M50
MATHEMATICA
Policy Research

White PAPER

BY JAMES D. RESCHOVSKY AND KATHARINE BRADLEY

Planning Section 1115 Demonstration Implementation to Enable Strong Evaluation Designs

March 2019

This white paper was prepared on behalf of the Centers for Medicare & Medicaid Services (CMS) as part of the national evaluation of Medicaid section 1115 demonstrations (contract number: HHSM-500-2010-00026I). Under the contract, Mathematica Policy Research and IBM Watson Health are conducting an independent national evaluation of the implementation and outcomes of Medicaid section 1115 demonstrations. Mathematica also provides technical assistance focused on states' demonstration evaluation designs and reports. This paper is a guide intended to support states and their evaluators by describing how states can plan the implementation of their demonstrations to enable rigorous evaluations.

CONTENTS

TABLES

FIGURES

INTRODUCTION

This guide describes how states can plan the implementation of their section 1115 Medicaid demonstrations to enable rigorous evaluations. Though the strategies discussed here can apply to a broad range of demonstrations, we focus on evaluating eligibility and coverage (E&C) policies. Recent policies of this type include community engagement requirements, premiums or monthly contributions to beneficiary health accounts, non-eligibility periods as a consequence for noncompliance with program requirements, healthy behavior incentives, and retroactive eligibility waivers.[1] E&C demonstrations often apply more than one of these policies to the target population. They also influence beneficiaries' likelihood of separating from Medicaid. This means E&C evaluations must examine beneficiaries' outcomes after they disenroll or are removed from Medicaid rolls, which in turn creates a need for longitudinal beneficiary surveys. For these reasons, E&C demonstrations are particularly relevant to a discussion of how implementation can strengthen evaluation design.

As part of the special terms and conditions for section 1115 demonstrations, states are typically required to submit an evaluation design to the Centers for Medicare & Medicaid Services (CMS) within 180 days of their demonstration being approved. This time frame may suggest that states should design their evaluations after they implement their demonstrations. Planning implementation to support a strong evaluation design can, however, give states more options for analytical approaches and comparison groups, which improve the quality of evaluation evidence. Proposing strong designs may streamline the process of having evaluation plans approved by CMS. In addition, integrating implementation and evaluation planning may give state Medicaid agencies opportunities to systematically refine demonstration implementation, increasing the chances that states realize their demonstration's goals.

In the four sections that follow, this guide discusses key benefits of coordinating demonstration implementation and evaluation design:

I. Coordination is a prerequisite for using an **experimental evaluation design**, which is the strongest option available.

II. Coordination facilitates **baseline data collection** that supports the evaluation.

III. Coordination permits phased implementation, which can **increase comparison group options** for quasi-experimental evaluation designs.

IV. Coordination enables evaluation designs that **disentangle the effects of specific demonstration features.**

[1] States can use this guide to supplement the evaluation design guidance available for community engagement requirements, premiums, non-eligibility periods, and retroactive eligibility waivers, available at https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html. The guidance provides suggested hypotheses and research questions for these policies and evaluation methods appropriate to address them.

In addition to considering the suggestions in this guide and related policy-specific E&C evaluation design guidance, CMS recommends that states and their independent evaluators refer to complementary technical assistance on comparison group selection, evaluation design, and causal inference in evaluations of section 1115 demonstrations, available at www.Medicaid.gov.[2]

Section 1115 Medicaid Demonstrations

Medicaid is a health insurance program that serves low-income children, adults, individuals with disabilities, and seniors. Medicaid is administered by states and is jointly funded by states and the federal government. Within a framework established by federal statutes, regulations and guidance, states can choose how to design aspects of their Medicaid programs such as benefit packages and provider reimbursement. Although federal guidelines may impose some uniformity across states, federal law also specifically authorizes experimentation by state Medicaid programs through section 1115 of the Social Security Act. Under section 1115 provisions, states may apply for federal permission to implement and test new approaches to administering Medicaid programs that depart from existing federal rules, yet are consistent with the overall goals of the program, likely to meet the objectives of Medicaid, and budget neutral to the federal government.

