Conducting Robust Implementation Research for Section 1115 Demonstration Evaluations

White Paper

October 2020

Laurie Felland and Katharine Bradley

This white paper was prepared on behalf of the Centers for Medicare & Medicaid Services (CMS) as part of the Medicaid 1115 Demonstration Support Contract (contract number: HHSM-500-2014-00034I/75FCMC19F0008). Under the contract, Mathematica provides technical assistance focused on states' section 1115 demonstration evaluation designs and reports. This paper is intended to support states and their evaluators by describing how states can conduct implementation research as part of the overall section 1115 demonstration evaluation.

Contents

I. Introduction

This is a guide to best practices in implementation research and practical considerations for using it to evaluate section 1115 Medicaid demonstrations. Implementation research performs several important functions as part of robust demonstration evaluations. It shows whether the policies envisioned during the demonstration design process are working as expected, reveals the factors that act as facilitators of—and barriers to—implementation, allows policymakers to assess changes to demonstrations over time, and informs improvements to current and future demonstrations.

Implementation research is a standard component of section 1115 demonstrations. Existing evaluation guidance developed by the Centers for Medicare & Medicaid Services (CMS) describes implementation research as one of several necessary evaluation components:

> The principal focus of the evaluation of a section 1115 demonstration should be obtaining and analyzing data on the process (e.g., whether the demonstration is being implemented as intended), outcomes (e.g., whether the demonstration is having the intended effects on the target population), and impacts of the demonstration (e.g., whether the outcomes observed in the targeted population differ from outcomes in similar populations not affected by the demonstration).[1,2]

Together, these components help states fully understand a demonstration's performance and effects, and improve the overall rigor of their evaluations.

States can use this guide to make their implementation research as useful as possible, enhance their overall demonstration evaluation, and prepare informative evaluation reports. In this brief, we describe implementation research in detail and explain how it enhances the overall evaluation and the demonstration itself (Section II), how to design implementation research (Section III), best practices for qualitative data collection (Section IV), quantitative data sources for implementation research (Section V), analyzing and summarizing data (Section VI), and integrating implementation findings with other evaluation findings (Section VII).

---

[1] See "Section 1115 Demonstrations: Developing the Evaluation Design" (available at https://www.medicaid.gov/medicaid/downloads/developing-the-evaluation-design.pdf) and "Section 1115 Demonstrations: Preparing the Evaluation Report" (available at https://www.medicaid.gov/medicaid/downloads/preparing-the-evaluation-report.pdf). Policy-specific evaluation design guidance for substance use disorder demonstrations and serious mental illness/serious emotional disturbance demonstrations is available at https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-state-monitoring-evaluation-resources/index.html.

[2] Program evaluation can employ a variety of methods that have similar, sometimes overlapping terms. Implementation research can include formative assessments (intended to improve the design of a program before it is fully implemented) or process assessments (intended to understand and improve performance). We use the term "implementation research" as a general label that encompasses a broad set of potential measures and data sources. The U.S. Centers for Disease Control and Prevention provides a guide with more information on common terms used to describe components of program evaluations: https://www.cdc.gov/std/Program/pupestd/Types%20of%20Evaluation.pdf.

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.

II. What is implementation research, and how can it be used?

Implementation research is the study of how an intervention—such as a section 1115 demonstration—is put in place, how it proceeds over time, and the reasons why it does or does not proceed as expected. Implementation research acknowledges and accounts for the complexities of the intervention and the context it operates in, with a goal of understanding how context affects an intervention's processes and, ultimately, its success (Peters et al. 2013). Although an intervention's design anticipates certain outcomes, and an impact evaluation assesses in retrospect whether the intervention achieved those outcomes, implementation research intentionally tracks the intervention closer to real time to understand the reasons behind its progress and give the implementers an opportunity to improve it (Werner 2004) (Figure 1).

Figure 1. Using implementation research to track and improve the demonstration

Implementation research can help answer questions such as:

- What is changing on the ground to allow implementation of the intervention? Who is involved, and where?
- Is the intervention being carried out as planned and proceeding as expected? Why? What factors are helping (facilitators), and what factors are posing challenges (barriers)?
- What is responsible for the success of those parts of the intervention that are going particularly well?
- Does the intervention's progress vary depending on where it is taking place and who is affected? Are there differences in context or characteristics that might explain this variation?
- What are potential reasons why the intervention did or did not achieve its goals?
- What else is changing at the same time that could be affecting the intervention or its outcomes?
- What conditions are necessary to sustain the intervention and replicate it somewhere else?

Implementation research on section 1115 demonstrations has several benefits for states:

1. **It gives them prompt feedback on any problems that might be emerging in the demonstration.** Implementation research typically starts early in the overall program evaluation process because it is not necessary to wait for expected outcomes to occur, and because early findings from implementation research can help a state refocus implementation or even modify the demonstration itself. For example, a state can collect data on beneficiaries' understanding of and participation in the demonstration or on provider practice changes in the first demonstration year. Findings about beneficiaries' experiences and provider behavior offer an opportunity for states to reflect on their logic models to identify where a demonstration policy might not be playing out as expected (possibly because of incorrect assumptions about how change happens or how the demonstration can influence change), and where changing the demonstration could help achieve desired outcomes. States should inform CMS about the need for changes to demonstration implementation and describe them in quarterly monitoring reports.
2. **It takes advantage of required monitoring metrics to understand implementation.** Demonstration monitoring metrics are defined by CMS and reported by states for certain demonstration types, including substance use disorder and serious mental illness/serious emotional disturbance demonstrations. Monitoring metrics provide information on demonstration processes and performance, and could shed light on near-term effects (known as proximal effects) that help states understand why the demonstration is or is not achieving its intended outcomes. An example is a state whose demonstration focuses on reducing use of emergency departments and inpatient hospital settings for substance use disorder. This state could use the CMS-defined metric capturing follow-up after emergency department visits for substance abuse as an early indicator of whether the demonstration is likely to achieve its goal. However, monitoring metrics alone are not enough to understand how implementation is proceeding (see Section V for a detailed discussion). States with demonstrations that do not have standardized metrics can develop other measures of demonstration processes, such as the numbers of beneficiaries subject to and actively participating in different stages of an intervention.
3. **It can help states complete timely, informative evaluation reports.** Because implementation research can start relatively early in the demonstration period, findings from implementation research may have a prominent place in interim evaluation reports, which are written before the end of the demonstration period and which form the basis for CMS decisions to award demonstration extensions. Due to data lags or the time it takes some outcomes to occur, interim reports might not include all planned analyses of demonstration outcomes, and they do not include data for the full demonstration period. States can use implementation findings to help ensure that interim reports provide as much information as possible to inform extension decisions.
4. **It strengthens the design and interpretation of outcomes research.** Combining and synthesizing implementation and outcomes research improves overall evaluation rigor. In the case of section 1115 demonstrations, starting implementation research early in the demonstration period can highlight important subgroups or control variables to use in outcomes research. In some cases, early implementation findings might suggest the need for more significant changes to the evaluation design, for example when participation in a given policy is low. Such a finding might suggest that planned analyses of outcomes would result in null findings or suggest ways to refocus evaluation resources on emerging questions. In such cases, states should alert CMS to the need for evaluation design changes and should describe them in quarterly monitoring reports as well as in evaluation reports. In addition, the findings of implementation research provide a context for the outcomes that

are ultimately observed and help the state interpret the meaning of those outcomes. States can conduct this type of synthesis as part of both interim and summative evaluation reports. More details are in Section VII.

5. **It articulates and shares lessons learned to inform changes.** Implementation findings can inspire changes to the state's future demonstrations and may also be useful for informing federal Medicaid policy and policy decisions in other states. Specifically, implementation research can reveal how to replicate changes in new environments by identifying which elements of implementation were critical to the demonstration's success.

