*Journal of MultiDisciplinary Evaluation*
Volume 18, Issue 42, 2022

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

# Conditions to Consider in the Use of Randomized Experimental Designs in Evaluation

George Julnes
*University of New Mexico*

Melvin M. Mark
*Penn State University*

Stephanie Shipman
*U.S. Government Accountability Office (retired)*

**Abstract:** Debates about the role of randomized experiments in evaluation have been heated at times, which likely has not facilitated and possibly has hindered thoughtful judgments about whether and when to use a randomized experimental design. The challenges of thoughtful deliberation may be especially great for funders and others who influence the choice of an evaluation design but are not immersed in methodological literatures. The current paper offers a non-technical summary of general factors to take into consideration when determining the appropriateness of a randomized design in a forthcoming evaluation or set of evaluations. Four general conditions are described that should be considered with respect to the specific context for the upcoming evaluation(s). These are, first, the expected value of the information that a well-implemented experiment can provide in the specific context; second, the legal and ethical issues that apply in the circumstances at hand; third, the practical constraints (or facilitating factors) that would apply to a randomized experiment in that context; and fourth, the likely value of the experimental findings in relation to and as part of a portfolio of evaluative studies in the specific context.

***Keywords:*** *randomized experimental design*

## Prologue

The document that follows was developed with the intention of supporting thoughtful judgments about whether and when to use a randomized experimental design in an evaluation or set of evaluations. The intended users of the document include U.S. government staff and others, such as the staffs of foundations or other funding agencies, who are charged with making or contributing to decisions about how evaluations will be done. The document is premised on the ideas that (a) relative to alternative methods, randomized experimental evaluations can be very effective in estimating the effects of a program or other intervention, (b) this benefit is not a given, but depends upon the specific circumstances that hold for a particular evaluation, and so (c) the choice of a randomized experimental design should be based on careful consideration of a set of relevant factors as they apply in specific circumstances.

The document is brief by design. It is devoid of references. It avoids detailed comparison of randomized experiments and other designs or methods. Nor does it attempt to sort out whether a situational factor applies to all randomized evaluations (or only to a subset), or whether a factor applies only to randomized designs (or also applies to other comparative designs). While these and other elaborations would be useful for some readers, the resulting complexity would undercut the document's value in providing a succinct overview of the major issues that should be attended to by those who are responsible for selecting or funding a randomized experimental design in a specific evaluation context. We anticipate that this brief overview may be especially useful for individuals without extensive training in research methods.

The accompanying document was developed by the authors as a subgroup of the evaluation policy task force (EPTF) of the American Evaluation Association (AEA). Thanks go to other members of EPTF for their comments. To be clear, however, this is not an official document of the EPTF or of AEA.

While we offer the document for use in its current form, we also welcome its modification. We hope that any modifications will be in keeping with our goal, which is neither to idealize nor to disparage randomized experimental designs, but rather to encourage thoughtful deliberation in support of the appropriate and effective use of such designs.

## Prologue, Part II

The development of this document was initiated and led by George Julnes. As a member of AEA's EPTF, George was familiar with the document "An Evaluation Roadmap for a More Effective Government (2019)," which the EPTF first produced in 2009 and updated in 2019. That document was originally developed in part as a "leave-behind" to be provided to congressional or agency staff or others with whom EPTF members met to discuss evaluation policy. Given past debate about the role of randomized experiments in evaluation, Julnes suggested that a comparable document be developed regarding the factors that should be taken into account when experimental methods are being considered. He contended that the need for such a document was especially acute for random assignment methods, given they had been the focus of past controversy, including claims in some quarters of their "gold standard" status.

Over time, and with valuable input from EPTF colleagues and others, George, along with Mel Mark and Stephanie Shipman, prepared the document that follows. It has not become an official AEA document. In hindsight this seems quite reasonable, as the document deals with only one approach to evaluation and AEA members employ a wide range of methods. At the same time, decisions about randomized experiments continue to be made, so we believe the document has value.

George drew on the document for an editor's note in the American Journal of Evaluation introducing a set of papers on experimental methodology (Julnes, 2020). Table 1 of that note is essentially a one-page summary of the general points of the current document, along with some related references in the accompanying text. While the condensed version in Julnes's note is not a substitute for the current document, it may serve as a useful complement for those who want an overview and some citations of relevant sources.

