Mathematica
Progress Together

WHITE PAPER

Beneficiary Survey Design and Administration for Eligibility and Coverage Demonstration Evaluations

June 2019

By Holly Matulewicz, Katharine Bradley, and Sarah Wagner

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 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 best practices in designing and fielding beneficiary surveys.

We thank the state-based 1115 demonstration evaluators who provided comments on this paper: Susan Dorr Goold and Renuka Tipirneni of the Healthy Michigan Plan evaluation team at the University of Michigan; Natoshia Askelson, Elizabeth Momany, and Brooke McInroy of the Iowa Wellness Plan evaluation team at the University of Iowa Public Policy Center; and Erica Dixon and Atheendar Venkataramani of the Kentucky HEALTH evaluation team at the University of Pennsylvania Perelman School of Medicine. Within Mathematica, Eric Grau and Barbara Carlson contributed to sections on sampling, weighting, and power calculations. Maggie Samra and Andy Weiss provided quality assurance reviews. Effie Metropoulos edited the paper, and Stephanie Barna provided secretarial support. We also thank staff within the State Demonstrations Group in the Center for Medicaid and CHIP Services, Centers for Medicare & Medicaid Services, for their support of this work: Danielle Daly, Paula Kazi, Teresa DeCaro, Emerson Carvalho, and Aaron Nadeau.

CONTENTS

I. INTRODUCTION 1

II. SURVEY AND SAMPLE DESIGN 3

A. Data needed for section 1115 evaluations 3

B. Survey design options 5

1. What type of survey design would generate the desired data? 5
2. When and how many times should the survey be fielded? 6
3. How long should the field period be? 7

C. Creating a sampling plan 8

1. How can states and evaluators design a representative sample? 8
2. How big should the sample be, and what is the target response rate? 10
3. What are the survey eligibility criteria? 11
4. How can survey weights increase the representativeness of the sample and adjust for nonresponse? 12
5. What can states do to ensure the sample selection process runs smoothly? 13

D. Importance of mode in collecting survey data 13

1. What are the key considerations when choosing a survey mode? 14
2. How can mixed-mode designs overcome barriers to participation and achieve target response rates? 15
3. How do different modes affect data quality and facilitate quality assurance checks? 17

III. DEVELOPING THE INSTRUMENT 21

A. Moving from research questions to survey questionnaires 21

1. What is the best way to start developing the instrument? 21
2. Which survey topics warrant one survey item, and which should have several? 24
3. What are the advantages to using existing resources in developing survey items? 25
4. What strategies can inform development of new items or response options? 26
5. What are best practices for wording a survey item? 28
6. How should items be modified for different modes? 28

B. Preparing a survey instrument that can collect high quality data 28

1. How should completed and qualified partial cases be defined to facilitate survey administration and data analysis? 29
2. How can surveys collect high quality, reliable data from non-English speakers? 29
3. When should surveys allow proxy interviews? 30
4. How can computer-assisted formats optimize survey instruments? 30

5. Why is it important to test the programmed instrument, and what does testing entail? ... 31

C. Testing survey instruments and procedures ... 32

1. Why is pre-testing important? ... 32
2. What are the logistics involved in a pre-test? ... 33
3. How can evaluators use pre-test findings? ... 33
4. What is a pilot test, and how can it help improve the design of the survey? ... 33

IV. FIELDING THE SURVEY ... 35

A. Importance of sample management ... 35

1. How should evaluators manage the sample from the initial draw to the final wave of data collection? ... 35
2. How does tracking survey response help evaluators measure progress toward survey goals? ... 35
3. How does the tracking mode of completed surveys support sample management and prepare for future waves of the survey? ... 36

B. Survey field periods with more than one release of sample ... 37

C. Strategies to minimize nonresponse and sample attrition ... 38

1. How can evaluators promote participation by sharing information on the survey? ... 38
2. How can branding and consistent messaging encourage responses? ... 39
3. What kinds of incentives work best? ... 39
4. How should evaluators approach nonresponse follow-up? ... 40
5. How can interim contacts keep sample members engaged in longitudinal surveys? ... 41
6. What strategies can states use to locate sample members when data from administrative records are no longer current? ... 41
7. How can states and evaluators support voluntary participation in the survey? ... 42

D. Monitoring the data collection process for quality ... 42

1. How can evaluators monitor for differential nonresponse by subgroup? ... 42
2. How can evaluators monitor the quality of survey responses? ... 43