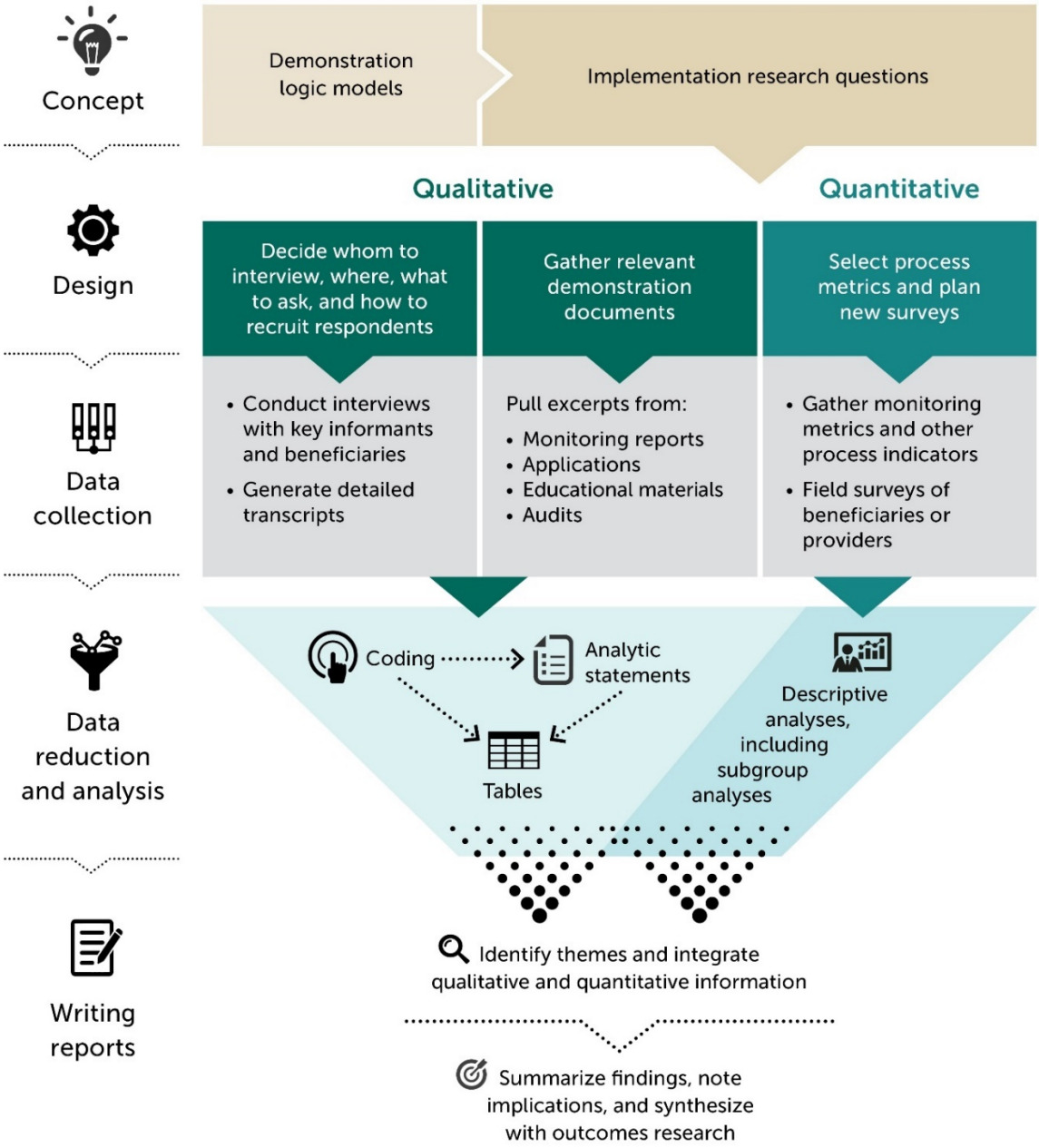
III. Including implementation research in evaluation designs for section 1115 demonstrations

Implementation research is part of the overall section 1115 evaluation, not a separate or additional component. The steps involved in designing and conducting implementation research are shown in Figure 2, and they are all discussed in this guide. The bottom of the figure denotes the synthesis of implementation findings and outcomes research that strengthens the rigor of the overall evaluation.

The first step is to develop research questions. States can use the logic model (or driver diagram) that guides the overall evaluation to plan implementation research.[3] Evaluators should identify (1) how each outcome in the logic model could be affected by how the demonstration is implemented and (2) what steps are necessary for the outcome to be realized. For example, if the demonstration offers beneficiaries a financial incentive to get preventive care (Point A), and the short-term outcome is the likelihood of a beneficiary receiving preventive care (Point B), there are steps between points A and B that would affect that outcome, and those steps should be the focus of the research questions. These might include whether the state educates beneficiaries on the incentive, whether beneficiaries understand the incentive, whether partners such as managed care plans and providers are taking necessary actions to participate in the incentive program, and whether they are doing so consistently.

[3] See "Best Practices in Causal Inference" for a brief discussion of using logic models to design evaluations, available at https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-state-monitoring-evaluation-resources/index.html.

Figure 2. Steps in designing and conducting implementation research



The next step is to determine the best research method(s) to answer each question. Qualitative approaches include interviews with key informants, who have a unique perspective and stake in the demonstration. Examples of key informants are program staff, leaders of managed care plans, and providers. Individual or group interviews with beneficiaries are another source of qualitative data. These approaches can be complemented with reviews of demonstration documents such as beneficiary education materials or provider guidance. Quantitative approaches include surveys of beneficiaries or providers and analysis of monitoring metrics collected as part of the demonstration. Using a variety of approaches and data sources (known as triangulation; see Box 1) helps states develop a complete and robust picture of the demonstration's implementation (Farmer et al. 2006).

Box 1. Triangulation

Triangulation is an important guiding principle of program evaluations. It should be part of each stage of the evaluation, from design and data collection through analysis. Triangulation means drawing from multiple sources and perspectives to create a complete, unbiased, and accurate picture.

In the data collection phase, triangulation involves (1) using both qualitative and quantitative data sources to increase the validity of the findings, and (2) getting the perspectives of different types of respondents to enrich the state's understanding of implementation.

In the analysis phase, triangulation involves considering the number of different respondents and/or types of respondents who made similar comments to determine whether those comments constitute a theme, or whether they do not reflect the views of many people. If information from only a few respondents is important enough to report, evaluators should note that the perspective is limited to a particular type of respondent or circumstance.

Qualitative and quantitative approaches to implementation research have different advantages and disadvantages that complement each other (Table 1). Qualitative approaches enable in-depth exploration of topics and often reveal unexpected or emerging issues and the reasons behind them (Cunningham et al. 2011). They usually sample a small subset of respondents and are less expensive and faster to execute than quantitative approaches like surveys. Quantitative approaches typically involve collecting information from a larger sample of beneficiaries or providers. If the sample is representative of the overall population, it can yield data on prevalence, or how common certain issues or perceptions are. States may find it helpful to link data collection and analysis from qualitative and quantitative sources. For example, findings from interviews can inform survey design or help interpret process measures, and quantitative findings can highlight topics that should be explored in more depth through interviews (Sofaer 1999; Patton 1999).

Table 1. Advantages and disadvantages of qualitative and quantitative approaches to implementation research

Qualitative approaches (particularly interviews)	Quantitative approaches (for example, surveys)
Advantages • Deepen understanding of what is happening and why, and explain complex processes • Hear unexpected but relevant information • Get prompt feedback (faster design, collection, and analysis of data) • Identify questions to ask of broader population through surveys, or patterns to look for in monitoring data • Help interpret and explain quantitative findings	**Advantages** • Get a breadth of perspectives when the population of interest is diverse • Find out how prevalent a characteristic, issue, or viewpoint is • Identify areas to probe in the qualitative research • Place qualitative findings in broader context • State-based surveys can collect data on both implementation and outcomes
Disadvantages • May require intensive per-person recruiting because qualitative interviews usually take more of respondents' time than surveys do • Make it relatively difficult to collect the same type of information from every respondent • Do not generate representative, generalizable data	**Disadvantages** • State-based surveys take longer and are more expensive than qualitative studies • Quantitative surveys and monitoring data do not offer the opportunity to get more explanation of answers or explore topics that weren't thought of ahead of time

IV. Best practices for collecting qualitative data

The main sources of qualitative data for implementation research are the people who are involved in the demonstration or affected by any changes it brings about. Interviewing them can produce important information about the status of implementation and about facilitators (the factors that are helping implementation) and barriers (the things that are making implementation harder). This section describes best practices in collecting qualitative data through interviews and other methods.

A. Selecting respondents to interview

The first step is to decide which kinds of respondents (or which categories of respondents) to interview, and then choosing individuals within each category.

1. Who are the right types of interview respondents?

In deciding whom to interview, states and their evaluators should think about which respondents might have the most important perspectives to help answer the evaluation's research questions. This could include beneficiaries and key informants who have a role in implementing the demonstration or who can share their expertise.