For evaluations of programs and policies in the U.S. context, the Foundations for Evidence-Based Policymaking Act of 2018 (P.L. 115-435) and related guidance from the Office of Management and Budget provide a more inclusive view of evaluation methods than did some earlier statements, which expressed a preference for randomized experiments and their closest quasi-experimental cousins. This shift suggests the need for a document that facilitates thoughtful decision-making about whether and when to use randomized experiments in evaluation may be less acute. However, even without heated debates as

part of the background context, decisions about evaluation methods in a specific circumstance will need to be made. Moreover, the general role of randomized experiments in evaluation is probably not settled everywhere and for all time. In addition, Julnes (2020) points out that attention of the sort suggested in this document will usually be more important for evaluations of social programs than for assessments of minor administrative procedures, such as when alternative versions of a recruitment letter are sent randomly to potential service recipients.

Thus we offer the current document. As noted in the first part of the prologue, it was designed primarily for a non-evaluation audience, such as government agency or foundation staff who are involved with commissioning evaluations. Of course, evaluators may play a role in sharing this document with these or other audiences who need to think about when, why, and whether to use a randomized design in an evaluation. Such deliberations might involve a single upcoming evaluation or an organization's overall evaluation policy. Our hope is that publication in the online, open-source journal JMDE will facilitate this document's dissemination, including to international audiences. Evaluators may also benefit from the document directly, for example, in evaluation training, in continuing professional development, and in reflecting on the mental models that guide their method choices. Such uses may be enhanced by awareness of other resources addressing the feasibility of social experiments, including a paper by Bell and Peck (2016) in JMDE.

Sadly, George Julnes passed away November 24, 2021. Mel Mark and Stephanie Shipman subsequently wrote the abstract and the second part of this prologue. Otherwise, with the exception of minor corrections and clarifications, including a handful suggested by reviewers and the addition of the brief reference section that follows, the text is that of the last version on which George worked. It is offered in honor of his memory and his commitment to the field of evaluation as a means of contributing to human well-being.[1]

[1] For those using this as a resource for non-evaluators, feel free to consider whether to exclude Prologue, Part II, which would result in a simpler document, essentially that which George Julnes left us with.

## References (for Prologue, Part II)

AEA. (2019). *An evaluation roadmap for a more effective government*. Revised. https://www.eval.org/Portals/0/Docs/AEA%20Evaluation%20Roadmap%202019%20Update%20FINAL.pdf

Bell, S. H., & Peck, L. R. (2016). On the feasibility of extending social experiments to wider applications. *Journal of MultiDisciplinary Evaluation*, *12*(27), 93–111. https://journals.sfu.ca/jmde/index.php/jmde_1/article/view/452

Julnes, G. (2020). Editor's Notes: Experimental Methodology. *American Journal of Evaluation*, *41*(4), 486–493. https://doi.org/10.1177/1098214020954305

## Overview

The Foundations for Evidence-Based Policymaking Act of 2018 calls for federal agencies to establish policies regarding how and when to conduct program and policy evaluation. In developing such policies, it is challenging to specify which methods to use without identifying the contextual issues that need to be considered before choosing one evaluation approach over another. This challenge is especially important for policies about which methods to use when decisions require evidence about the effectiveness of a program, or about which of two or more alternatives appears better. For example, should a pilot program be expanded nationwide? What are the effects of making Naloxone widely available, compared to current practices? A primary approach to answering these types of questions is to randomly assign persons (or other units, such as schools) either to a treatment or to a control condition (or to alternative treatment conditions). Such assignment is the key feature of a randomized experimental design, sometimes called a randomized controlled trial (RCT) or simply an experimental design.

The advantage of randomly assigning people, or other units, to treatment and control conditions is that it ensures the groups are essentially similar initially. This reduces the likelihood of bias that would result if instead one compared outcomes for groups that were quite different before the treatment (e.g., comparing the health outcomes of people in poor health who received a "treatment" of vitamin supplements with a control group of initially healthy people). With random assignment to conditions, the only initial differences between the groups are due to random chance. This reduction of bias can be very helpful, for example, when it is important to know how much better—if at all—an innovative program is compared to the status quo.

