REQUEST FOR PROPOSALS:

Building Rigorous Evidence on Effective Student Support Strategies in Online Postsecondary Education

September 1, 2021

I. Overview:

Arnold Ventures (AV) is a nonpartisan philanthropy whose core mission is to invest in evidence-based solutions that maximize opportunity and minimize injustice. This Request for Proposals (RFP) seeks grant applications to build the evidence base on effective student support strategies for students enrolled in online postsecondary programs, either fully-online or in hybrid modalities. As a first step, we are requesting initial letters of interest by October 11, 2021 (please see more details below).

We are committed to improving the return on investment in higher education for students - especially those who have been historically underserved - and for taxpayers. Even as access to higher education has significantly expanded, we still struggle to help students complete their credentials and secure a strong return on their investments. Colleges need sound evidence to identify ways to support students' financial, social, and academic needs. As part of our work in higher education, we support research to uncover and scale up the most effective programs and practices that will pave the way for success among all students.

RFP Objectives

Our ultimate goal in this RFP is to build a base of credible evidence about "what works" for students enrolled in online higher education programs. Online enrollment has steadily grown over the past decade: in fall 2018, 5.7 million students - more than a third of undergraduates - enrolled in at least one online education course, while 2.3 million enrolled exclusively online. If done effectively, online options could expand access for students who require flexible scheduling or lack in-person options, such as adult students and rural students.

But much of the evidence suggests that many students who enroll online are less likely to complete college and earn lower grades, and particularly those students who are traditionally underserved and less academically-prepared.[1] There is also relatively limited evidence on effective supports for online students - such as advising or other wraparound services - to help improve outcomes, even as institutions and other stakeholders have begun to develop promising strategies.

Project Scope

AV has committed **up to $2 million** for research studies that identify and hone promising student success interventions for students enrolled in online education to set the stage for larger-scale, rigorous randomized controlled trial (RCT) evaluations. We are primarily interested in innovative student

[1] See Protopsaltis & Baum (2019), "Does Online Education Live Up to Its Promise? A Look at the Evidence and Implications for Federal Policy," https://www.researchgate.net/publication/330442019_Does_Online_Education_Live_Up_to_Its_Promise_A_Look_at_the_Evidence_and_Implications_for_Federal_Policy.

supports that promote success among online students, such as novel advising or coaching strategies. To be clear, this RFP is intended to support efforts to refine or adapt programs to support students enrolled in online in advance of rigorous evaluation, as opposed to foundational research to study online learning patterns or to design interventions from the ground up.

We are seeking to support the following types of projects through this RFP:[2]

- *Feasibility studies* that set up a prospective evaluation and consider whether it would be productive (e.g., assessing program maturity and variation in implementation, identifying potential sites to participate, refining research methodology)
- *Rapid-cycle evaluations* that refine program models and provide preliminary evidence of impact
- *Well-conducted quasi-experimental design (QED) studies* that can provide a reliable assessment of the program's impact at lower cost before conducting a full-scale RCT[3]
- *Pilot randomized controlled trials (RCTs)* that provide a test run for a fully-powered RCT—for instance, the pilot study might have a lower sample size or only produce short-term results—by laying important groundwork (e.g., testing data collection, assessing program fidelity) and offering preliminary evidence of impact

We will prioritize research that adapts the most effective in-person comprehensive support and advising interventions, as well as particularly promising new practices specific to online education, with a focus on solutions that have the potential to scale up quickly. We will pay particular attention to models that promise to improve outcomes among historically underserved populations across both 2- and 4-year settings, including low-income students, students of color, and adult students. As we discuss in more detail below, we are seeking to prioritize support for research, so are only open to funding a limited proportion of direct program delivery costs.

Commitment to Racial Equity and Diversity

Promoting diversity and equity in research practice is a key facet of AV's mission of maximizing opportunity and minimizing injustice. While racial disparities persist across the higher education system, we recognize that there are too few researchers of color leading studies that are intended to result in successful outcomes for students of color. We are committed to funding work conducted by research teams that are diverse.

In addition, AV's commitment to building the field of scholarship means that applicants should propose teams that include early-career researchers and scholars with meaningful and active roles in the work. AV also equally welcomes new potential partners and existing grantee partners to apply to this RFP.