- **Medicaid beneficiaries** can reveal whether the demonstration is reaching the desired population and share their experience with demonstration policies. Beneficiaries can say how aware they are of demonstration policies, how receptive they are to them, and what barriers hinder compliance. Beneficiaries can also describe how they expect the demonstration to affect their access to care and their overall health.
- **People working in agencies and organizations directly involved in carrying out the demonstration** can give important policy and operational perspectives and might have insights to share on variation in demonstration operations across parts of the state and different populations. These respondents might include staff from the Medicaid agency, participating Medicaid managed care organizations, and organizations involved in outreach and advocacy.
- **Providers who treat Medicaid beneficiaries,** such as hospitals, community health centers and other clinics, physician practices, and others, can shed light on changes to health care delivery, adequacy of Medicaid payments, and how beneficiaries access and interact with the health care delivery system.
- **Academics, other researchers, and consultants who are familiar with the demonstration, Medicaid policy, and the general population of beneficiaries** can discuss the demonstration within the state's broader historical and health care context.

2. Which people would best represent the desired types of respondents?

Because it is not feasible to interview everyone who is involved with or potentially affected by a demonstration, creating a sampling plan helps states and their evaluators decide on the right number of respondents and the specific people to seek out to represent each respondent type. In the case of beneficiaries, sampling plans can also stratify the set of potential respondents to ensure that people with different demographic characteristics, geographic locations, and/or exposure to demonstration policies are represented, along with any other subgroups of interest. State Medicaid agencies can provide contact lists of beneficiaries that the evaluator can stratify before randomly selecting potential respondents in each subgroup.

For state Medicaid agency staff, providers, and other key informants, sampling is typically purposive and not random, meaning that evaluators will select respondents based on their particular expertise or role in the demonstration. As they can with beneficiaries, state Medicaid agencies can give lists of providers to evaluators. For other key informants, in addition to ideas from the Medicaid agency, evaluators can review websites and background materials about the demonstration to ensure they include a broad range of organizations and individuals. Supplementing an initial respondent list with respondents' recommendations for additional respondents, or "snowball sampling," can be a useful strategy to fill gaps in the list. Referrals like this can also encourage the referred respondents to participate.

There is no formula that generates the right number of people to seek out for interviews. That number varies depending on the research questions, respondent type, number of respondent types or subgroups, and the evaluation budget. One to two dozen key informant interviews (sometimes fewer) are usually enough. However, if a demonstration focuses on changing providers' behavior, providers are key to answering the research questions, and evaluators might want to aim for two to three dozen provider interviews to get a good range of perspectives.

Evaluations often include a larger number of interviews with beneficiaries than key informants. Because it will be difficult to get everyone to participate, evaluators should start out with a larger sample than the ultimate desired number and should maintain a list of alternates. If it becomes clear that evaluators are reaching saturation in the types of information and perspectives they are hearing before they finish all interviews planned for a respondent type, they could stop recruiting more respondents. "Saturation" is the point at which no new information or perspectives emerge from successive interviews (Guest et al. 2006).

B. Choosing the best forum for interviews

States and their evaluators will also need to decide whether to interview respondents individually or in groups (also called focus groups), and whether to conduct the interviews in person, by telephone, or virtually.

1. When is it better to conduct individual interviews, and when are focus groups or group interviews better?

Factors that drive this decision include the type and sensitivity of the information being sought, the ease of scheduling, and the evaluation budget.

Depending on the number of respondents, **individual interviews** can require more time and resources than group interviews. However, they may be more comfortable for respondents when covering potentially sensitive topics, such as an organization's competitive strategies or a beneficiary's health status or concerns. Individual interviews are easier to schedule and reschedule, which can be particularly important for providers and beneficiaries with unpredictable schedules. Individual interviews are usually best for key informants because learning about their roles in and perspectives on the demonstration may require focused conversations and probes (follow-up questions) that are too time-consuming for a group setting.

Group interviews (or focus groups) can take a lot of time to schedule and organize, but they save time on data collection. Group interviews are also preferable when evaluators are uncertain about each respondent's level of knowledge about and experience with the demonstration (Morgan 1998). For example, when evaluators do not know which beneficiaries have encountered certain demonstration policies or to what degree, a group of beneficiaries can generate discussion on a range of experiences, and

a point made by one beneficiary can help prompt another's understanding or memory or help link different experiences together to build broader themes. There are pitfalls in conducting a group interview, and the facilitator should be prepared for them. Sometimes one or more participants dominate the conversation, or some individuals in the group are uncomfortable speaking up. In addition, it takes skill to assess areas of agreement and disagreement in real time.

2. Should interviews take place in person or remotely?

Interviewing individuals and groups in person is often preferable because meeting in person helps build trust and rapport, which in turn encourages candor. Sitting down with respondents in person also enables evaluators to observe nonverbal cues to help interpret a response, or to understand whether a pause means a respondent is still thinking or is uncomfortable. Being physically present also reduces the potential for evaluators and respondents to multitask or get distracted.

When interviewing key informants, visiting them where they work or in a location convenient to where they live helps convey how important their participation is to the evaluation, which in turn can increase their interest and engagement. **Site visits**, which are a series of in-person interviews in the places where respondents work or live, give interviewers the opportunity to: (1) immerse themselves in the community to see the surroundings in which the demonstration is taking place, and (2) observe firsthand the organizations providing services to Medicaid beneficiaries and the processes they use. These observations enrich interviewers' understanding of respondents' perspectives and can help generate more questions and ideas about factors affecting the demonstration.

Conducting interviews remotely can save travel costs, provide scheduling flexibility, or accommodate travel limitations or safety concerns that might arise during a natural disaster or public health emergency such as the COVID-19 pandemic. Increased availability of user-friendly virtual technology such as WebEx or Zoom can help simulate an in-person interaction. However, technology might not be accessible or familiar to Medicaid beneficiaries, and connectivity and audio issues can make data collection challenging or delay the interviews.

C. Creating discussion guides

A semi-structured discussion guide helps evaluators systematically collect information from a range of respondents in different interviews. "Semi-structured" means that all interviews for a particular category of respondent cover the same topics, but the questions are flexible enough to promote conversation, follow the respondents' train of thought, and explore areas that interviewers had not considered in advance. Some best practices to consider when developing discussion guides include the following. (See also Castillo-Montoya [2016] for a more detailed framework and examples.)

- **Use the research questions as a framework.** Research questions about the demonstration can be a guide to developing interview questions. Writing interview questions that clearly map to a research question focuses the interview on the most important topics and supports analysis later.
- **Use open-ended questions.** In contrast with surveys, which may consist mostly or entirely of questions that invite respondents to choose between pre-set response options, interview questions should be open-ended, meaning that they invite people to respond in their own words. Interviews are an opportunity to not only collect descriptive information but also to get detailed accounts from people about their decision making processes, strategies they have pursued or abandoned, and the

barriers and facilitators they encountered. (Examples of the types of questions an implementation evaluation can address are in Section II.)

- **Avoid leading language.** Questions should be worded objectively and not lead respondents to answer in a particular or socially desirable way, or to try to please the interviewers. For example, neutral language would be, "What do you think of [a given demonstration requirement]?" instead of "Has it been hard to comply with the requirement?" which could make respondents think they should say something about a difficulty even if they have not had one.
- **Avoid "double-barreled" questions.** It is also important to avoid asking about more than one thing with a single question. This can confuse respondents and generate muddled responses that are difficult to attribute to either question. For example, if they are asked about both facilitators and challenges in the same question, respondents could mention factors that are not clearly positive or negative. Similarly, it is a good idea to start with simple questions that establish a point of reference before asking complex questions that could be interpreted or answered in different ways. For example, interviewers could first ask about the current status of a particular issue, and then ask in a new question how it might have changed from a point in the past.
- **Order questions strategically.** The order in which interviewers ask questions can affect the quantity and quality of responses. Starting with more straightforward questions can help respondents warm up to both the interviewer and the topics. Once interviewers have built rapport, respondents will be more comfortable answering challenging or sensitive questions. Leaving more complex questions until later in the interview gives respondents the chance to reflect on what they have said so far and elaborate on or emphasize key points. At the same time, discussion guides should avoid placing high-priority questions last in case interviewers run out of time.
- **Include sub-questions and probes.** These elements of the discussion guide clarify the level and type of details interviewers should pursue and help them guide the respondent to provide information in a logical, gradual sequence. Probes can give interviewers ideas about the types of information they should try to collect and specific examples of how respondents might answer a question, but that the interview need not ask explicitly such as through a sub-question (Box 2). Similarly, adding instructions throughout the guide helps interviewers navigate the questions and manage their time (for example, by indicating how to prioritize questions if time is short).
- **Customize discussion guides by respondent type.** After developing a master discussion guide, the next step is to compose separate versions for each type of respondent. One way to approach this is to first place all questions into rows in a table, arrange the different respondent types as columns, and check off which questions apply to which respondent types. This technique helps evaluators see whether each question is likely to achieve triangulation, whether each type of respondent has a reasonable number of questions, and where additional wording changes might be necessary for a given type of respondent.
- **Test the questions.** It is good practice to pre- test the discussion guide before conducting the interviews. A pre-test can reveal unclear content or wording, confirm whether the number of questions seems reasonable for the allotted interview time, and uncover redundancies or gaps in the questions. Pre-testing is particularly important to ensure that beneficiaries understand questions. Reflecting on how the discussion guide is working after the first few interviews can help determine whether to modify the guide for the remaining interviews.