In recent years, because of their advantage in reducing bias in estimating program impacts, calls for more rigorous evaluation have often been paired with calls for the use of randomized experimental designs. However, in some situations experimental designs can be difficult to implement successfully or may not be appropriate. Some concerns apply also to evaluations without random assignment, but some of these are more acute for randomized assignment evaluations. It is important that those involved in evaluation decisions understand these potential challenges and are prepared to take them into account before planning an evaluation that employs an RCT design.

There have been differing views in the evaluation community about the appropriateness of RCT evaluations in different contexts. In recent years, however, there has been progress toward some consensus on both the strengths of and concerns regarding experimental evaluations of government policies and programs. Specifically, there is broader recognition of the value of using RCTs when estimating program or policy impacts, as well as acknowledgment of lessons learned about which conditions are favorable or unfavorable for experimental evaluations.

With these lessons in mind, this document was prepared to assist federal evaluators, as well as legislators and agency administrators, in making decisions about when, or whether, they might want to promote or require the use of experimental designs for evaluating program effectiveness. This document highlights four sets of conditions that should be considered when assessing the appropriateness and value of experimental evaluations in specific contexts.

The four sets of conditions are:

1. the potential information value of well-implemented experiments for estimating policy or program impacts;
2. the current legal and ethical context;
3. the practical constraints on conducting an experiment in the specific context; and,
4. the value of the experimental findings as part of a portfolio of evaluative studies conducted to address a wide array of stakeholder information needs.

## Conditions to Consider in the Use of Randomized Experimental Designs in Evaluation

Government officials often want to know if a program, intervention, or policy is effective in terms of the impacts it is having. For this, there is considerable agreement that experimental designs with random assignment to treatment and control (or alternative treatment) conditions can make important contributions in addressing the question of program impact. There is also considerable agreement that experimental evaluations can be of less value, or even misleading, in some contexts. It is important, therefore, to understand the emerging consensus on the conditions in which experimental designs are more, and less, appropriate.

First, evaluation methods must be selected on the basis of the primary questions that need to be addressed. The first section below describes the kind of question that a randomized experiment is well suited to answer. This focus on the primary question is followed by sections discussing the contextual issues that need to be considered before deciding to move forward with an experimental evaluation. Some of these issues apply equally to all evaluations of a similar scale, policy area, and complexity; however, situations exist in which issues can be more challenging for an experimental evaluation. Where there are possible challenges, thoughtful consideration is warranted before proceeding.

### *Potential Information Value*

Well-implemented experiments, conducted under proper conditions, are indicated when (A) a primary stakeholder question requires estimating the size of intervention impacts, (B) enough evidence exists to suggest the intervention is promising or effective, but there is not enough conclusive evidence as to make an additional impact evaluation unnecessary, and (C) non-randomized methods that may, in the given context, have other advantages (e.g., less resource intensive or with fewer constraints) are judged to be inadequate for providing credible impact estimates with the needed precision in the particular case.

*A. Estimating Impacts.* Experiments are potentially most valuable when a primary stakeholder question involves the causal impact(s) of a specified policy, program, or other well-defined intervention. That is, "What difference did the implemented program or policy cause in observed outcomes of interest?" (This is in contrast to questions about the causes, or reasons, for an observed state of affairs, such as exploring why an increase in drug abuse occurred. This second type of causal question, exploring the causes of observed effects, is often well addressed with other methods, including qualitative and/or epidemiological investigations.) Estimating the causal impacts of an intervention is important when one needs to choose among discrete alternatives, such as when making funding decisions (i.e., increase, decrease, maintain, or discontinue funding), or when selecting the best intervention (while recognizing that "best" often depends on context, effectiveness, costs, and unintended negative effects). Other causal impact questions that could benefit from experimental designs include assessing the consequences of alternative ways of implementing programs, administering regulations, or enforcing rules or requirements. Because experimental designs are most informative when they compare the effects of a clearly defined intervention, they are less useful for assessing the effects of a federal block grant program, for example, that funds a mix of differing activities in different state or local jurisdictions.