I. PLANNING IMPLEMENTATION TO ENABLE EXPERIMENTAL EVALUATION DESIGNS

Experimental designs are the gold standard for program evaluation. In experiments—also called randomized controlled trials, or RCTs—individual beneficiaries are randomly assigned to either a treatment group (which is exposed to the demonstration) or a control group (which is not subject to demonstration policies).[3,4] Because states will need to know which beneficiaries should and should not be subject to demonstration policies, states must plan RCTs and conduct random assignment before implementation. States must also consider how to assign beneficiaries who will become eligible for the demonstration after it has started, either because they are newly enrolled in Medicaid or because they transferred to the demonstration's target group from another eligibility group.

[2] See "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations" and "Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations," both available at https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html. CMS is developing separate guidance on beneficiary survey design, which will be made available at the same link in 2019.

[3] In an experiment, the group intentionally withheld from the intervention is typically called the control group, whereas in quasi-experimental evaluation designs, the group not subject to the intervention is referred to as the comparison group. Quasi-experimental designs are observational studies that—without randomization—identify an existing comparison group that is not subject to the intervention but is similar to the treatment group. For both study types, these groups provide the counterfactual against which the treatment group's outcomes are compared.

[4] Random assignment means the use of chance procedures—often computer-generated random numbers—to ensure that each beneficiary will have the same chance of being assigned to the treatment or control group. States can conduct an RCT that includes the demonstration's entire target population or only a portion of it, if that portion is large enough to show program impacts with adequate statistical precision. The state should select the portion to participate in the RCT by using a random sample of the target population and then conduct random assignment to either the treatment or control group. There are many existing resources on RCT procedures and design options for interested states, and this paper therefore provides only a high-level summary. See, for example, Orr (1999) and Boruch (1997).

The main advantage of RCTs is that the results are less prone to bias than the results of quasi-experimental and non-experimental evaluation approaches. Random assignment ensures that treatment and control groups are similar to one another along all dimensions that could affect demonstration outcomes—including those that are not readily observable. This advantage is particularly important for E&C demonstrations, because a beneficiary's response to demonstration policies is very likely to be influenced by unobserved characteristics. For example, community engagement requirements are designed to encourage beneficiaries to seek or retain employment. Some beneficiaries may be inherently more motivated to seek employment than others are, or better equipped with relevant labor market skills that cannot be readily measured. Another advantage of RCTs is that data analysis is more straightforward than the statistical techniques required for many quasi-experimental evaluation approaches.

Experiments are not free from all sources of bias. States should be aware of the possibility of bias from (1) altered behavior on the part of people in the treatment group because they know they are in an experiment (known as the Hawthorne effect), (2) changed behavior by control group members in anticipation of future demonstration coverage, or confusion about whether demonstration policies apply to them (also known as contamination), and (3) differential survey nonresponse patterns for the demonstration treatment and control groups, which is also a concern for quasi-experimental designs that use surveys.[5]

States should weigh the benefits of RCTs against their drawbacks, which include costs. RCTs may be expensive depending on the scope of changes to data systems needed to allocate and track members of the treatment and control groups, as well as administer benefits to and interact with two groups with different messages or program rules. This cost may be at least somewhat offset, however, by the fact that RCTs do not rely on baseline (pre-demonstration) data collection to the same degree that quasi-experimental evaluations do.[6] Efficient data collection is an important consideration for evaluating E&C demonstrations because beneficiary surveys are a key data source, and states may wish to save the cost of a baseline survey.

II. PLANNING IMPLEMENTATION TO OBTAIN BASELINE DATA FOR THE EVALUATION

Most evaluations of section 1115 demonstrations rely on Medicaid administrative data to test the impacts of demonstration policies on beneficiaries' access to care, care quality, health outcomes, or program costs. Administrative data are typically available both before and after implementation and therefore support rigorous quasi-experimental evaluation designs and

[5] See "Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations" for a discussion of other pitfalls to avoid in conducting RCTs, available at https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html.