Box 2. Examples of provider interview questions, sub-questions, and probes designed to address the hypothetical research question "*Will incentive payments increase providers' willingness to treat more Medicaid patients?*"

- What are the main factors affecting your decision to participate in the Medicaid program?
  - What made it easier for you to participate?
  - [Probe on mission, services, need in the community, availability of other providers]
  - What makes it harder for you to participate?
    [Probe on cost relative to payment, challenges meeting particular patient needs]
- Have any of these factors changed since [year intervention started]?
- How, if at all, has the number of Medicaid patients you have treated changed since [year intervention started]?
  - What are the main reasons for this [change/ lack of change]?
    [Probe on changes in capacity, demand, motivation]
- To what extent have the new Medicaid incentive payments affected the number of Medicaid patients you have treated?
  - What are the main reasons for this?
    [Probe on level of incentives received as proportion of payments; administrative issues; patient demand]

D. Getting respondents to participate in interviews

Even though people selected for interviews could have a vested interest in the demonstration, they also have competing demands on their time, and they might be skeptical about the value of speaking with evaluators. Evaluators can use several strategies to encourage participation, but depending on their location and affiliation, they might need Institutional Review Board approval to use them (Box 3). As discussed in a companion paper on conducting beneficiary surveys, sending notification letters that announce the evaluation (possibly in languages other than English), providing nominal financial incentives, and following up after the initial contact are all strategies that encourage responses.

Box 3. Institutional Review Board approval

The U.S. Department of Health and Human Services requires assurance that research involving human subjects is ethical and will not harm the people involved—particularly vulnerable populations (Protection of Human Subjects Regulations in the Code of Federal Regulations (Title 45 Part 46)). Institutional Review Boards (IRBs) led by states, universities, or other entities, assess and approve proposed research activities. For Medicaid evaluations, IRB submissions typically include a description of the proposed research and proposed respondents; a consent form for respondents to sign that explains the research; data collection materials; and, potentially, plans to secure data, and disclosure of any conflicts of interest. See Mathematica [2016] for more details about the IRB process and sample submissions and forms.

These strategies can be good to use in both surveys and qualitative interviews, although there are some differences in applying them to each type of data collection and for different respondent types. For example, incentive payments for qualitative interviews may be larger than those for surveys, because interviews usually take longer, up to an hour.[4]

[4] There is more research on incentive amounts for surveys, which may involve telephone or in-person interviews, than for qualitative interviews. Typical amounts for provider interviews can be up to $75 to $100, although a review by Cho et al. (2013) found that much smaller incentives can encourage providers to respond. Erring on the side of

Likewise, following up on nonresponse can be different for qualitative interviews because samples are smaller, and key informants might not easily substitute for each other. Evaluators could decide to customize outreach to each person, reiterating the importance of their views and reassuring them that the interview can be scheduled when it is convenient for them. Additional strategies for getting people to participate in interviews include:

- **Ask for enough time, but not too much.** The norm is to ask respondents to take 30 to 60 minutes for the interview, depending on the number of topics and questions that have to be covered. More time can allow evaluators to cover more topics in depth, but a shorter interview might encourage participation.
- **Assure respondents of confidentiality.** States and their evaluators should decide on the type and level of interview data that will be shared in reports. If evaluators do not reveal who is interviewed and avoid associating comments with respondents' names and organizations (either verbally in talking with the state or in written materials), respondents may be more likely to agree to participate and be forthcoming. If a respondent asks, "Who else are you interviewing?" evaluators can give them a general sense of the types of people who are being interviewed without violating the other respondents' confidentiality.

E. Conducting the interviews

Qualitative interviewing involves three main activities: asking questions, listening and probing for more information, and documenting responses. Best practices for each of these activities are as follows. (See Weiss [1995] and McGrath et al. [2019] for introductions to successful interviewing and more information on these three activities.)

1. **Asking questions to elicit relevant information for the evaluation**
    - **Build rapport.** Interviewers should (1) express appreciation for respondents' time; (2) briefly introduce the study, reiterating its purpose, how the information will be used, the confidentiality policy, and the main topics to cover; and (3) ask respondents if they have any questions before starting. Interviewers can also build rapport by being friendly and approachable and maintaining regular eye contact if the interview is in person, or by using verbal cues to indicate active listening.
    - **Set a conversational tone and be flexible.** To avoid sounding like they are administering a survey, qualitative interviewers should not read the questions verbatim, instead using their own words and syntax while being careful not to change the meaning or intent of the question. Ideally, when respondents segue to another topic that is still relevant, interviewers will follow the respondents' train of thought. Alternatively, interviewers can stick to the order of the questions in the discussion guide but let respondents know they will return to that topic later.
    - **Know when to skip questions and when to refer back to previous answers.** If respondents have already covered the subject of one question when they answered a different question, the interviewer can decide not to ask the redundant question directly. This can help manage time and

giving beneficiaries more nominal amounts (for example, closer to $20) can prevent the ethical dilemma of whether people are being offered an unfairly large inducement to participate, which makes the invitation too difficult to turn down. The form of the incentive—for example, pre-paid versus post-paid—and offering nonmonetary incentives are also effective in encouraging provider response in certain circumstances (VanGeest et al. 2007).

avoid giving the impression that interviewers are not paying attention. When it seems possible that respondents will have more to say, interviewers can ask the written question with a preface like, "You already touched on this, but I'll ask the question directly to see if you have other thoughts about it."

- **Encourage detailed responses.** When respondents give too-brief answers or seem reluctant to share information, interviewers can try asking them if they could say more, restating the question to make sure they understand it, or coming back to the question later. Sometimes reminding people that there are no right or wrong answers encourages candor. If respondents seem to not know about a given topic, it is best to move on to another topic.
- **Avoid jargon.** When asking questions, interviewers should define any terms that might be new to respondents or that could have more than one meaning. Interviewers also should confirm that they understand the terms respondents use and mirror those terms, which shows respondents that interviewers are listening and helps build rapport.

2. **Listening and probing to collect comprehensive information in limited time**
   - **Interviewers should spend more time listening than talking.** After asking concise questions, interviewers should give respondents time to think and to answer a question before explaining it to them, probing for an answer, and moving on. Sometimes this means allowing longer silent periods than are common in typical social conversation.
   - **Follow the thread of the conversation and make links between points.** Taking brief notes by hand can help interviewers keep track of important issues to return to later in the conversation for more details or clarification. Briefly referencing previous answers can also help respondents complete their thoughts and reassure them that interviewers are engaged and listening.
   - **Clarify information.** When responses are complex or unclear, restating them back to respondents can clarify or confirm their meaning (for example, "Just to make sure I understand, are you saying that you find X difficult to participate in?"). Having such summary statements is also helpful when writing analytic statements during the analysis phase (see Section VI.B).
   - **Remain neutral.** Acknowledging responses with neutral statements such as "I understand," "okay," or "thank you," helps respondents feel heard and can keep the interview moving. These statements are especially important for telephone interviews because visual cues cannot be used. Interviewers should convey understanding but refrain from conveying judgment about the substance of responses. Examples of judgment would be saying, "That's what I was hoping to hear," or "I don't think the state wanted you to do that." Interviewers should also avoid commenting too often or for too long, which could disrupt respondents' own train of thought.
   - **Avoid referring to other respondents' answers.** If respondents say something that seems incorrect or in conflict with what others have said, it may help to probe and ask clarifying questions, but telling them their response conflicts with others can break confidentiality or make respondents feel self-conscious. It can be efficient to ask respondents directly about factual information from other interviews without revealing the precise source of that information. For example, the interviewer might ask, "I've heard that several new community health centers are being built. How do you expect that to affect access to care"?
   - **Manage available time.** When starting interviews, interviewers should check whether respondents are available for the entire time period they were scheduled for, so interviewers can prioritize their questions if they do not have as much time as expected. To make the most of

available time, interviewers can politely interject and redirect respondents when they veer off topic or provide more detail than needed for a question. Respondents do not necessarily know how helpful their responses are, so might appreciate cues about when they have provided enough information on a topic. If time is almost up but there are still important questions to ask, respondents should be asked if they are willing to continue for a few extra minutes.