*B. Drawing Conclusions.* The informational value of a randomized study depends on the existing information about the program or policy to be evaluated. There should be enough suggestive evidence of potential effectiveness to establish that the particular intervention is worth evaluating experimentally in order to obtain more conclusive evidence of its impacts, but not so much as to make the new results redundant regarding the nature of policy or program impacts. Relevant evidence can come from many sources, including analysis of program performance data, interviews of pilot project participants, or non-experimental evaluations.

*C. Providing Credible Impact Estimates.* The relative value of information from experiments is greatest when other methods, which may require fewer resources or place fewer constraints on participants, are not adequate for addressing questions about causal impacts. This can occur when (1) the outcomes studied are affected by many influences that cannot be distinguished with non-experimental methods (e.g., many current influences on adolescent at-risk behaviors or many external influences affecting behaviors over time), and (2) the desired statistical precision of, and level of confidence in, the results cannot be achieved with non-experimental methods (e.g., when there are no pre-existing groups that are comparable to the treatment group, or when a small but important impact might be missed).

### *Legal and Ethical Value*

Use of random assignment in evaluations depends on the (A) legal and (B) ethical considerations that determine their appropriateness in specific contexts.

*A. Legal Considerations.* Experimental designs can be considered when laws and regulations permit or require providing access to the treatment or intervention to some eligible persons or groups and not others. These issues are often a concern when individuals or groups are assigned

access, whether randomly or by administrative discretion, to differing programs or policies.

1. Absent specific permission, laws either requiring or prohibiting certain behaviors (e.g., highway speed limits) cannot be selectively applied to some individuals but not others. Thus, an evaluation with random assignment could be used to test alternative ways of enforcing a law or regulation, such as targeting specific locations or time periods for observing drivers' speeds, but not the application of different speed limits to drivers on the same road.
2. Some programs are entitlement programs and so cannot be withheld from those persons determined to be legally entitled to them. Thus, an evaluation could use random assignment to test alternative ways of delivering or enhancing entitlement program benefits, but not the denial of benefits to which persons are legally entitled or the imposition of additional restrictions on their receipt□unless a waiver specifically allows for that purpose. Federal-level waivers have been used, for example, to permit tests of the effectiveness of adding work incentives to welfare programs on an experimental basis. The argument that a waiver is ethically permissible involves the assertion that the experiment offers the potential for improved outcomes.

*B. Ethical Considerations*. Experimental designs can be considered when the experiment meets ethical standards, including (1) respect for persons, (2) social justice, and (3) procedural justice.

1. An experiment needs to maintain respect for participants in the study. Evaluators must abide by current professional ethics, standards, and regulations regarding confidentiality, informed consent, and potential risks or harms to individuals. Denying control group members a treatment known to be effective is a common example of potential harm. Another example of potential harm would be creating conditions that increase risks in order to evaluate prevention programs.
2. Consideration should be given to the possibility that random assignment will raise legitimate concerns about social justice, which focuses on the fairness of the resulting outcomes. On the one hand, rather than using a "first-come, first-served" approach when funding is not adequate to serve all who are eligible, random assignment to program enrollment or control group could well be viewed as more ethical and fairer. On the other hand, when there is tentative evidence that the program is effective, advocates and others often feel that it is unfair that some who are most in need of program benefits are assigned to the no-program control group while others less in need are assigned to the program. Compromises have been developed for this ethical issue, with, for example, random assignment only within a limited range of need. The general point, however, is that concerns about social justice can exist and warrant consideration.
3. Procedural justice concerns the fairness of processes used in the evaluation and is generally supported by ensuring transparency and consistency in study procedures, and by providing for and respecting participants' voiced preferences and needs. First, participation in the experiment is not to be a requirement for receiving normally available government benefits. Also, random assignment must respect cultural values, including the cultural values of some Indigenous peoples who view random assignment as disrespecting individual and community (e.g., Tribal Nation) rights to make choices. Further, it is important that the impartiality and integrity of the assignment process is perceived as fair.

## *Practical Value*

Experiments are most valuable when: (A) the policy, program, or other intervention is ready for evaluation of its effectiveness or value, (B) the resources available for evaluation are adequate, (C) it is feasible to establish and maintain the desired treatment and control conditions, and (D) it is reasonable to expect meaningful use of the experimental findings in the given social and political context.