[2] If applicants believe that the evidence base for a prospective model merits a full-scale RCT evaluation at this stage, please reach out directly to Chase Sackett (csackett@arnoldventures.org), Manager of Higher Education.

[3] For an overview of quasi-experimental design methods that we believe are most likely to produce valid estimates of a program's impact, please see the Appendix, "Which Comparison-Group ("Quasi-Experimental") Study Designs Are Most Likely to Produce Valid Estimates of a Program's Impact?"

II. Selection Criteria

Under the project scope discussed above, AV will consider letters of interest and proposals under four selection criteria:

1. **Prospective impact.** Is the applicant proposing to implement an intervention that:
    - Involves a model with clearly defined program elements (e.g., duration of the treatment, principles for student engagement)? (The proposed project may also clarify or hone the program elements to be included in a prospective full-scale RCT evaluation.)
    - Targets key recognized outcomes of postsecondary success, such as student learning, persistence, completion, time to completion, job placement, and post-college earnings?
    - Is supported by compelling logic, informed by the most relevant and rigorous evidence available, that the intervention could produce sizable impacts on student success? We will prioritize research on practices that have already demonstrated compelling evidence of impact in some format through as rigorous an evaluation as possible. For example, applicants could propose to test an intervention that combines the most promising elements of past effective models; describe the connection between the intervention's elements and top barriers to student success; and/or propose to assess an adaptation of an intervention that improved outcomes in a rigorous experimental evaluation in another setting, such as adapting a 2-year intervention to a 4-year setting.
    - Is likely, or directly designed, to improve outcomes for historically underserved populations, including low-income students, students of color, and adult students?
    - Is cost-effective and could feasibly be scaled to other programs and institutions in the future?
2. **Study design.** Does applicant's proposed study:
    - Explicitly outline core research objectives and lay out a strategy to prospectively, if successful, transition from this project to an RCT (including a statement of provisional approval from key participants)?
    - Consider the delivery of the intervention in a real world context?
    - Clearly document the key elements of the intervention (including the program model, training, supervision, and cost)?
    - Clearly document the predetermined population that will receive the intervention and the setting(s) in which the intervention is delivered as well as the size of the sample?
    - Measure—or demonstrate it is possible to accurately measure—both proximal outcomes to assess whether the intervention affects the behaviors it is intended to change (e.g., credit enrollment and accumulation, program participation), as well as ultimate outcomes (e.g., degree completion)?
3. **Research team.** Does the applicant's team include all parties needed to conduct a feasibility study that can transition to a full-scale RCT, including:
    - An experienced practitioner or institutional leader who has successfully implemented an innovative intervention in a real world setting, with documented adherence to the intervention's key elements?

- One or more researchers who have previously conducted implementation studies and documented key elements of a program model for use in replication studies, and a consultant/advisor who has carried out a well-conducted quasi-experimental design (QED) or RCT and who will help the team to design the feasibility study so it can flow seamlessly into a more rigorous evaluation?
- Personnel familiar with necessary administrative data who can assist researchers in accessing and understanding the data, if such data will ultimately be needed to measure outcomes?

4. **Scope of funding request.** Will the applicant's proposed project:
    - Use the funds requested in the most cost-efficient way to achieve the proposed research objectives? (While we do not have a specific funding amount or maximum amount in mind given project scopes may vary, information about past grants of similar scope funded under the AV higher education portfolio are available on our website.[4])
    - Evaluate an intervention whose delivery is primarily paid for by another funder, and does that funder (and/or other essential parties) agree to the study? (While AV may contribute funds towards intervention delivery, we expect such delivery to be primarily paid for by other funders and will prioritize funding for core research costs. To verify that the necessary funding and other commitments are in place, we will look for attached letters or other communications showing that the funders and/or other necessary parties assent to the study.[5])

III. Timeline and Application Materials

This table includes the full timeline for the RFP and application materials deadlines:

Stage	Date
Deadline for applicants to submit letter of interest (maximum 3 pages)	Deadline: October 11, 2021
AV notifies applicants on a rolling basis whether they are invited to submit a full proposal	By November 17, 2021
Deadline for invited applicants to submit a full proposal	December 20, 2021
AV notifies applicants whether they have been selected for an award	March 2022
Grants are awarded	April - May 2022

Under this timeline, we expect that projects funded under this RFP would start, at earliest, on May 1, 2022, and we are open to start dates through the end of 2022.