3. **Documenting responses completely and accurately to analyze later**

- **Generate a transcript of the interview.** The interview responses are the data that will be analyzed later, so it is important to document the conversation as completely and accurately as possible in a transcript. If audio-recording the interview for later transcription, it is best practice to ask respondents for their consent and reiterate the confidentiality policy for the research. Given the potential for technological glitches when recording, either through virtual platforms or with handheld devices, it is important to use a backup recording device. If an interview is not being recorded, a second person should attend the interview to take detailed notes. Interviewers should review the notes and revise them if necessary based on their own handwritten notes, recollections, and understanding of terms that a transcriptionist or second interviewer might not know.

F. Collecting other qualitative information

Demonstration documents can be good sources of qualitative information for the implementation evaluation. These could include annual and quarterly monitoring reports; applications to participate in aspects of the demonstration, such as pilots; beneficiary education materials such as member handbooks or beneficiary account statements; program web sites; managed care contracts; and provider guidance. These documents can shed light on how the state is implementing a demonstration policy and how participants and stakeholders may learn about that policy. Evaluators should take inventory of the available documents and extract excerpts relevant to the research questions. This process may be most helpful if it precedes development of interview guides as document contents could inform interview questions.

When audits are part of the demonstration, evaluators should consider reviewing audit summary materials and extracting relevant information. An example might be a substance use disorder demonstration using audits. As part of the demonstration a state might conduct routine visits to participating providers and review clinical records to determine whether the provider appropriately used assessments and placement criteria, and validate the processes used to credential network providers.[5]

V. Quantitative data sources for implementation research

Quantitative data can reveal how common or typical an observed implementation factor is (its prevalence). Together, qualitative and quantitative data can give a full picture of demonstration implementation. For example, qualitative data might suggest that some beneficiaries (or providers, or state staff) have difficulty with a certain demonstration process. Quantitative data can then show how common that issue is, which can indicate (1) how extensively the issue might affect expected demonstration outcomes and/or (2) how important it would be for the state to modify implementation in response. States can also take these data collection steps in the opposite order—for example, a process

[5] University of Michigan Institute for Healthcare Policy and Innovation (IHPI), "Proposed Evaluation of Michigan 1115 Behavioral Health Demonstration," Substance Use Disorder 1115 Demonstration, September 12, 2019.

measure or demonstration monitoring metric might show an unexpected trend for the overall demonstration, and states can then collect qualitative data to understand the reasons for that trend.

There are four sources of quantitative data that are especially useful for implementation research on section 1115 demonstrations: (1) demonstration monitoring metrics, (2) Medicaid administrative data, (3) policy-specific process indicators, and (4) state-based survey data.

Monitoring metrics for section 1115 demonstrations are now defined for several demonstration types, including substance use disorder demonstrations and demonstrations focused on treating serious mental illness (SMI) or serious emotional disturbance (SED).[6] Most section 1115 monitoring metrics are CMS-defined measures of demonstration processes and performance that can provide early signals of the need to improve implementation or protect beneficiaries. Some are based on established quality measures. Most monitoring metrics are calculated from Medicaid administrative data on a monthly, quarterly, or annual basis, and CMS requires states to report some of them for specific demographic subgroups. Monitoring metrics can support implementation research in several ways:

- Some metrics, such as prescriptions for high-dose opioids for people without cancer in substance use disorder demonstrations, could highlight a trend worthy of investigation and/or provide context that helps the state interpret observed demonstration outcomes.
- Monitoring metrics defined for specific policies can show states how well demonstration processes are working. For example, states with substance use disorder demonstrations can choose to report a metric that reflects whether beneficiaries are assessed for SUD treatment needs using a standardized screening tool. If a state observes that the use of such assessments is lower than expected, or does not increase over time, the state could investigate why, or use the information to interpret demonstration outcomes, or both. Using required monitoring metrics as a source of implementation data is efficient because the same metrics can support both monitoring and evaluation.

  Monitoring metrics are not, however, appropriate measures of demonstration outcomes. States are expected to apply more robust econometric methods to answer research questions about outcomes, such as using comparison groups and statistical techniques that control for observed beneficiary characteristics.
- States can also use monitoring metrics to identify factors that could modify or confound expected outcomes. For example, if a state with a substance use disorder demonstration analyzes a required metric of provider capacity, and sees that capacity is low or decreasing, the metric provides important contextual information for interpreting observed outcomes.

Medicaid administrative data can yield measures of Medicaid program operations and performance that may be useful for understanding demonstration implementation. Administrative data can be used in addition to standardized monitoring metrics or for section 1115 demonstrations that do not have standardized monitoring metrics. States may define many of their own performance measures based on administrative data and can consider which of those measures might be relevant to demonstration implementation. For example, grievances and appeals may be useful for assessing beneficiary acceptance of or problems with section 1115 demonstration policies. As another example, data on provider

[6] See links to lists of monitoring metrics for these demonstration types at this page: https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-state-monitoring-evaluation-resources/index.html.

participation in the demonstration and provider availability to beneficiaries might be useful for understanding the extent to which the policies are affecting access to care.

Policy-specific process indicators for which there are no CMS-defined section 1115 monitoring measures can help document the operational steps necessary to implement a demonstration policy. Some policies require several steps to work in concert if they are going to have their intended effects. For example, if a state implements a copayment for non-emergency use of emergency departments, hospitals must screen patients to determine whether a visit is for an emergency condition, inform patients of their cost-sharing obligation if the visit is for something that is not an emergency, suggest an alternative source of care that the patient can access promptly, and charge the copayment if the patient declines to use the alternative source of care.[7] Consistently implementing these steps across hospitals and across patients—and over time within hospitals—makes it more likely that the policy will have its intended effects. To understand whether hospitals are implementing the policy as intended, a state could define a process indicator reflecting how many copayments are charged by hospitals in the state to patients who are subject to the policy. Quantitative measures that track each operational step would be even more informative, but states must balance reporting burden with the need for data. If the copayment process indicator reveals a problem with implementation, the state could also consider interviewing hospitals to understand implementation at a more granular level.

Surveys of beneficiaries, providers, or other stakeholders in the demonstration are another key source of quantitative data for implementation research. Surveys are especially useful sources of information on stakeholders' understanding of and experience with demonstrations. For example, a beneficiary survey conducted as part of an evaluation of a copayment for nonemergent use of emergency departments could ask beneficiaries whether they are aware of the copayment and were charged the copayment. These are important data points for implementation research because they are necessary conditions for the desired policy effect—beneficiaries must understand the requirement and be able to comply with it for the requirement to work as intended. The same survey could also ask about employment and education outcomes, and the results would inform the state about whether the desired policy effects are taking place.

VI. Analyzing and summarizing data

Taking systematic steps to analyze the qualitative and quantitative data collected as part of the implementation evaluation is important for drawing accurate conclusions and improving the efficiency of the final writing process. This guide focuses on steps for analyzing qualitative data because existing evaluation resources for section 1115 demonstrations do not cover this topic in depth. Methods for qualitative data analysis vary; those described here are recommendations for states and their evaluators to consider, and not the only acceptable approaches. This section also covers ways to summarize and integrate findings from both qualitative and quantitative data to include in interim and summative evaluation reports. For more information on analyzing and summarizing qualitative data, see the seminal work by Miles et al. (2020).

