burden falls partly to the trialist to encourage identification of pertinent outcomes, and often to identify dependable ways to measure them. The programmatic or policy decision about what outcome to measure is usually not in the trialist's purview.

Snares in any such negotiation on this topic are common. The challenges for the trialist include the WAM's opining that a dozen outcomes are important. The health of the young woman and health of the baby are, to be sure, relevant. But so too are outcomes such as the women's knowledge and perhaps also her beliefs, attitudes, and observable links to the social networks that WAM facilitates.

Typically, all potential outcomes are not equally important to everyone. The trialist, must nonetheless seek agreement on which one or two or three outcomes have the highest priority and why. Setting priorities is basic because: (a) measuring all plausible outcomes well is likely to be infeasible, (b) measuring lots of plausible outcomes guarantees that some effects will be positive, some negative, and many will be negligible, which inevitably leads to squabbles about the implications of results, and (c) identifying a couple of primary outcomes does not mean that others be ignored. The result of negotiation usually entails classifying outcomes into the primary ones and the ones of secondary importance. This step could be done with the advisory panel.

Absent a negotiated agreement that a couple of outcomes are high priority and can be measured dependably, and absent agreement that there are secondary outcomes that are important and measurable, a prospective trialist may excuse himself or herself without disgrace.

Random Allocation and Fairness

Designing an RCT in this context requires some agreement among stakeholders about random allocation of pregnant young women, to WAM or to a control condition (which might be an existing community program). The random allocation will yield defensible evidence that the estimated effect is transparent and fair.

For example, if the demand for WAM's services far exceed WAM's capacity to supply services, then random allocation of young pregnant

women to the WAM program may be acceptable to stakeholders. The negotiation on this may involve ethical, managerial, and political values. It depends of course on the realities of the pipeline.

When demand exceeds supply, and when lottery allocation is acceptable to stakeholders, the technical issues in the design are easy. For instance, trialists have easy access to software for substantial power analysis, assuring that a WAM effect can be detected, for example, Dong and Maynard (2013). Trialists have access to technology that incorporates blocking factors and covariates, such as types of pregnant women (single mothers, drug users, diabetics), into the experiment's design.

Statistical power analysis is essential. But it is not enough. In designing the RCT, good trialists will review work on earlier trials on programs that seem relevant. Systematic reviews, of the sort published by the Cochrane Collaboration in health and the Campbell Collaboration in social sectors, are important to anticipate WAM's effect size, attrition rates and other issues related to the statistical power of the trial.

Ethics and Random Assignment

WAM's description does not tell us that the program is mandatory, for example, ordered by a family court or some other legal authority. Assume, as is usually the case, that the women's participation in the program and the trial is voluntary. In such cases, the idea of informed consent is crucial. But there are other ethics issues in this scenario.

For instance, the ethical trialist might well ask whether the program is important enough to test in a trial. If it is not, then the trial results are likely to be trivial at best and waste people's time at worst. Assume that WAM is important enough to justify a trial. Ethics questions remain.

If WAM's important ingredients are known to be effective, based on earlier dependable evidence, deploying a randomized trial may not be warranted. For instance, no one does Salk vaccine trials nowadays to test the vaccine's effectiveness because effects have been well established. One might however ethically test WAM's variation under the assumption that it is potentially better than others, just as testing different doses of vaccines can be ethical.

Depending on the local preferences and constraints, trap door approaches can be tailored to the trial. For instance, a prespecified fraction of the entering cohort of women may a priori be identified as in absolute need for the program, granted entry to it, and consequently not be part of a lottery allocation. The remaining women in the cohort, who do not meet these advance criteria but who nonetheless are eligible for program admission, are subject to the lottery allocation. This tactic is a variation on more general regression discontinuity designs that do not rely on complete random allocation of individuals in a cohort.

Would methods other than a randomized trial yield dependable evidence about its effects? If so, randomization may not be necessary and even be unethical. For instance, times series data on women in WAM's catchment area might be sufficient, when the interruption of the series is obvious, without plausible competing explanations, the program is sharp, and the time series data are adequate. These conditions are likely to be unmet in the scenario at hand.