[6] Although the expectation is that random assignment ensures the treatment and control groups are similar, there is some random chance they will not be—particularly if sample sizes are small. Hence, baseline data can be used to provide statistical adjustment for any remaining differences after random assignment. Baseline surveys can also help identify beneficiary subgroups of interest.

associated statistical techniques, such as difference-in-differences.[7] However, many E&C demonstrations test outcomes that cannot be measured with Medicaid data or other administrative data, or that are expected to happen after beneficiaries are separated from Medicaid. Examples are transitions to commercial health insurance or long-term improvements in health.[8] In these cases, states will need to field surveys of beneficiaries to collect data on their outcomes. To use beneficiary survey data in quasi-experimental evaluations, states and their contracted evaluators should plan to sample and collect data from beneficiaries at demonstration baseline.[9]

Ideally, states should collect baseline data just before they implement the demonstration. However, evaluators responsible for beneficiary surveys may not have enough time to design survey instruments, develop computer code for survey administration (for computer-assisted telephone or online surveys), and train interviewers before the demonstration starts. In these cases, it may be acceptable to conduct a baseline survey after demonstration implementation has started, as long as data collection takes place before demonstration policies have had time to affect beneficiaries' behavior or other outcomes. For example, a state testing the effect of premiums could consider the baseline period to be the months between initial implementation and the distribution of premium invoices to beneficiaries.[10]

III. PLANNING IMPLEMENTATION TO EXPAND OPTIONS FOR QUASI-EXPERIMENTAL DESIGNS

Selecting a valid comparison group is arguably the most critical aspect of planning a quasi-experimental evaluation design.[11] Selecting an in-state comparison group can be a challenge if

[7] Administrative data may not support difference-in-differences analysis for demonstrations that coincide with eligibility expansions, because it is not feasible to collect retrospective data on health care outcomes and costs for individuals newly eligible for Medicaid.

[8] In some states, all-payer claims databases and/or non-Medicaid administrative data may support examination of outcomes for beneficiaries who have separated from Medicaid, and could be available for the period before implementation. The availability and quality of data from all-payer claims databases and the feasibility of collecting non-Medicaid administrative data vary by state.

[9] Although post-implementation surveys could ask for information about beneficiaries during the periods both before and after the demonstration is implemented, responses to retrospective questions about health conditions or activities in earlier time periods are subject to recall bias, such as omission or underreporting of events, or to "telescoping," in which respondents cognitively displace past events, either perceiving recent events as being farther away than they are or remembering distant events as being more recent than they are. CMS is developing separate guidance on designing beneficiary surveys to support rigorous evaluations. This guidance will be made available at the following link in 2019: https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html.

[10] States should, however, consider the possibility that publicity surrounding new demonstrations could influence beneficiary behavior even before specific demonstration policies take effect.

[11] See "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations" (at https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html), for a description of quasi-experimental evaluation designs and a discussion of a broad range of comparison group options.

Medicaid beneficiaries who are not subject to the demonstration differ markedly from the demonstration beneficiaries on characteristics like health status or income.[12] Although states could consider selecting a comparison group of beneficiaries from a different state, they may not want to rely exclusively on this strategy because states differ from each other—for example, in their labor market characteristics, their populations, and their state Medicaid programs. In this section, we describe how states can develop strong in-state comparison groups by staging or rolling out the implementation of their demonstrations to different beneficiary cohorts over time. This supports the use of "stepped wedge" evaluation designs, and can also facilitate the use of regression discontinuity designs. Both designs can generate robust evaluation evidence.

A. Phased implementation by cohort using a stepped wedge design

A state can overcome common problems involved in identifying a good comparison group by staging demonstration implementation so that beneficiaries in cohorts selected for later implementation can serve as a comparison group for beneficiaries selected for earlier implementation. When clusters of beneficiaries are randomized to each cohort, the evaluation design is called a stepped wedge.

Although stepped wedge designs typically contain an element of randomization, they are considered quasi-experimental because they do not randomize individual beneficiaries to treatment and control groups. A state might find this approach preferable to an RCT because stepped wedge cohorts are typically made up of people who fall into naturally occurring clusters—for example, different geographic areas—that are randomly assigned into implementation cohorts. It might be easier for states to implement the demonstration by geographic area instead of implementing it for randomly assigned individuals. This also minimizes the threat of contamination, whereby treatment and comparison group members interact with each other, and members of both groups become confused about whether demonstration requirements apply to them.