*A. Readiness*. Programs are generally viewed as ready for an effectiveness evaluation when: (1) there is consensus on program goals and intended activities; (2) the program has had sufficient opportunity to benefit from earlier evaluative feedback and appears to operate effectively with regard to these goals and activities; and (3) the program—even if a pilot program—is relatively stable. These conditions of needed program readiness can apply equally to experimental and non-experimental evaluations. However, caution is warranted as government-sponsored experiments

sometimes assign people to new policies or programs and then proceed to evaluate impacts before the innovative policies or programs can be tried out and modified. Not only would this waste evaluation resources, but it could also lead to the premature rejection of what might have become an effective innovation. On the other hand, decision timelines sometimes necessitate an impact assessment before a program has been refined. Such situations call for thoughtful consideration of the trade- offs involved.

*B. Resources.* All impact evaluations require sufficient resources. These needed resources include (1) sufficient time and availability of people for planning the evaluation; (2) access to relevant quality data and staff with expertise in analyzing the data; and (3) personnel and funding needed for recruiting and assigning participants for the experiment and for monitoring the implementation of the experimental and control conditions. For simpler experiments (e.g., exploring different outreach procedures), the resource requirements may be modest, but larger-scale and more complex evaluations are often more resource intensive and so require more attention to verifying, prior to implementation, the availability of needed resources.

1. Large-scale evaluations of program or policy impacts, whether involving random assignment or not, often require considerable time for planning. Among the issues requiring planning is how to ensure that the intended distinction between the treatment and control conditions is faithfully established. For example, if people are assigned to a new program or policy with novel options, additional outreach may be required to improve understanding of the options. More generally, the groups that should be involved in planning commonly include not only evaluators and program staff and administrators, but also relevant stakeholder groups. Especially when there are many such groups, the time needed for planning may be greater and the skills for eliciting and balancing different groups' interests more important.
2. To be of value, impact evaluations, experimental and otherwise, require access to, and the ability to analyze, the information needed for outcome measures appropriate for assessing program impacts. This information may include existing administrative records as well as new data collected from program participants or staff and, earlier on, pilot testing procedures. This calls for a prior review of data availability and the ability to link the required data sources. With recent advances in conducting RCTs, this also requires staff with sufficient experience to assess the quality and relevance of the data, including whether outcome measures are sensitive enough to detect meaningful changes, as well as staff with the skills to analyze the data. Internal program administrative experiments are often advantaged with regard to data access and analysis.
3. Evaluations that make use of treatment and comparison or control groups to estimate program impacts require sufficient and appropriately justified sample sizes in the groups. While larger sample sizes yield more precision and improved ability to detect impacts, overly large sample sizes waste resources. Relative to alternatives such as comparing intact groups using administrative data (e.g., states with different Medicaid policies), employing random assignment may at times add extra challenges to achieve adequate samples, because volunteers need to be recruited and their informed consent will include consent to be randomly assigned to experimental and control conditions.

*C. Feasibility*. In addition to the resources noted above, experimental evaluations also require the practical ability to establish and maintain the desired treatment and control (or alternative treatment) conditions. This involves the abilities to:

1. Assign study subjects to treatment(s) and control(s) conditions, with procedures that meet the technical requirements of random assignment. Further, the control (or alternative treatment) condition should provide a sufficient level of contrast to the program so as to represent the policy or program alternatives under consideration. When evaluating innovative public policies, which often are the focus of experimental evaluations, establishing the treatment and control conditions can require resource-intensive monitoring to ensure that the conditions are being implemented as intended. For example, if those in the treatment condition are confused about a new policy or new program features, questions exist as to whether the treatment– control contrast was fully established.
2. Maintain the assigned treatment and control conditions, which includes ensuring that: (a) Contamination/diffusion across groups can be kept within acceptable limits, which can

require isolating treatment and control groups and monitoring the groups' treatment-related experiences throughout the study. The risk of contamination can be compounded when people responsible for maintaining the desired treatment–control contrast are committed to providing the highest level of services to all participants. It can also be a problem if treatment group providers "drift" back to the standard control group procedures. (b) Attrition, particularly differential attrition where those leaving the treatment group(s) are different from those leaving the control group(s), can be kept within acceptable limits. This problem can result when people assigned to the control group leave the study to receive the desired treatment services elsewhere.