[4] Available at: https://www.arnoldventures.org/grants-search?amount=100000%253A499999&topics=Higher%2520Education

[5] Such agreement(s) may be tentative at the time the letter of interest is submitted, but should be finalized before submission of the full proposal. We especially encourage agreements in which the necessary parties not only assent to the study, but also provide a credible description of how they or others would use the study findings to inform program or policy decisions.

Letter of Interest (LOI)

As a first step, we request that applicants submit an initial letter of interest by October 11, 2021. Please submit the letter and any attachments by email to onlineRFP@arnoldventures.org.

Please limit letters of interest to 3 pages or less (not including attachments); applicants may use their own format with single or double spacing, and a font size of 11 or higher. In the letter of interest, please address these core questions:

- Directly address each of the selection criteria above; to be clear, we recognize that applicants may not have fully finalized all aspects of the Study Design criterion at the LOI
- Specify the amount of funding requested, and if feasible, any major cost components; if additional funding from other sources is needed to carry out the study, please describe 1) the total study cost and 2) the portion of that cost to be covered by AV, and please attach a letter or other communication showing that the additional funding will be in place prior to AV's grant award
- Include a brief discussion of the anticipated project timeline; while we do not have a hard requirement for project durations, we expect most projects will be at most 2 years and prospectively shorter in order to transition to a prospective follow-on RCT

And please also:

- Confirm that 1) an appropriate practitioner/institutional leader has agreed to participate in the study, and 2) that a data source has agreed to provide the research team with access to administrative data necessary measure study outcomes (please attach letters or communications reflecting these arrangements)
- Briefly address how the study meets recognized ethical standards for research with human subjects
- Specify the proposed recipient of the grant award, which we generally expect to be a tax-exempt organization (e.g., nonprofit organization, university, or governmental unit); if your organization is not tax-exempt and is interested in applying, please contact Chase Sackett (see contact information below)

Proposal

AV will review letters of interest on a rolling basis and will invite full proposals by November 17, 2021, at the latest. At that time, AV will share feedback to inform the proposal. While we do not have a page length restriction for the full proposal, we strongly prefer concise submissions.

At this stage, we also ask that applicants complete two additional documents (available in advance on request):

- A full project-based budget subject to AV's indirect cost policy (this policy is included in full in the appendix below); we also ask that in the full proposal narrative, you share a preliminary

estimate or prospective range of the anticipated cost for a prospective follow-on RCT evaluation, in the event the initial project is successful

- Complete a research methodology appendix for the project (e.g., describing data sources, research design)

IV. What to Expect in the Grant Agreement

As a condition of the grant award, we ask that grantees:

- Pre-register the study on the Open Science Framework (OSF) website (https://osf.io/) and, prior to the commencement of the study, upload a copy of the research and analysis plan in their proposal. For more details on AV's preferences for pre-registration, please see this overview: https://docs.google.com/document/d/1uA0MOTQdmZvTs8cYy_tnvtIlvUtlWdZ2M0hKnaPP-tU/edit?usp=sharing
- Provide AV with brief updates on the study's progress on a periodic basis, and before making any key decisions that could materially affect the study's design or implementation.
- Submit concise reports on the findings at appropriate intervals. These reports should make it easy for readers to see the study's main results and gauge their credibility.
- Make their datasets and related materials (e.g., survey instruments, code used to clean and analyze datasets) publicly available on the OSF site. Applicants will be asked to do this within one year of the last data collection, and only to the extent allowed under any confidentiality/privacy protections. If anonymization/de-identification would make data sharing possible, please add a brief section in the proposal discussing whether extra time and/or funding is needed.

V. Questions

For any questions, please contact Chase Sackett, AV Manager of Higher Education (csackett@arnoldventures.org).

Arnold Ventures
Indirect Cost Policy
Effective February 1, 2018

Policy Purpose

Arnold Ventures (*"AV"*) requires that any resources awarded by AV to an organization be dedicated to the costs necessary to accomplish the purpose of a grant.