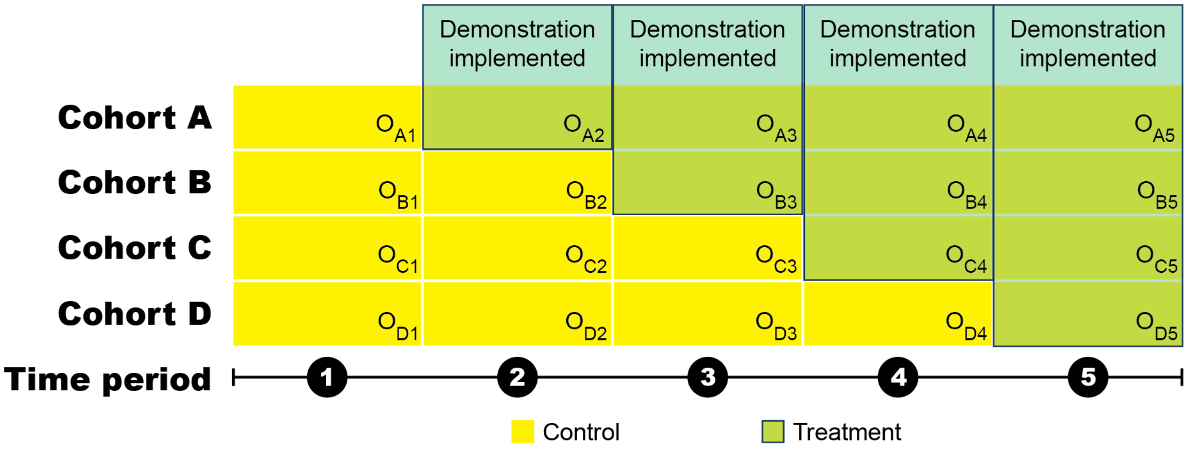
Figure 1 illustrates a stepped wedge design. Demonstration beneficiaries are divided into four cohorts: A, B, C and D.[13] Observations (O) on outcomes and control variables are measured for each cohort five times. Yellow shaded cells indicate time periods during which cohorts are not subject to the demonstration. After the baseline period (Time Period 1), the demonstration is implemented for beneficiaries in Cohort A starting in Time Period 2 (as indicated by green shading of the cell). During Time Period 2, beneficiaries in the other cohorts are available to serve as a comparison group. In Time Period 3, the demonstration is implemented for Cohort B, with beneficiaries in Cohorts C and D available to serve as the comparison group, and so on. In

[12] For example, because income can be considered a proxy for attachment to the labor force, comparing demonstration beneficiaries in the expansion adult group with section 1931 parents and caretaker relatives, who have lower incomes, would be inappropriate for demonstrations designed to encourage more independence through more focus on working. Similarly, demonstration beneficiaries who are exempt from individual E&C policies on the basis of characteristics such as medical frailty are usually not comparable to demonstration beneficiaries who are subject to the E&C policies.

[13] For more information on stepped wedge designs, see Copas et al. (2015); Handley, Schillinger, and Shiboski (2011); and Hussey and Hughes (2007).

Time Period 5, the demonstration is implemented for all beneficiaries. This design gives evaluators the ability to estimate short-, medium-, and longer-term outcomes.[14]

Figure 1. Stepped wedge evaluation design


Demonstration implemented
Demonstration implemented
Demonstration implemented
Demonstration implemented
Cohort A
O_{A1} O_{A2} O_{A3} O_{A4} O_{A5}
Cohort B
O_{B1} O_{B2} O_{B3} O_{B4} O_{B5}
Cohort C
O_{C1} O_{C2} O_{C3} O_{C4} O_{C5}
Cohort D
O_{D1} O_{D2} O_{D3} O_{D4} O_{D5}
Time period
1
2
3
4
5
Control
Treatment

Note: Observations (O) on outcomes and control variables are made for each cohort five times, once at the end of each time period.

Stepped wedge designs have several advantages over other quasi-experimental approaches. First, they give new administrative processes more time to get established, because only a portion of the ultimate target population will be part of the demonstration in the first time period. The logistics of small-scale implementation may be more manageable for states than immediate full-scale implementation. Second, states can use early evaluation results and implementation experiences to iteratively improve the implementation for later cohorts.[15] Third, evaluators can observe and control for secular trends because beneficiaries join the demonstration at different times. For example, the impacts of community engagement requirements might be sensitive to the state of the economy. If unemployment rates change significantly during the course of the evaluation, a phased approach to implementation may allow evaluators to isolate the influence of those changes on program impacts.