A. Qualitative data coding

Qualitative data are voluminous by nature. To reveal the themes they might contain, it is important to break them down. This process (known as data reduction) involves extracting pertinent pieces of

[7] States' ability to implement cost sharing for non-emergency use of the ED is regulated by 42 CFR § 447.54. The steps listed here summarize the steps outlined in the rule.

information and organizing and grouping them into categories, for example, by general topic, question, or specific topics that arose in the interviews. Coding the data makes this process easier.

Creating a code list. In this context, codes are short labels (no more than a few words) that succinctly capture a topic, attribute or characteristic that is relevant to the demonstration and meaningful to the evaluation. Codes can be generated from the topics covered in the interview guides, based on information that emerges from the interviews, or both. Common types of codes include *descriptive codes*, which describe characteristics of a demonstration or refer to implementation research topics, and *analytic codes,* which reflect evaluators' assessments about what is being said. Examples of a descriptive code might be "completion of health risk assessments" or "providers' use of health risk assessment data." A straightforward way to create descriptive codes is to make a code for each question and sub-question in the discussion guide(s). Helpful analytic codes capture relationships between descriptive codes and barriers and facilitators to implementation. Examples of analytic codes might be "barrier" and "facilitator" or something more specific, such as "many steps to complete" or "supportive leadership." Other helpful codes to consider using are "good quote" and "good example" to find those easily when doing the final write-up.

Applying the codes. Next, evaluators review the interview transcripts (or notes) and apply the relevant codes from the code list to different segments of the transcript. Evaluators could do a second round of coding, applying new codes to reflect notable concepts that emerged during their initial review of the data and that are relevant to the analysis. However, it is good practice to apply no more than a few codes to the same text passage—the more codes a passage has, the more likely evaluators are to see the same information multiple times during analysis, which is bothersome at best and potentially misleading at worst because it could give information more weight than it should have.

Using qualitative software. Coding by hand could be done by applying different colors to the text or using the comment function in MS Word. The other option is to use qualitative software (such as Atlas.ti or NVivo, or a free cloud-based option like Dedoose), particularly if there are a large number of interviews. These software packages help evaluators store detailed qualitative information, code it, sort it, and analyze it. After applying codes, the software can run queries to generate reports on all of the raw data linked to each code.

Improving inter-rater reliability. Deciding which codes to apply to a given text passage is not always straightforward and is subject to differences in interpretation. To improve consistency in coding, it can be helpful to define and describe when to apply each code. If multiple researchers are involved in coding information, it is helpful to have them review each other's work to ensure they all share the same understanding of the codes and are interpreting the passages consistently as they apply the codes.

B. Qualitative data analysis

After data are coded, analysis involves reviewing them, organizing them, and developing common themes and key findings. Recommended steps are as follows.

Developing analytic statements. Analytic statements break qualitative information down even further and help evaluators start to analyze it. Constructing these statements involves reviewing segments of qualitative data that have been assigned a particular code, determining the crux of the response that is relevant to the study, and summarizing it in the evaluators' own words. (See Table 2 for examples of analytic statements for two common codes.) Condensing respondents' verbatim answers or other materials into more succinct passages to clearly state the main points makes the volume of information

more manageable for later organization and review. Although the analytic statements should be succinct, including particularly relevant examples and details reduces the need to return to the original transcripts later during the outlining or writing phase. Analytic statements might also include quotes if they are particularly illustrative and eloquent. Evaluators can also code the analytic summaries to further categorize and organize the data, although this step is not always necessary.

Constructing tables. Creating tables (with qualitative data software or by using word processing or spreadsheet applications) is a helpful way to further organize and analyze qualitative data. For example, evaluators can organize their analytic statements (and other qualitative information from document review) using tables with rows for each respondent. Retaining the source of statements in analytic tables, using file names that include specific information about respondents, puts findings in context and isolates themes by salient respondent characteristics (Ayers 2003). Table columns can reflect individual codes or groups of codes organized by topic areas. Some table cells might have several analytic statements and other cells could have none, making it easy to see what types of responses there were and how common they are. Evaluators might count the number of times a finding appears to inform their understanding of how widely key informants shared a particular perspective, but refrain from including counts of qualitative data in the written findings to avoid conveying false precision or confusing readers about the nature of the data source (qualitative versus quantitative) (see Section VI.C). Evaluators can also use tables to synthesize information for specific units of analysis (for example, by community, provider type, beneficiary type) or for different steps in a demonstration's logic model. See Table IV for an abbreviated sample of how to organize data in this way.

Table 2. Sample analytic statements from providers about the barriers and facilitators involved in providing a service incentivized by the demonstration (imaginary study)

Respondent ID	Facilitators	Barriers
NAME_HOSP_RURAL_REGION1	Three of their providers participated in the training required for providing the service and found it efficient and useful. "My providers were reluctant to attend, but were surprised at how much they learned from it." The incentive payments are valuable to the provider's bottom line because other state subsidies to care for Medicaid and uninsured patients declined over the same period, and the new payments make up about half of that loss.	Would like to provide the service to more patients to generate more incentive payments, but at capacity and cannot find another qualified clinician in their rural region
NAME_PHYSICIAN_URBAN_REGION1	None mentioned	The paperwork required for receiving the incentives was lengthy and not straightforward, which resulted in some back and forth with the state and a six-month delay in receiving the first payments.
NAME_FQHC_URBAN_REGION1	Already provided this service to most of their Medicaid patients on a regular basis, so there wasn't much room for growth. Still, the payments have enabled them to more systematically remind their patients to come in, which they think has contributed to an uptick in volume.	Wish the training were available to non-physician staff, who increasingly help support the provision of the service and would benefit from understanding it more. As a medical assistant said, "Patients ask me for details of what will be involved, and I can't answer all of their questions."

Developing themes. Once qualitative data have been reduced and coded, the next step is to assess and interpret coded data to develop themes—the recurring ideas and concepts emerging from the data (Sandelowski and Leeman 2012). Codes (especially analytic codes and second-level codes) start to identify and document commonalities in the data, but are not themselves themes. Themes are higher-level findings based on an analysis of what the codes convey as a group, the frequency with which certain codes are applied to the data, and the content of coded data. For example, if providers with a new incentive to perform a certain service describe facilitators coded as "large enough payment," "clear and helpful training," and "not a big change," the theme might be: "Providers appreciated the new payments and training, but because many of them already performed most of the required activities, the payments and trainings largely rewarded and reinforced existing practice instead of sparking practice change." Developing such thematic statements can clearly summarize key ideas in a way that preserves and expresses the complexity and nuance behind them. Thematic statements are helpful in the writing phase and can potentially serve as subheads in the evaluation reports (Sandelowski and Leeman 2012).

The tables created in the previous step can support the development of themes. Evaluators can review the tables to look for the range of the types of responses obtained and patterns and topics with frequent or few responses. For example, if only one analytic statement relates to a particular implementation barrier, that statement would not constitute a theme. However, a specific barrier mentioned in a few statements could become an example of a broader group of barriers that constitute a theme. For example, if a few provider

respondents cited slow payment from the state as a reason not to participate in the demonstration, and a few others cited burdensome paperwork requirements, a theme might be that state administrative issues pose a barrier to participation. Evaluators should also determine whether themes apply broadly or to only certain groups or circumstances, and whether themes vary or even conflict across groups (Bradley et al. 2007). Evaluators might choose to end analysis if they reach saturation of themes, meaning the same themes emerge across many interviews so additional analysis is unlikely to change the themes.

C. Summarizing key findings and integrating qualitative and quantitative data

After analyzing qualitative and quantitative data, the next steps are to summarize the main findings and integrate qualitative and quantitative data to tell a comprehensive and compelling story about the demonstration implementation. The following steps can help produce useful findings for state Medicaid agencies. (See also Sandelowski [2012] for more on considerations in writing up findings.)

Connecting the dots and nesting themes. When writing up findings, it is important to link key points to give the findings context for findings and explain how one finding relates to another. One way to do this is to "nest" the themes to convey and highlight the main findings as well as informative variation or nuance for each of them. Nesting means to organize information in layers, typically starting with broader themes and information that applies generally and then sharing more specifics and information that applies more narrowly. For example, nesting can show which findings apply generally across respondent groups, how findings differ by group, where they are most or least present, whether there are outliers, how findings change over time, and so on.