Direct & Indirect Cost Definitions

AV permits grantees to request funding for all of the direct costs associated with a project, including salaries and federally required benefits for employees, travel, meetings and conferences, data access fees, and payments to third-party consultants and sub-grantees that are directly attributable to or created specifically for the purpose supported by a particular grant. Moreover, AV also recognizes that in order to successfully accomplish the purpose of a grant, grantees often need additional financial support to cover a portion of their indirect costs. AV's Indirect Cost Policy (the *"Policy"*) defines indirect costs as organizational costs incurred for a common or joint purpose benefitting more than one project and not exclusively attributable to or created for the project supported by a particular AV grant. Please see Appendix A for examples of indirect costs covered under this Policy.

Allowable Indirect Cost Rates

The Policy permits institutions of higher education, including community colleges, to receive an indirect cost rate of 15 percent (15%) of total direct project costs; all other organizations (e.g., non-profit, governmental, for-profit, etc.) may receive an indirect cost rate of 20 percent (20%) of total direct project costs. [1, 2]

Requirements

For each grant proposal, grantees must provide: (i) a project budget, (ii) a corresponding budget narrative that clearly outlines and defines[3] the total direct project costs, and (iii) fringe rate calculation detail for all personnel allocated to the project within the project budget.

Each new grant request received by AV will be independently reviewed and approved subject to the provisions set forth in this Policy. AV maintains the sole discretion to determine the approved classification of direct and indirect costs for each grant.

Please contact Ruchi Agrawal, AV Finance Manager, at RAgrawal@arnoldventures.org with any questions regarding this Policy.

---

[1] Grants with the primary purpose of providing general operating support are not subject to this Policy.
[2] Third-party consulting or subcontract expenses, sub-awards, data acquisition, and tuition (if applicable) shall not be included as part of the total direct project cost base for the indirect cost calculation.
[3] Please review AV's Budget Template for additional guidelines.

Appendix A
Examples of Indirect Costs

The examples listed in this Appendix A are for general guidance. The list is not exhaustive, and AV, in its sole discretion, will make the final determination on the approved classification of direct and indirect costs for each grant.

Expense Type	Indirect Expense Examples
Personnel	Executive Management (*e.g.*, CEO, COO, CFO, Executive Director, etc.), Central Operational Functions (*e.g.*, Accounting, Administrative Support, Grants Management, HR, IT, Legal, etc.)[4]
Consultants	Contracted work for general operational functions (*e.g.*, legal work or audits)
Travel and Accommodations	Any travel not required to achieve the grant's purpose; accommodation costs over and above the market rate for a specific area
Equipment	Equipment that can be used by an institution for other purposes or projects (*e.g.*, computers, telephones, office furniture)
Rent	Office space rental and utilities associated with central operational functions (*i.e.,* rent expenses incurred whether or not the subject grant is awarded)
Other	All materials and supplies used for more than one purpose or project, printing and postage costs, memberships and subscriptions, hardware and software programs for general operational functions, organizational insurance, etc.

Note: Direct and indirect costs awarded to grantees may only be used for charitable, educational, and/or scientific purposes as such purposes are generally defined by those authorities interpreting the provisions of Section 501(c)(3) of the Internal Revenue Code, and may not be used to carry on propaganda, influence legislation, fund any political campaign, influence the outcome of any election, carry on any voter registration drives, or violate any applicable local, state, federal, or foreign law.

[4] To the extent members of an executive management team are contributing to the project beyond their normal role as an organizational leader, a grantee may request a direct allocation with a corresponding justification explaining the additional contributions of such individuals.

Which Comparison-Group ("Quasi-Experimental") Study Designs Are Most Likely to Produce Valid Estimates of a Program's Impact?:

A Brief Overview and Sample Review Form

Originally published by the Coalition for Evidence-Based Policy with funding support from the William T. Grant Foundation and U.S. Department of Labor, January 2014

Updated by the Arnold Ventures Evidence-Based Policy team, December 2018

We welcome comments and suggestions on this document (jbaron@arnoldventures.org).

Brief Overview:

Which Comparison-Group ("Quasi-Experimental") Studies Are Most Likely to Produce Valid Estimates of a Program's Impact?