Drawbacks to stepped wedge designs include the cost associated with multiple rounds of data collection and the length of time needed to complete the evaluation. States and their evaluators must make a number of interrelated design decisions that involve trade-offs between cost, length of the evaluation period, and statistical power. We next turn to a brief review of the major design decisions.

[14] Using the design shown in Figure 1, for instance, regression models could estimate demonstration impacts after Time Periods 2, 3, and 4.

[15] It is possible that states might want to use early evaluation results to not only alter implementation of the demonstration (for example, by giving beneficiaries more help as they look for jobs), but to amend the demonstration itself. We caution that early results may be too premature to serve as a basis for demonstration amendments, but states and CMS can weigh the available evidence against the potential benefit of amendments and the time needed to approve them.

1. How many cohorts should states create, and how long should the "step length" be?

In general, increasing the number of cohorts will increase the statistical power of the design, but states will need to weigh the advantages of statistical power against higher costs and time constraints. The combination of the number of cohorts and the length of time between exposing each cohort to the demonstration, called the step length, determines the total time between the first and last observation (between Time Periods 1 and 5 in Figure 1). States must ensure that the total evaluation time period is long enough to allow observation of the full impacts of the demonstration and short enough to comply with the state's reporting obligations to CMS. In addition, states must ensure that the step length is long enough to observe meaningful changes in outcomes. For example, a step length of six months might be long enough to observe short-term changes in employment resulting from community engagement requirements, but not long enough to observe transitions to employer-sponsored insurance after gaining employment, because some employers have waiting periods before workers become eligible for coverage. Similarly, six months might be long enough to observe greater use of preventive services in response to healthy behavior incentives, but not long enough to observe health benefits resulting from the change in utilization of preventive care. These timing decisions should be informed by the demonstration's logic model, relevant published research, and the experiences of states that have already implemented similar policies.

2. How often should states collect data, and when?

Figure 1 shows a "closed cohort" stepped wedge design, in which data are collected five times. Five rounds of data collection could get expensive, particularly if surveys are used to collect data. However, there are other "open cohort" stepped wedge designs that require less frequent data collection. For example, states might collect baseline data only during the time period immediately preceding each cohort's implementation, and they could postpone treatment period data collection to measure only longer-term demonstration impacts. These options have implications for the evaluation design's statistical power; evaluators should carefully weigh data collection frequency against sample size requirements.[16]

3. How should states form beneficiary cohorts?

Defining clusters based on geographic area—such as county of residence—may be a logical approach to creating stepped wedge cohorts for section 1115 demonstration evaluations. The number of clusters and the degree to which beneficiaries are similar within clusters have implications for the sample size required to provide adequate statistical power. Even if states randomize geographic areas to implementation cohorts, there will likely be differences in the number of beneficiaries in different areas and in their socio-demographic characteristics. For example, urban and rural areas may have strikingly different economic bases and labor market characteristics. Moreover, local Medicaid offices serving these areas may differ with respect to their capabilities to administer the demonstration, or the availability of services that support compliance with community engagement requirements may vary across areas. Before randomization, evaluators should stratify areas along these or other dimensions to ensure each cohort has a similar set of geographic clusters, and they should also consider using propensity

[16] For more information, see Copas et al. (2015).

score matching or other statistical tools to ensure that that each cohort's sample is as similar as possible to the others.[17,18]

States could also create implementation cohorts by randomly assigning beneficiaries. This option is in essence an experimental design, although rather than randomly assigning beneficiaries to a single treatment and control group, states would randomly assign them to multiple cohort groups. Compared to a conventional experimental design, which uses a single demonstration baseline, a stepped wedge experimental design can better adjust for external events (for example, a recession) that may affect the external validity (generalizability) of evaluation results. The advantages and disadvantages of an experimental design discussed above would apply here as well. States might also consider opportunities for "semi-random" assignment of beneficiaries to cohorts. For example, a state could implement its demonstration sequentially for beneficiaries at their annual eligibility redetermination date, collecting baseline and year-end data on each cohort to estimate the relationship between outcomes and time in the program.

It is not appropriate to use beneficiary characteristics such as age or income to form cohorts in a stepped wedge design. Those characteristics could be related to demonstration outcomes, and that approach would bias evaluation results. However, states can take advantage of beneficiary characteristics to support use of regression discontinuity designs, as described in the next section.