V. CONCLUSIONS ... 45

REFERENCES ... 47

APPENDIX A: SUMMARY OF SURVEY DESIGNS AND RESPONSE RATES FOR RECENT SURVEYS USED IN SECTION 1115 EVALUATIONS ... 51

APPENDIX B: CONDUCTING POWER CALCULATIONS ... 53

APPENDIX C: EXAMPLE ITEM WORDING FOR MEASURES LISTED IN THE CMS EVALUATION DESIGN GUIDANCE FOR SECTION 1115 ELIGIBILITY AND COVERAGE POLICIES ... 57

TABLES

II.1. Example survey schedule for a longitudinal survey in a five-year demonstration period 7

II.2. Illustration of random sampling strategies on distribution of subgroups in a 10% sample of a population of 2,000 10

II.3. Mode considerations in designing state beneficiary surveys 15

II.4. Potential barriers to survey participation for low-income populations, by mode 16

II.5. Proactively addressing data quality concerns in each mode 18

III.1. Survey domains and topics relevant for evaluations of section 1115 eligibility and coverage demonstrations 23

III.2. Example topic table to guide instrument design 24

III.3. Examples of exploratory interview questions to inform survey development 27

IV.1. Illustration of a 20-week field period with two sample cohorts 37

A.1. Summary of survey design and response rates for recent surveys used in section 1115 evaluations 51

C.1. Example survey items that collect demographic data 58

C.2. Example survey items for evaluations of community engagement requirements 61

C.3. Example survey items for evaluations of premiums and beneficiary account payments 73

C.4. Example survey items for evaluations of non-eligibility periods 75

C.5. Example survey items for evaluations of retroactive eligibility waivers 79

EXHIBITS

II.1. Longitudinal survey data collection with replacement samples 6

This page has been left blank for double-sided copying.

I. INTRODUCTION

This guide is an overview of best practices in the design and administration of beneficiary surveys for use in evaluations of section 1115 Medicaid demonstrations. Beneficiary surveys are particularly important data sources for evaluating demonstrations with eligibility and coverage provisions. Recent section 1115 policies of this type include community engagement requirements, premiums or monthly beneficiary account contributions, non-eligibility periods as a consequence of noncompliance with program requirements, healthy behavior incentives, and waivers of retroactive eligibility.[1] Surveys can yield rich information on beneficiaries' understanding of and experiences with each of these policies. They can also help states assess changes in beneficiary outcomes over time, including for those who disenroll or transition to other coverage.

States can use survey data on beneficiaries' perspectives to inform their interpretation of other evaluation results, and potentially to improve their implementation of the demonstration. For example, if a state finds that some beneficiaries do not understand a demonstration's incentives, this finding might explain why behaviors and outcomes have not changed as much as the demonstration intended them to. As another example, surveys can shed light on whether certain beneficiaries are facing particular barriers to complying with demonstration requirements, and on how the state could alleviate those barriers to increase both compliance and the chances of achieving the demonstration's goals.

Longitudinal information on beneficiaries' outcomes, collected through repeated observations of the same people over time, is valuable because it can reveal how long it takes for desired outcomes to occur and how persistent those outcomes are. Longitudinal data can also help states understand whether immediate outcomes (such as employment) lead to expected long-term outcomes (such as improved health status). In addition, surveys that follow people over time can collect information on outcomes (such as transitioning to commercial health insurance) that are expected to occur after beneficiaries are separated from Medicaid. Some states could use all-payer claims databases or other non-Medicaid administrative data to observe outcomes for former beneficiaries, but the availability and quality of such data vary by state.

As valuable as they are, beneficiary surveys are resource-intensive and require advance planning and skillful deployment to yield reliable evidence. The Centers for Medicare & Medicaid Services (CMS) expects that states that incorporate beneficiary surveys in their evaluations will submit detailed survey plans to CMS as part of their proposed evaluation designs. Survey plans should include the sampling strategy, planned sample size, power calculations, subgroups of interest, frequency and timing of data collection, survey mode (the method of data collection), and survey instruments. Survey plans should also describe how the state will meet challenges such as reaching hard-to-reach populations, achieving high enough response rates, limiting and accounting for nonresponse, and minimizing the number of

[1] States can use this guide to supplement the evaluation design guidance 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 suggests hypotheses and research questions for these policies and describes the evaluation methods that are appropriate to address them.

beneficiaries who drop out of the survey. States and their independent evaluators can use the information in this guide to help them make decisions about each of these aspects of survey design and administration, and can consult the references listed at the end of the guide for more in-depth information. If states' independent evaluators do not have demonstrated experience with the functions outlined in this guide, they should plan to contract with a separate firm that specializes in survey data collection and that can contribute to the survey plan submitted to CMS.