I. A number of careful investigations have been carried out to address this question.

Specifically, a number of careful "design-replication" studies have been carried out in education, employment/training, welfare, and other policy areas to examine whether and under what circumstances non-experimental comparison-group methods can replicate the results of well-conducted randomized controlled trials.

These studies test comparison-group methods against randomized methods as follows. For a particular program being evaluated, they first compare program participants' outcomes to those of a randomly-assigned control group, in order to estimate the program's impact in a large, well-implemented randomized design – widely recognized as the most reliable, unbiased method of assessing program impact. The studies then compare *the same program participants* with a comparison group selected through methods other than randomization, in order to estimate the program's impact in a comparison-group design. The studies can thereby determine whether the comparison-group estimates replicate the benchmark estimates from the randomized design.

These design-replication studies have been carried out by a number of leading researchers over the past 20 years, and have tested a diverse range of non-experimental comparison-group designs.

II. Three excellent systematic reviews have been conducted of this design-replication literature; they reached largely similar conclusions, summarized as follows.

The reviews are Bloom, Michalopoulous, and Hill (2005)[1]; Glazerman, Levy, and Myers (2003)[2]; and Cook, Shadish, and Wong (2008)[3]; their main findings include:

A. If the study compares program participants to non-participants who differ markedly in demographics, ability/skills, or other characteristics, it is unlikely to produce valid results. Such studies often produce erroneous conclusions regarding both the size and direction of the program's impact. This is true even when the study tries to equate the two groups using statistical methods such as regression (to adjust for pre-program differences between the two groups) or matching (to identify and compare subsamples of participants and non-participants who have similar characteristics). In other words, if the participants and non-participants differ in key characteristics *before* such statistical methods are applied, applying these methods is unlikely to rescue the study design and generate valid results.

As Cook, Shadish, and Wong (2008) observe, the above finding "indicts much of current causal [evaluation] practice in the social sciences," where studies often use program and comparison groups that have large differences, and researchers put their effort into causal modeling and statistical analyses "that have unclear links to the real world."

B. The comparison-group designs most likely to produce valid results contain all of the following elements:

1. **The program and comparison groups are highly similar in observable pre-program characteristics, including:**

   - **Demographics** (e.g., age, sex, ethnicity, educational attainment, employment status, earnings).

   - **Pre-program measures of the outcome the program seeks to improve.** For example, in an evaluation of a program to prevent recidivism among offenders being released from prison, the offenders in the two groups should be equivalent in their pre-program criminal activity, such as number of arrests, convictions, and severity of offenses.

   - **Geographic location** (e.g., both are from the same area of the same city).

2. **Outcome data are collected in the same way for both groups – e.g., the same survey administered at the same point in time to both groups;**

3. **Program and comparison group members are likely to be similar in motivation.** One type of comparison-group design in which the two groups are likely to have similar motivation is a cutoff-based study, also known as a "regression-discontinuity" study. In such studies, the program group is comprised of persons just above the threshold for program eligibility, and the comparison group is comprised of persons just below (e.g., families earning $19,000 per year versus families earning $21,000, in an employment program whose eligibility cutoff is $20,000). Because program participation is not determined by self-selection, and the two groups are very similar in their eligibility score, there is reason to believe they are also similar in motivation. There are other types of comparison-group designs in which the two groups are likely to have similar motivation.[4]

   However, many comparison-group designs use a program group comprised of persons who volunteer for the program, and a comparison group comprised of non-volunteers. In such studies, the two groups are unlikely to be similar in motivation, as the act of volunteering signals a degree of motivation to improve (which could then lead to superior outcomes for the program group even if the program is ineffective).

4. **Statistical methods, such as matching or regression adjustment, are used to adjust for any minor pre-program differences between the two groups**. Although such methods are highly useful in improving a study's impact estimates, no one method performed consistently better than the others across the various design-replication studies.

C. The three reviews reach varying conclusions about whether comparison-group studies meeting the preferred conditions above can consistently produce valid results, replicating the results of large, well-conducted randomized controlled trials. Consistent with Cook, Shadish, and Wong (2008), we believe additional design-replication studies, testing the most promising comparison-group designs against benchmark randomized controlled trials, are needed to convincingly answer that question.[5] What is clear, however, is that meeting the preferred conditions above greatly increases the study's likelihood of producing valid results.