Conveying prevalence. Because qualitative sampling is typically not representative, and respondents are not asked exactly the same questions, qualitative data do not support statements about the prevalence of any given finding. Evaluators should avoid stating specific numbers or fractions when reporting a finding. For example, if 4 in 10 providers who were interviewed mentioned that financial incentives were too small for them to provide a service, but the interviewers did not directly ask all providers whether the incentives were large enough, the evaluator cannot be confident that 40 percent of all providers participating in the demonstration thought the incentives were too small. However, it is still important to give states a sense of the size of the group that each finding is relevant for to help them weigh each finding. Evaluators can define and use terms such as "most" (to denote more than three-fourths of the relevant group); "many" (to denote more than half but fewer than three-fourths of the relevant group); "several" (to denote between one-fourth and one-half of the relevant group); and "few" (less than one-fourth of the relevant group) (Peterson et al. 2020). In this example, the evaluator could say, "Several providers said the incentives were not large enough to encourage them to provide X service."

Providing examples and quotes. Qualitative data by nature are rich in detail, description, and nuance. It is important to leverage this richness in the final write-up to paint a picture of how the demonstration looks on the ground. Examples and quotes can help bring findings to life and make them more accessible to the reader. However, examples and quotes should be used to illustrate findings reported by evaluators, not to replace them. It is good practice to put quotes and examples in the context of the findings by stating whether they convey the sentiment or experience of many respondents, or a few, or even that they are outliers. To respect confidentiality, the write-up should set up quotes by describing the type of respondent (for example, "a practitioner at a community health center"), and not the name of the respondent or organization. Placing quotes in call-out boxes can highlight them, and placing several in a table can show the array and diversity of perspectives on a given theme.

Integrating qualitative and quantitative findings on demonstration implementation. As noted, qualitative and quantitative findings can inform each other to create a fuller picture of demonstration implementation. Returning to the earlier example of copayments for non-emergent use of the emergency department, quantitative data might show that only a small percentage of eligible patients are charged the copayment, and qualitative data analysis could reveal barriers that keep hospitals from charging the copayment. In this example, the evaluator could state:

> According to outcomes measures, only 20 percent of patients using the emergency room for non-emergent conditions were charged a copayment. Interviews with hospitals revealed that a lack of available staff to screen patients and few alternative sources of care in the community were key barriers to charging the copayment. Smaller rural hospitals reported these barriers more than larger urban ones, and the evaluation of demonstration outcomes revealed that the smaller rural hospitals indeed charged copayments only half as often as their larger urban counterparts.

Simple tables that show quantitative data points by hospital type, coupled with a quote or two from hospitals to illustrate the barriers they face, could help elevate the key findings (for examples of ways to display qualitative and quantitative data together, see Guetterman et al. [2015]). At the same time, quantitative and qualitative information might not always align with each other and could even conflict, in which case evaluators should point out the differences and explore potential reasons for them (Patton 1999).

Drawing conclusions and analyzing their implications. After reporting the findings, evaluators should also draw conclusions about what the findings mean for the overall evaluation and the demonstration, clearly indicating where the write-up switches from reporting findings based on the data to discussing conclusions and implications. Findings can give the state ideas for ways to improve the implementation or design of the demonstration. In addition to conveying the findings that clearly emerge from the data, it is also useful to point out important gaps in information, and places where findings are inconclusive.

VII. Integrating implementation and outcomes research

As noted, implementation findings are not only useful for understanding implementation, but also for improving the rigor of outcomes research by suggesting ways to refine the evaluation design and interpret observed outcomes. Synthesizing implementation and outcomes research increases the credibility and usefulness of the evaluation findings for the demonstration state, other states, and CMS in several ways:

1. **Suggests subgroup analyses to perform using the outcomes data.** If interviews, surveys, or monitoring data reveal that certain demographic groups have different perceptions, experiences, or trends than other groups, those findings can prompt states to include subgroup analyses in the outcomes research. Disaggregating quantitative analyses by subgroup could show that demonstration policies have the expected effects for some groups but not others, leading a state to a much different conclusion about its demonstration than it would have reached if it had only examined average effects for the entire population subject to the demonstration.
2. **Finds differences in implementation that lend themselves to indicators in regression analyses, or that suggest ways to refocus evaluation resources.** The method or intensity of demonstration implementation could vary across different areas in a state or across different groups of actors with a role in implementation, such as providers or health plans. For example, implementation research might reveal that some health plans provide more beneficiary education materials or stronger encouragement for particular beneficiary behaviors than other health plans do. As another example,

states could learn that only some hospitals collect copayments for nonemergency use of emergency departments. In such cases, evaluators could develop indicators for high, medium, and low intensity, and could use those indicators in regression models to develop a more precise understanding of linkages between the intervention and observed outcomes. Although this is like the idea of using implementation research to identify subgroups, this idea focuses on the intervention itself and not the people receiving it. If implementation of demonstration policies is very minimal among some health plans or providers, such that planned analyses of outcomes might result in null findings, the state should consider refocusing evaluation resources on emerging questions. In such cases, states should notify CMS and describe changes to the evaluation design in quarterly monitoring reports as well as in evaluation reports.

3. **Identifies additional control variables to include in regression analyses.** Implementation findings can also point to confounding variables or moderating factors that evaluators should include in regression analyses to increase their accuracy. For example, a state conducting implementation research on its SMI/SED demonstration might find that access to needed treatment and the efficacy of screening instruments vary throughout the state. If the state can construct indicator variables that reflect these factors and include them in a regression framework, it could reach a more accurate conclusion about the ability of the demonstration to reach its desired policy goal of reducing preventable readmissions to residential mental health treatment settings. Not including such control variables could lead to omitted variable bias in the regression estimates.[8]
4. **Helps establish or refute causation.** Finally, any analysis of observed outcomes should always incorporate available implementation findings. This is an important part of making causal inferences about demonstration impacts. This is especially important when implementation research shows that a demonstration's implementation did not go as planned, or that moderating or contextual factors were likely to have a substantial effect on demonstration outcomes. To return to the example of copayments for non-emergency use of emergency departments, implementation findings showing that hospitals did not implement the copayment policy should help the state interpret an outcome measure showing that non-emergent visits decreased. In this case, the emergency department copayment could not have caused the observed outcome because it was not implemented, so the state should seek alternative explanations for the observed outcome. In contrast, if most beneficiaries understand a demonstration policy focused on their own behavior, such as a reward for obtaining primary care, that finding can increase confidence that the reward helps to explain an observed increase in primary care use.

VIII. Conclusions

Implementation research is an essential component of evaluating section 1115 Medicaid demonstrations because it helps states and CMS understand whether demonstration policies are working as expected and informs and strengthens research on demonstration outcomes. Because implementation research takes place while the demonstration is underway, it can help track changes over the course of the approval period and reveal possible modifications that could increase the demonstration's chances of achieving its policy goals. In addition, implementation research can generate nuanced information about the factors

[8] See "Best Practices in Causal Inference" for further discussion of the need to control for confounding factors to support causal inference, available at: https://www.medicaid.gov/medicaid/section-1115-demonstrations/1115-demonstration-monitoring-evaluation/1115-demonstration-state-monitoring-evaluation-resources/index.html.

necessary for success. Together, these benefits can provide valuable lessons for a state's future demonstrations, for other states interested in similar policies, and federal Medicaid policy more broadly.

This guide expanded on existing CMS evaluation resources to offer best practices in implementation research using complementary qualitative and quantitative methods. It provided strategies for collecting and analyzing rich, in-depth qualitative information and using it together with quantitative data to develop a comprehensive picture of the demonstration. This guide also suggested ways to synthesize implementation and outcomes research to increase the overall rigor of section 1115 demonstration evaluations. States can use their understanding of these techniques to select qualified evaluators and to collaborate with selected evaluators on activities such as documenting demonstration implementation, developing data collection strategies, and interpreting findings for state and federal stakeholders.

Sources for more information on qualitative research methods

In addition to the papers referenced throughout this guide, states and their evaluators can find more information on qualitative research concepts and methods on university websites. Below are links to each website and short descriptions of the types of relevant and publicly accessible information they feature:

Duke University: Guidance and reports from qualitative conferences and workgroups

Harvard University: Written guidance and slide decks on interviewing skills and how to compare qualitative software platforms; links to webinars on qualitative software platforms (NVivo, Atlas.ti); links to blogs and YouTube videos on conducting virtual interviews

Nova Southeastern University: Comparisons of mobile research applications; links to online research communities and websites; curated list of online text resources (papers, books, presentations) on wide range of qualitative methods topics

University of North Carolina: Links to free trials and online tutorials of qualitative analysis software; links to audio, video and transcription applications for conducting and recording interviews; links to online articles on sampling, data saturation, and writing up qualitative findings

University of Washington: Presentation slides and webinar recordings on topics including an introduction to qualitative research methods, conducting interviews, coding and data analysis

References

Bradley, Elizabeth H., Leslie A. Curry, and Kelly J. Devers. "Qualitative Data Analysis for Health Services Research: Developing Taxonomy, Themes, and Theories." *Health Services Research,* vol. 42, no. 4, August 2007.