B. Staged implementation to facilitate a regression discontinuity design

Beneficiary populations subject to E&C policies are typically defined by demographic characteristics such as age or income. This raises the possibility of using a regression discontinuity (RD) design, which is a strong quasi-experimental evaluation design. In basic terms, if there is a threshold value that delineates which beneficiaries are subject to the demonstration, those who are not subject to it but close to the threshold can serve as a comparison group. For example, suppose a state receives approval for a community engagement demonstration that applies to individuals ages 19–49 who do not meet various exemption criteria. The lower age bound could not be used as a threshold value in an RD evaluation because children are not expected to be working and independent, but the upper age bound does create an RD opportunity to compare outcomes between those just below (ages 45-49) and above (ages 50-54) the threshold, who may be fairly similar.

In this example, an RD design would not necessarily require the state to coordinate implementation and evaluation. However, there is an important limitation of the RD design that states can overcome with phased implementation: RD designs can only provide estimates of the demonstration's impacts on those beneficiaries who are close to the eligibility threshold value;

[17] States should avoid placing beneficiaries served by better-prepared local Medicaid offices or support services in earlier demonstration cohorts, as this could bias results.

[18] For more in-depth discussion, see Hawkins et al. (2007).

the estimates are not generalizable to beneficiaries with values far from the threshold.[19] In addition, the threshold value must be within the outer bounds of the demonstration's eligibility criteria. Consider a demonstration policy applied to all working-age (that is, ages 19–64) members of a state's Medicaid expansion population. The lower and upper age bounds would not support an RD design, because children and the elderly are not suitable comparison groups. But if a state phased in implementation for three age groups sequentially—for example, those ages 20–34, 35–49, and 50–64—this would create two thresholds, at age 35 and 50, amenable to an RD design. The state should implement the demonstration for new age cohorts only after enough time has passed to assess full program impacts on earlier cohorts.

IV. PLANNING IMPLEMENTATION TO DISENTANGLE THE EFFECTS OF SPECIFIC DEMONSTRATION FEATURES

Section 1115 E&C demonstrations often apply multiple policies to the same beneficiary population. In some cases, these policies are intended to influence the same outcome or set of outcomes. For example, states and CMS expect that both retroactive eligibility waivers and non-eligibility periods will increase the number of beneficiaries who are continuously enrolled in Medicaid. However, state evaluators will not be able to assess which of these demonstration policies is most effective if they are implemented at the same time and applied to the same people. States can strengthen their demonstrations and inform new policy if they understand which policies are and are not effective in achieving desired outcomes, which are most responsible for unintended adverse impacts, and how policies may interact with each other to influence outcomes.

States can disentangle the impacts of multiple demonstration policies by using individual randomization, cluster randomization, or other approaches to construct multiple treatment groups. This allows states to test the impacts of individual policies or combinations of policies, each administered to a different treatment group. Such designs can allow a state to estimate the contribution of alternative policies to a common policy goal (as in the case of retroactive eligibility waivers and non-eligibility periods) as well as whether the combination of approaches is reinforcing. This approach is also useful for testing the effect of setting a single policy at different levels. For example, if a demonstration includes premium requirements, a state could create different treatment groups to assess the effects of different premium amounts.

States can also test the impact of multiple policies that are intended to reinforce each other by designing evaluations of sequentially implemented demonstration policies. A state could design its evaluation to first estimate the effects of an initial demonstration policy and then, after its full effects can be measured, implement an additional demonstration policy (or policies) for the same treatment group in order to assess the marginal effects of these additional requirements on outcomes.

A more systematic approach to the evaluation of demonstrations with multiple policies is to use a factorial design. These designs typically randomize individuals to form treatment groups

[19] The results of an RD evaluation are not generalizable to beneficiaries far from the threshold value when the variable that defines eligibility moderates the demonstration's impact on outcomes. For example, an E&C demonstration may have greater impacts on work effort among younger beneficiaries than it does on older ones.

that are subject to all possible policy combinations, along with a comparison group that is not subject to any of the policies. For example, an E&C demonstration with three policies—community engagement requirements, premiums, and non-eligibility periods—would evaluate eight (2^3) different combinations (Table 1). In this example, evaluators would estimate the impact of community engagement requirements by comparing outcomes for Groups 5–8 versus Groups 1–4, while estimating the effect of premiums by comparing Groups 3, 4, 7, and 8 to Groups 1, 2, 5, and 6. Factorial designs also allow evaluators to explore any interactions between policies. For example, to assess whether the impact of premiums on an outcome varies depending on whether community engagement activities are also required, evaluators would compare outcomes in Groups 7 and 8 with those in Groups 3 and 4.