*D. Usibility*. Conducting evaluations, including experimental ones, is more justified when it is expected that the findings will be used, recognizing that there are many ways of making use of evaluation findings. For example, evaluation findings may be used to recommend changes in program management or design, support budget requests or reallocation of resources, assess previous decisions, expand understanding of the program, or share lessons learned or promising practices with others. Regardless of whether an experimental or another evaluation design is used, several factors need to be considered prior to implementation regarding the likelihood that evaluation findings will find meaningful use.

1. Timeliness. Experimental evaluations are most appropriate when the results are expected to be available in advance of major decisions about the program or policy being evaluated, or in more general debates about future directions. Whenever there is random assignment to new programs, sufficient time should be allowed for the programs to be fully functional prior to the evaluation, and this could affect the timeliness of the findings.
2. Organizational, political, and policy contexts. Whether random assignment is used or not, evaluation findings are of limited value if they enter into organizational, political, or social environments antithetical to their potential, appropriate use. If misuse or nonuse is likely for the kind of findings for which experimental designs are best suited, then evaluation resources may be better allocated to other evaluative studies. Effective use of findings also depends on how the findings fit into the current policy discourse around the policy or program being evaluated.
3. Relevance of the experimental context. A problem can exist if the circumstances in which random assignment is possible differ in important ways from the settings where the findings would be applied. A common example involves attempting to project the effects observed in a study of volunteers to the expected effects when the evaluated intervention or policy is applied to all future program participants. In some cases, there are options for enhancing generalizability by conducting analyses across subtypes of participants, in different geographic localities, or across other variations of interest to stakeholders. More generally, this issue of applicability of findings to intended policy settings needs to be considered.

## *Portfolio Value*

Those responsible for allocating scarce resources for multiple evaluations need to manage a portfolio of studies that make use of various methodologies to best address the information needs of key stakeholders. This involves (A) managing a balance of methodologies appropriate for answering the various questions that need to be addressed and (B) seeking ways to combine various methodologies, both in individual evaluations and across localities or programs, to more efficiently and effectively answer major questions.

*A. Managing Balance of Methodologies*. Given the always limited funding available for evaluation, managers need to balance the need for randomized experimental studies that provide bottom-line estimates of project impacts with the need for other methodologies, often descriptive, to ensure that major stakeholder information needs are met. For example, the use of only experimental methodologies would result in inadequate attention to evaluations of implementation challenges, which are a valuable form of evaluation prior to experimental evaluations. On the other hand, conducting only implementation studies would neglect addressing important impact questions. This potential for an overemphasis on one methodology, with a corresponding reduction in funding for other studies that address different questions, should always be considered with regard to the desired balance within a portfolio of evaluation studies.

*B. Effectively Combining Methodolgoies*. Managing a portfolio of studies with the aim of

optimizing resources also requires considering how funded studies using different methodologies can support and complement each other. For example, an experimental impact evaluation with serious methodological limitations, such as high sample attrition, might be effectively and efficiently complemented with interview-based narrative evidence, in this case from people leaving the study, to help understand whether attrition is a major factor responsible for the observed treatment effects.

### *Overall Value*

The four sets of conditions addressed above should always be considered in judging the value of funding and implementing a randomized experimental evaluation. If one or more of these conditions is questionable for supporting experimental evaluations, there should be added scrutiny and deliberation about the appropriateness of such an evaluation. The deliberation, based on the specific context of the evaluation, would include consideration of how the problematic conditions(s) could be addressed, the value of less-than- perfect impact information, and the adequacy of alternative designs that might be used, including whether an alternative design suffers in terms of the same condition(s). This the type of balancing of the supporting and disqualifying factors in specific contexts is analogous to what effective institutional review boards (IRBs) are supposed to do when reviewing the acceptability of research proposals. Those charged with assessing the appropriateness of funding randomized experimental evaluations should engage in similar, and probably more intensive, due diligence.