Castillo–Montoya, Milagros. "Preparing for Interview Research: The Interview Protocol Refinement Framework." *The Qualitative Report.* vol. 21, no. 5, May 1, 2016.

Cho, Young Ik, Timothy P. Johnson, and Jonathan B. Vangeest. "Enhancing Surveys of Health Care Professionals: A Meta–Analysis of Techniques to Improve Response." *Evaluation and the Health Professions,* vol. 36, no. 3, 2013, pp. 382–407. doi:10.1177/0163278713496425.

Cunningham, Peter J., Laurie E. Felland, Paul B. Ginsburg, and Hoangmai H. Pham. "Qualitative Methods: A Crucial Tool for Understanding Changes in Health Systems and Health Care Delivery." *Medical Care Research and Review,* vol. 68. no. 1, February 2011.

Farmer, Tracy, Kerry Robinson, Susan J. Elliott, and John Eyles. "Developing and Implementing a Triangulation Protocol for Qualitative Health Research." *Qualitative Health Research,* vol. 16, no. 3, March 2006.

Guest, Greg, Bunce, Arwen, and Laura Johnson. "How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability." *Field Methods*, vol. 18(1), 59–82 (2006).

Guetterman, Timothy, Michael Fetters, and John Creswell. "Integrating Quantitative and Qualitative Results in Health Science Mixed Methods Research Through Joint Displays." *Annals of Family Medicine,* vol. 13, no. 6, 2015, pp. 554–561. doi:10.1370/afm.1865.

Mathematica. "Guidance for Institutional Review Board (IRB) Approval of Healthy Marriage and Responsible Fatherhood Grantee Activities." Washington, DC: Mathematica, April 20, 2016.

McGrath, Cormac, Per J. Palmgren, and Matilda Lijedahl. "Twelve Tips for Conducting Qualitative Research Interviews." *Medical Teacher,* vol. 41, no. 9, 2019.

Miles, Matthew B., A.M. Huberman, and Johnny Saldaña. *Qualitative Data Analysis: A Methods Sourcebook.* Fourth edition. Los Angeles, CA: SAGE, 2020.

Morgan David L., *The Focus Group Guidebook.* London: SAGE Publications, 1998.

Patton, Michael Quinn. "Enhancing the Quality and Credibility of Qualitative Analysis." *Health Services Research,* vol. 34, no. 5, part II, December 1999.

Peters, D. H., T. Adam, O. Alonge, I. A. Agyepong, and N. Tran. "Implementation Research: What It Is and How To Do It." *BMJ,* vol. 327, f6753, 2013. doi: 10.1136/bmj.f6753.

Peterson, Dana, Ann O'Malley, Arkadipda Ghosh, et al., "Independent Evaluation of Comprehensive Primary Care Plus (CPC+), Second Annual Report, Supplemental Volume." Princeton, NJ: Mathematica, March 2020.

Sandelowski, Margarete, and Jennifer Leeman. "Writing Usable Qualitative Health Research Findings." *Qualitative Health Research,* vol 22, no. 10, 2012.

Sofaer, Shoshanna. "Qualitative Methods: What Are They and Why Use Them?" *Health Services Research*, vol. 34, no. 5, part 2, 1999, pp.1101–1118.

VanGeest, Jonathan B., Timothy P. Johnson, and Verna L. Welch. "Methodologies for Improving Response Rates in Surveys of Physicians: A Systematic Review." *Evaluation & the Health Professions,* vol. 30, no. 4, pp. 303–321. doi: 10.1177/0163278707307899.

Weiss, Robert S. *Learning from Strangers: The Art and Method of Qualitative Interview Studies.* New York, NY: Free Press, 1995.

Werner, Alan. *A Guide to Implementation Research*. Washington, DC: The Urban Institute Press, 2004.

Appendix: Glossary

Analytic codes are short text labels (no more than a few words) that succinctly describe concepts that emerge from qualitative text (interview responses or program materials). They differ from descriptive codes in that they convey concepts, categories, or interpretation instead of topics. Examples of common analytic codes in implementation research include "barrier" and "facilitator." They are useful in reducing qualitative data to help organize them for further analysis.

Analytic statements are brief passages (no more than a few sentences) that summarize the crux of an interview response or text from other qualitative materials. Analytic statements are helpful in reducing qualitative data to main points, descriptions, examples, and/or brief quotes.

Confounding (or contextual) variables are variables that may influence policy implementation or outcomes and can bias evaluation results if the evaluation approach does not control for them. Beneficiaries' underlying health status is an example of a confounding variable that evaluators should control for in any regression model of the effect of demonstration policies on health outcomes.

Data reduction is the process of extracting pertinent pieces of information from long qualitative text (for example, interview transcripts and program materials) and organizing and grouping it into categories, for example, by topics that arose in interviews. Coding qualitative information, writing analytic statements, and organizing information into tables are helpful data reduction strategies.

Descriptive codes are short text labels (no more than a few words) that succinctly describe an attribute or characteristic relevant to the research topic being studied. They are typically a first line of coding to reflect what is stated directly in the qualitative material (as distinct from analytic codes, which reflect concepts or interpretation). They are useful in reducing data to help organize them and identify key themes.

Double-barreled questions are interview questions that attempt to gather information on multiple concepts in a single question. Their use could result in respondents skipping over an important concept or providing information that is unclear as to which part of the question they answered. Evaluators should avoid using double-barreled questions.

Moderating factors are important preliminary outcomes that evaluations should consider because they affect the relationship between the demonstration policy and one or more hypothesized outcomes. They are not themselves the policy goals. For example, beneficiary (or provider) understanding of demonstration policies is a moderating factor that may affect—and should inform interpretation of—observed demonstration outcomes.

Nesting means to organize information in layers, typically starting with broader themes and information that applies generally and then sharing more specifics and information that applies more narrowly. Nesting is important for conveying the evidence that contributes to a theme, showing how a theme might vary across groups of interest, and making findings more digestible for the reader.

Probes are words or short phrases that are helpful to include in qualitative discussion guides to provide interviewers ideas about the types of information they should try to collect and specific examples of how respondents might answer a question. Probes differ from questions and sub-questions in that interviewers do not ask them directly of all respondents, but could name them if the respondent is not sure of the types of information the question is attempting to illicit.

Purposive sampling in this context involves selecting specific respondents for interviews based on their particular expertise or role in the demonstration. This type of sampling is often used in selecting key informants.

Random sampling refers to the selection of respondents so that any individual within a specific group of interest has an equal chance of being selected. Random sampling is commonly used in selecting respondents when many individuals fit within a given group of interest and the evaluator has no other specific criteria for selecting an individual.

Sampling plans describe the types of respondents necessary to answer research questions, how many of each type to target, the type of sampling used, and any specific criteria for selecting individuals within each type.

Saturation is the point at which no new information or perspectives emerge from gathering more data (for example, conducting additional interviews) or analyzing more data (for example, continuing to analyzing data from the interviews conducted).

Snowball sampling involves asking respondents for ideas for other individuals who would be useful to interview. That is, the respondent sample builds gradually, respondent by respondent. This is a useful technique when identifying key informants for interviews, particularly when the evaluator does not have full information about who is involved in the demonstration and in what ways.

Themes are ideas, concepts, or findings based on analysis of coded data. Although codes are an initial step in identifying themes, themes are higher-level findings based on an analysis of what the codes convey as a group, the frequency with which certain codes are applied to the data, and the content of coded data.

Triangulation is the process of drawing from multiple sources and perspectives in both the data collection and analysis phase of the evaluation to create a complete, unbiased, and accurate picture.

Mathematica

Princeton, NJ • Ann Arbor, MI • Cambridge, MA
Chicago, IL • Oakland, CA • Seattle, WA
Tucson, AZ • Woodlawn, MD • Washington, DC

EDI Global, a Mathematica Company

Bukoba, Tanzania • High Wycombe, United Kingdom

mathematica.org