D. Subsequent design-replication evidence has strengthened the case for cutoff-based comparison-group designs as a valid alternative when a randomized trial is not feasible. Such designs are described above (see B.3). Shadish et. al. (2011)[6], summarizing the most recent design-replication evidence, conclude as follows: "First, the [design-replication evidence] generally supports the hypothesis that the [cutoff-based design] produces similar causal estimates to the randomized experiment across the majority of comparisons attempted. Second, whether tested by statistical or practical significance, a nontrivial percentage of the[se] comparisons did not yield the same results." In other words, the cut-off based designs produced impact estimates that were similar to those of the benchmark randomized controlled trials in most, but not all, cases. Chaplin et. al. (2018)[7] reach similar conclusions.

Shadish et. al. emphasize that this somewhat hopeful conclusion applies to cutoff-based designs that limit their sample to persons just above and just below the cutoff for program eligibility (who are most likely to be equivalent in motivation and other characteristics), rather than including persons well above or below that cutoff. The resulting impact estimates thus apply to sample members near the eligibility cutoff, and may not generalize to those further away.

III. Other factors to consider in assessing whether a comparison-group study will produce valid impact estimates:

A. Preferably, the study chooses the program and comparison groups "prospectively" – i.e., *before* the program is administered.

If the program and comparison groups are chosen by the researcher *after* the program is administered ("retrospectively"), the researcher has an opportunity to choose among numerous possible program and comparison groups. For example, the researcher might select a group of program participants from community A or community B, from years 2007 or 2008, or from age-group 16-20 or 20-24; and might select a comparison group from community A or B or other communities in the county, state, or nation. Each of these choices would likely yield a somewhat different estimate of the program's effect. Thus, a researcher hoping to demonstrate a program's effectiveness can often try many different combinations of program and comparison groups and, consciously or unconsciously, select those that produce the desired result, even in cases where the true program effect is zero. Furthermore, it is generally not possible for the reader of such a study to determine whether the researcher used this approach.

For this and other reasons, retrospective comparison-group studies are regarded by social policy evaluation experts, such as Cook, Shadish, and Wong (2008), and scientific authorities, such as the National Cancer Institute and Food and Drug Administration,[8] as providing less confidence than prospective comparison-group studies and randomized controlled trials (where the composition of the program and control/comparison groups are fixed in advance). Their susceptibility to investigator bias may make them particularly unreliable when the researcher has a financial stake in the outcome.

B. The study follows the same practices that a well-conducted randomized controlled trial follows in order to produce valid results (other than the actual random assignment).

For example, the study should have an adequate sample size, use valid outcome measures, prevent "cross-overs" to or "contamination of" the comparison group, have low sample attrition, use an "intention-to-treat" analysis, and so on.

Appendix:

Sample Form Used to Review a Comparison-Group Study of an Employment and Training Program

Main question to address in your review: Did the study produce scientifically-valid estimates of program impact?

Specific items to be rated by reviewer:

1. **Please assess whether the study produced valid estimates of the program's impact, using the "Brief Overview" document (attached) as a reference.**

   Specifically, please rate the study on a scale of 1 to 5 (5 being strongest, 1 being weakest) on the following categories in the Brief Overview:

	Study Ratings:
The program and comparison groups were highly similar in key pre-program characteristics *before* statistical methods were used to equate the two groups. Please give one composite rating based on items in section 2 of the Brief Overview, including: ▪ Members were selected from the same local labor market. ▪ The two groups were similar in pre-program employment rates and earnings, and demographic characteristics. ▪ Outcome data were collected in the same way, and at the same time, for both groups. ▪ Members of the two groups were likely to be similar in motivation.	
Appropriate statistical methods were used to adjust for any pre-program differences (hopefully minor) between the two groups. Please give one composite rating.	
Preferably, the study chose the program and comparison groups prospectively – i.e., before the program was administered. Please give one composite rating.	
The study had a valid design/implementation in other areas, such as the following. Please give one composite rating. ▪ Adequate sample size ▪ Minimal cross-over, or contamination, between the two groups ▪ Low sample attrition and/or differential attrition ▪ Sample members kept in original group assignment (program or comparison), consistent with intent-to-treat ▪ Valid outcome measures that are of policy or practical importance ▪ Study reports size of effects, and conducts appropriate tests for statistical significance ▪ Study reports effects on all outcomes measured	

   Comment briefly on the reasons behind your ratings.