One might explore propensity score approaches, which are more sophisticated than older covariance analyses. These require identifying the right matching variables (covariates), measuring them in the right way, and putting them into the right functional form/model for reasonably defensible statistical estimates of effect. Identifying and discussing alternatives to random assignment as a device to produce fair comparisons are typically debatable matters. Assume for the sake of argument, that the alternatives are not feasible, will not yield persuasive evidence, or will yield evidence that is far less transparent than randomized trial results would be. The trialist may then proceed.

Explaining Random Allocation to Young Pregnant Women

How does one explain to pregnant young women the fact that they will be randomly allocated to a special program? Experience from other sectors suggests that explanations have to be tailored to the people involved. To cut to the rhetorical chase, the word "experiment" in some contexts is unacceptable. The phrase "randomized controlled trial" is meaningless in others. The word "lottery" may be tolerable, and familiar, in common parlance. The idea of a lottery is fair in many cultures.

Learning how to explain well can depend on focus groups of potential WAM participants in anticipation of the RCT, and on the judgments of stakeholders of course. See for instance Rockefeller Foundation's video tape documentary on the Minority Female Single Parent Program. The taping grew out of initially flawed efforts to explaining the trial and shows how explanation may differ with culture.

Executing RCTs Under Uncertain or Unstable Conditions and Learning From Failure

Absent pre-RCT efforts and evidence, the trialist will be a troubled soul. Less troubled than Hamlet, but troubled nonetheless. In uncertain conditions, one option is obvious. The trialist might run the trial in two stages. The first stage, a "run-in trial" in engineering parlance, is deployed so everyone makes all the mistakes one can make. The trialist and everyone else can then learn how to deploy the second stage and do a better trial.

In olden days, for instance, cops, nurses, and other service providers subverted the random allocation at times, for example, by simply ignoring the assignment and treating the client as they pleased. Audits and

mechanisms can be built to suit the particular setting so as to reduce or eliminate the problem. Similarly, information that permits tracking of women over time can be built based on experience so as to reduce attrition, missing data, and so on. Doing RCTs in stages is a simple tactic for reducing uncertainty gradually for trialists in the social sciences, health, engineering, physical sciences.

Innovative programs that are expected to produce an expected effect do not always do so. Failures to detect effects, despite a well-designed trial, in are not uncommon in medical, educational, and social services evaluations. The state of the art in designing trials so as to better learn from inevitable failure is advanced in the engineering and medical sciences. Though not so advanced in the social sector, there are promising approaches including the stage-wise tactic described above.

For instance, Clara-Christina Gerstner, Claire Allen-Platt, and others are reviewing reports on results of large scale randomized trials in education and the post facto analyses of the trials' qualitative and quantitative results (Allen-Platt & Gerstner, 2019). The literature in the education sector is inchoate and not entirely coherent as one might expect. Much of it fails to be orderly and transparent in discussing plausible reasons for failure to detect effects, and the challenge is to find and distill the discussions worth attending to. Low participation rates in programs are discussed, for instance, but few reports inform the reader about the usefulness of incentives and the authors try to identify them. Their taxonomy includes attention to relationships among targets of the program and program staff and management, the nature of cooperative agreements and contracts, insufficient or variable implementation of programs, instability of the kinds discussed above that can be anticipated based on prior data, and other factors. This empirical and taxonomic work in education updates and expands on earlier work by Boruch, Dennis, and Carter-Greer (1988) among others in the case of minority single parents.

Small Sample Issues

WAM's pipeline is modest, involving hundred young women in each cohort. For the trialist, this invites two themes. One concerns rhetoric and policy. The second directs attention to technical approaches to handling the matter. They are interrelated.

Small samples of program participants invite the trialist, and WAM's stakeholders, to denominate the study as exploratory, or as a pilot test. This nomenclature reflects a reality. Its use is important in reducing the temptation to over advertise potential results. More importantly, it lays the groundwork for scaling up or for scaled up trials if indeed the pilot yields promising results.

Even in small samples, random assignment guarantees that there will be no *systematic* bias in comparing WAM's outcomes to a control condition

or to other programs. This is fundamental. Nonetheless, chance outcomes should be taken into account. This is despite the fact artificial thresholds such as *p* values and statistical power recede in their importance for small samples in pilot efforts, and that the magnitude of a discernable difference is important.