Table 1. Experimental conditions in the 2^3 factorial design for a hypothetical demonstration with three policies

	Community engagement		Premiums		Non-eligibility periods	
Evaluation group	Yes	No	Yes	No	Yes	No
1		X		X		X
2		X		X	X	
3		X	X			X
4		X	X		X	
5	X			X		X
6	X			X	X	
7	X		X			X
8	X		X		X	

One disadvantage of experiments with multiple treatment groups is that they require a larger sample size than experiments using a single treatment group. Factorial designs require larger total sample sizes than simple experiments with a single treatment and control group do, but may require smaller sample sizes than those needed for a non-factorial experiment with several treatment groups.[20]

CONCLUSIONS

CMS is committed to supporting rigorous evaluations of section 1115 demonstrations because they can inform both federal and state policy and strengthen states' efforts to provide better, more efficient health care for their Medicaid beneficiaries. By planning implementation to support effective evaluation design, states can improve the quality of their evaluations and avoid troublesome limitations like the lack of a strong comparison group strategy. It is therefore important to consider evaluation design options early in the demonstration process, before finalizing implementation plans. Coordinating implementation and evaluation design in this way can help states meet the standards of rigor communicated in newly released guidance on evaluating E&C policies, and may also reduce the time and resources needed to gain CMS approval of evaluation designs.

[20] See Collins et al. (2014) for a discussion of factorial experiments.

REFERENCES

Boruch, Robert F. *Randomized Experiments for Planning and Evaluation: A Practical Guide*. Applied Social Research Methods Series, vol. 44. Thousand Oaks, CA: Sage, 1997.

Contreary, Kara, Katharine Bradley, and Sandra Chao. "Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations." June 2018. Report submitted to the Centers for Medicare & Medicaid Services, June 2018. Available online at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/causal-inference.pdf. Accessed February 12, 2019.

Collins, Linda M., John J. Dziak, Kari C. Kugler, and Jessica B. Trail. "Factorial Experiments: Efficient Tools for Evaluation of Intervention Components." *American Journal of Preventive Medicine,* vol. 47, no. 4, 2014, pp. 498–504.

Copas, Andrew J., James J. Lewis, Jennifer A. Thompson, Calum Davey, Gianluca Baio, and James R. Hargreaves. "Designing a Stepped Wedge Trial: Three Main Designs, Carry-Over Effects and Randomisation Approaches." *Trials,* vol.16, no. 1, 2015, p. 352.

Handley, Margaret A., Dean Schillinger, and Stephen Shiboski. "Quasi-Experimental Designs in Practice-Based Research Settings: Design and Implementation Considerations." *The Journal of the American Board of Family Medicine,* vol. 24, no. 5, 2011, pp. 589–596.

Hawkins, Nathan G., Robert W. Sanson-Fisher, Anthony Shakeshaft, Catherine D'Este, and Lawrence W. Green. "The Multiple Baseline Design for Evaluating Population-Based Research." *American Journal of Preventive Medicine,* vol. 33, no. 2, 2007, pp. 162–168.

Hussey, Michael A., and James P. Hughes. "Design and Analysis of Stepped Wedge Cluster Randomized Trials." *Contemporary Clinical Trials,* vol. 28, no. 2, 2007, pp.182–191.

Orr, Larry. *Social Experiments: Evaluating Public Programs with Experimental Methods*. Thousand Oaks, CA: Sage, 1999.

Reschovsky, James D., Jessica Heeringa, and Maggie Colby. "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations." Report submitted to the Centers for Medicare & Medicaid Services, June 2018. Available at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/comparison-grp-eval-dsgn.pdf. Accessed February 12, 2019.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ SEATTLE, WA ■ TUCSON, AZ ■ WASHINGTON, DC ■ WOODLAWN, MD

M50
MATHEMATICA
Policy Research