2. **Based on your ratings and comments above, do you believe this study produced scientifically-valid estimates of program impact?** [Yes / No] Please comment briefly.

References

[1] Howard S. Bloom, Charles Michalopoulos, and Carolyn J. Hill, "Using Experiments to Assess Nonexperimental Comparison-Groups Methods for Measuring Program Effects," in *Learning More From Social Experiments: Evolving Analytic Approaches*, Russell Sage Foundation, 2005, pp. 173-235.

[2] Steve Glazerman, Dan M. Levy, and David Myers, "Nonexperimental Replications of Social Experiments: A Systematic Review," Mathematica Policy Research discussion paper, no. 8813-300, September 2002. The portion of this review addressing labor market interventions is published in "Nonexperimental versus Experimental Estimates of Earnings Impact," *The American Annals of Political and Social Science,* vol. 589, September 2003, pp. 63-93.

[3] Thomas D. Cook, William R. Shadish, and Vivian C. Wong, "Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons," *Journal of Policy Analysis and Management,* vol. 27, no. 4, 2008, pp. 724-50.

[4] An illustrative example is a study of a new teaching method in a college course, in which (i) sections of the course that meet on Monday and Wednesday employ the new teaching method while sections that meet in Tuesday and Thursday employ the old teaching method; (ii) students were unaware which sections would employ which method when they enrolled in the course; and (iii) the study estimates the impact of the new teaching method by comparing end-of-semester exam scores in the Monday-Wednesday sections (i.e., program group) to that in the Tuesday-Thursday sections (i.e., comparison group). Because students did not choose their sections knowing whether the new or old teaching method would be used, there is no reason to believe that students in the program group are any more or less motivated or willing to try new methods than students in the comparison group. A study with a design similar to this is Scott E. Lewis and Jennifer E. Lewis, "Departing from Lectures: An Evaluation of a Peer-Led Guided Inquiry Alternative," *Journal of Chemical Education*, vol. 82, no. 1, January 2005, pp. 135-139.

[5] We also strongly support the recommendation of Cook, Shadish, and Wong (2008) that, in design replication studies, the researchers obtaining the comparison-group estimates be kept unaware ("blinded") as to the benchmark randomized estimates they are seeking to replicate. This would rule out the possibility that they consciously or unconsciously choose parameters for the comparison-group design so as to achieve a successful replication (which would cast doubt on the design's ability to produce valid results in real-world application, where benchmark randomized estimates are unavailable).

[6] William R. Shadish, Rodolfo Galindo, Vivian C. Wong, Peter M. Steiner, and Thomas D. Cook, "A Randomized Experiment Comparing Random and Cutoff-Based Assignment," *Psychological Methods*, vol. 16, no. 2, 2011, pp. 179–191.

[7] Duncan D. Chaplin, Thomas D. Cook, Jelena Zurovac, Jared S. Coopersmith, Mariel M. Finucane, Lauren N. Vollmer, and Rebecca E. Morris, "The Internal and External Validity of the Regression Discontinuity Design: A Meta-analysis of 15 Within-Study Comparisons," *Journal of Policy Analysis and Management*, vol. 37, no. 2, spring 2018, pp. 403-429.

[8] Gary Taubes and Charles C. Mann, "Epidemiology Faces Its Limits," *Science*, vol. 269, issue 5221, July 14, 1995, pp. 164-169. Among other things, this journal article contains a clear description of the issue by Robert Temple, Director of the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration: "The great thing about a [prospective control or comparison-group study] is that, within limits, you don't have to believe anybody or trust anybody. The planning for [the study] is prospective; they've written the protocol before they've done the study, and any deviation that you introduce later is completely visible." By contrast, in a retrospective study, "you always wonder how many ways they cut the data. It's very hard to be reassured, because there are no rules for doing it" (p. 169).