Instead of formal tests of hypotheses and the conventional power standards, the trialist may employ simpler approaches invented by Fisher, and elaborated and accompanied by software by Edgington and Orghena (2007) among others. Consider the following illustration.

Suppose that the outcome variable indicates each woman's wellbeing on a score with range 1-5. Further suppose that the sample is really small, with five young ladies being randomly assigned to WAM and five randomly assigned to a control condition. Such a sample is far too small for conventional power analyses or tests of hypotheses unless the expected effect size is huge, that is, the trialist and stakeholders are absurdly optimistic.

In simple randomization tests, one ignores the actual assignments and identifies all possible assignments of each woman to a group. In a scenario involving ten women assigned to two groups, there are 252 such assignments. One then computes the 252 mean differences in the scores of the two groups so composed. At one extreme, for instance, all women in one group may have a score of 5 and all women in the other condition have a score of 1, leading to a simple difference in average scores of 4. At another extreme, all women in each group may have the same scores, leading to an average difference in outcomes of zero. Beyond these extremes, there is lots of variation.

In this hypothetical scenario, 15% of all the average differences between groups may be "big" by chance, that is, an average difference of 1 or more. If the actual observed difference in the trial is 1, then one may conclude that the magnitude is big and further that the probability level is tolerably small if there is no real WAM effect. It is not statistically significant under a 5% threshold value, but the result is promising.

Concluding Remarks

As readers might surmise, developing fair and transparent estimates of the effects of new programs is hard work. This is the case for WAM as well as others. The steps needed to do the job right are numerous. To some, they are tedious and not fast enough.

In other scenarios, experiments can indeed be brisker. The Silicon Valley slogan, "Fail often and fast," embodies a different perspective on trials for instance. The context often differs. Computer chips are irrelevant to ethical issues. Web-based commercial experiments on influencing site user preferences, however, usually involve fewer steps and less complex "programs" than something like WAM.

The series of questions enunciated above can be reconfigured as checklist items, of course. Taking a checklist approach helps to assure (a) procedural simplicity, (b) action topics are covered, (c) the generation of a dependable scientific log on the experiments, and (d) ingredients for interim and final reporting on WAM's effects.

Questions other than the three that are posed here can be laid out. For instance, can WAM be replicated and might the results of this particular WAM trial be replicated again in the same and other contexts? Further, how might the results of repeated trials on WAM like programs be recorded and submitted to systematic review? Though organizations such as Campbell Collaboration, the Cochrane Collaboration, and others do a good job, they cannot do the job well if results of trials are not reported.

Making incremental progress to assist young pregnant women in low-income neighborhoods in their learning how to make progress on their own is important. The evidence on progress is a product of randomized trials. Absent dependable evidence, the young women and we will make little or no progress.

References

Allen-Platt, C., & Gerstner, C. (2019). *Toward a scence of failure analysis* (Unpublished report). Graduate School of Education, University of Pennsylvania, Philadelphia, PA.

Boruch, R., Chao, J., & Lee, S. (2016). Program evaluation policy, practice and use of results. *Comparative Education Research Quarterly*, *24*(4), 71–90.

Boruch, R. F., Dennis, M., & Carter-Greer, K. (1988). Lessons from the Rockefeller Foundation's experiments on the minority female training program. *Evaluation Review*, *12*(4), 396–426.

CONSORT. (2017). *Transparent reporting of trials*. Retrieved from http://consort.statement.org

Dong, N., & Maynard, R. A. (2013). *PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies* [Software]. Retrieved from http://www.causalevaluation.org/

Edgington, E., & Onghena, P. (2007). *Randomization tests*. Boca Raton, FL: Chapman and Hall/CRC/Taylor and Francis.

ROBERT BORUCH is the University Trustee Chair and Professor of Education and Statistics at the University of Pennsylvania's Graduate School of Education.

CLAIRE ALLEN-PLATT is a PhD student in Quantitative Methods at the Graduate School of Education, University of Pennsylvania, Philadelphia.

CLARA-CHRISTINA GERSTNER is an MPhil student in Quantitative Methods at the Graduate School of Education, University of Pennsylvania, Philadelphia.