In the following sections, we describe considerations and suggested approaches for designing beneficiary surveys, including developing a sampling plan and selecting survey mode(s) (Section II), developing survey instruments (Section III), and fielding surveys (Section IV). Appendix A has information on designs for recent Medicaid beneficiary surveys conducted for evaluations of section 1115 demonstrations, Appendix B is a brief technical discussion of power calculations, and Appendix C provides a set of suggested survey items that correspond to research questions provided in the CMS evaluation guidance for eligibility and coverage policies. Using or adapting these suggested survey items can cut down on the time and money states would otherwise spend to develop or identify items, and could also make it possible to reliably compare survey data across states.

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. SURVEY AND SAMPLE DESIGN

The survey planning process begins by considering what data are needed for demonstration evaluations (Section A), and what the corresponding survey design options (Section B), sample design (Section C), and mode of data collection (Section D) should be.

A. Data needed for section 1115 evaluations

The first step in the survey planning process is to consider the observations, or data structure, needed to support the planned evaluation design.[2] Depending on the analytic approach states intend to use to answer each research question, surveys may need to collect comparison group observations, observations on surveyed people at different points in time, and/or a series of observations on the same people over time.

Comparison group observations. In general, comparison groups are necessary to learn whether demonstration policies are responsible for observed changes in outcomes. Survey evidence will be strongest with an experimental evaluation design, which randomizes beneficiaries to either a treatment group (which is exposed to the demonstration) or a control group (which is not). States not using randomization can use quasi-experimental designs, which are observational studies that identify a comparison group that is not subject to the demonstration but is similar to the demonstration group.[3] Comparison or control group observations are not necessary to address research questions that focus only on the demonstration group. For example, all states with section 1115 eligibility and coverage demonstrations should assess how well beneficiaries understand demonstration policies, but they do not need comparison group data to answer that research question.

Observations at different time points. If states do not randomize assignment to demonstration and comparison groups, they can survey beneficiaries before and after the demonstration is implemented to support inferences about changes caused by the demonstration. Pre- and post-period observations lend themselves to rigorous modeling approaches like difference-in-differences.[4] Ideally, states that need pre-period observations would survey

[2] CMS has provided other resources that cover evaluation design in detail. See "Selecting the Best Comparison Group and Evaluation Design: A Guidance Document for State Section 1115 Demonstration Evaluations" for a detailed discussion of appropriate evaluation designs based on comparison group strategies (https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/comparison-grp-eval-dsgn.pdf). See "Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations" for more guidance on how states can determine the causality of demonstration policies (https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/causal-inference.pdf).

[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. For both study types, these groups provide the counterfactual against which the treatment group's outcomes are compared.

[4] It is possible to use a difference-in-differences approach with a single set of survey observations in the pre-implementation period, although this requires evaluators to (1) assume there were parallel trends between the comparison and demonstration groups before the demonstration, or (2) verify the parallel trends assumption using a different data source. Methods that require multiple observations in the pre-implementation period, such as interrupted time series models, are not well suited to state-based beneficiary surveys. See "Best Practices in Causal

beneficiaries at the start of implementation or before, also called the "demonstration baseline." States' evaluators might not have enough time to prepare and deploy surveys before the demonstration starts, however. Instead, it may be acceptable to do a baseline survey after implementation has started, as long as data collection takes place before demonstration policies have had enough 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 the start of implementation and the first time beneficiaries receive premium invoices.[5] It is also possible to field a beneficiary survey after implementation and to ask beneficiaries retrospective questions about pre-implementation outcomes, although responses to retrospective questions can be subject to recall bias.

Insights from the field

"Fielding a baseline survey takes a lot of preparation and lead time. We had identified a survey partner well in advance. It turned out to be very useful to have a baseline survey in Kentucky due to implementation delays, but states with less time should keep in mind that some evaluation design choices, like randomization, don't need a full baseline survey. Survey planning interacts with other aspects of evaluation design."

– University of Pennsylvania evaluation team for Kentucky HEALTH

States that are unable to gather baseline data, but do have policies that differ on either side of a threshold value, such as beneficiary age or income, could have the option of regression discontinuity analysis, another methodologically rigorous approach. Beneficiaries exposed to demonstration policies on one side of the threshold can be compared to beneficiaries on the other side of the threshold who are not subject to the policies.[6] States using regression discontinuity designs should plan to oversample beneficiaries who are close to the threshold between the demonstration and comparison groups.[7] Randomizing assignment to the demonstration or a control group is another approach that does not require a baseline to support causal inference (that is, to support a conclusion that the demonstration is responsible for an observed outcome).

Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations" for further discussion of the kinds of observations needed for different modeling approaches; available at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/causal-inference.pdf.

[5] As noted in a companion guide, "Planning Section 1115 Demonstration Implementation to Enable Strong Evaluation Designs" (available at https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html), states should also consider the possibility that the publicity surrounding new demonstrations could influence beneficiaries' behavior even before specific demonstration policies take effect.

[6] States can create in-state comparison groups by staging the implementation of their demonstrations so that those in different age or income categories are exposed to demonstration policies at different points in time. This strategy supports the use of regression discontinuity designs if implementation is paced to allow enough time to observe expected outcomes for each group before rollout to the next group. See "Planning Section 1115 Demonstration Implementation to Enable Strong Evaluation Designs" (available at https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html) for further discussion.

[7] For example, comparing beneficiaries with incomes from 90 to 100 percent of the federal poverty level (FPL) to beneficiaries with incomes from 100 to 110 percent FPL would allow evaluators to assess the effects of a policy on one of these groups, which are otherwise similar. Increasing the income range to compare beneficiaries with incomes from 60 to 100 percent FPL to those with incomes from 100 to 140 percent FPL would increase the sample size, but calls the similarity of the two groups into question. See "Best Practices in Causal Inference for Evaluations of Section 1115 Eligibility and Coverage Demonstrations" at https://www.medicaid.gov/medicaid/section-1115-demo/downloads/evaluation-reports/causal-inference.pdf for further discussion.

A series of observations on the same people over time. As noted, some section 1115 eligibility and coverage policies could bring about changes that take a long time to emerge, that are logically dependent on earlier outcomes, or that take place after beneficiaries are separated from Medicaid. The timing of expected outcomes, and the need to show whether they happen in a particular order, determine whether states should follow surveyed individuals over time. Multiple survey observations for the same people, potentially for several years after they separate from Medicaid, are also necessary if the state wants to know whether desired demonstration outcomes—like enrollment in commercial coverage—persist over time.

B. Survey design options

Next, depending on their data and analysis requirements, states can choose a survey structure from several basic options. Examples include cross-sectional surveys at a single point in time, repeated cross-sectional surveys, and longitudinal surveys. States that choose repeated cross-sectional or longitudinal designs must also decide how many times to field the survey and at what interval. For any survey design, states need to decide when to collect data and how long the field period should be. See Appendix A for a summary of the cross-sectional and longitudinal designs used in several recent surveys of section 1115 demonstration beneficiaries.

1. What type of survey design would generate the desired data?

Single cross-sectional surveys sample one or more groups at a single point in time. States have used cross-sectional designs to compare beneficiaries who receive different benefits, or to compare current, former, and never-enrolled beneficiaries. This approach is the least expensive, but it does not provide data on changes over time.

Repeated cross-sections sample different beneficiaries at different points in time. States can use this design to compare group-level outcomes before and after the demonstration is implemented or to understand how a demonstration's effects change over time. This approach does not follow the same people over time, and therefore does not reveal how outcomes change for those people at the individual level. However, repeated cross-sections are appropriate for difference-in-differences analyses, in which the state samples demonstration and comparison groups before and after the demonstration is implemented. Repeated cross-sections also require a smaller sample size than longitudinal surveys.

Longitudinal or panel surveys ask questions of the same beneficiaries at different points in time.[8] This design can help states determine whether intermediate demonstration outcomes (like employment) lead to longer-term outcomes (like enrollment in employer-sponsored insurance) for people in the demonstration. However, because they follow the same people over time, longitudinal surveys are vulnerable to attrition, or a drop in the number of completed interviews that can happen as some people fail to respond to successive waves of the survey. Attrition generally increases over time. This is a particular concern for surveys of Medicaid beneficiaries, because it can already be difficult and resource-intensive to reach them and achieve high response rates. For example, many Medicaid beneficiaries move often, have their phone or Internet service interrupted, or change their telephone number, making it difficult to keep them

[8] A panel is a group of people surveyed at multiple time points. A longitudinal survey can include more than one panel. We use these terms interchangeably.

engaged (see Table II.4 for details). Further, differential attrition between subgroups is problematic because it can bias the overall survey results, especially if the results do not adequately represent subgroups that experience different outcomes than other respondents. Section IV.C has guidance on how states can use follow-up mailings and locating strategies to limit attrition.

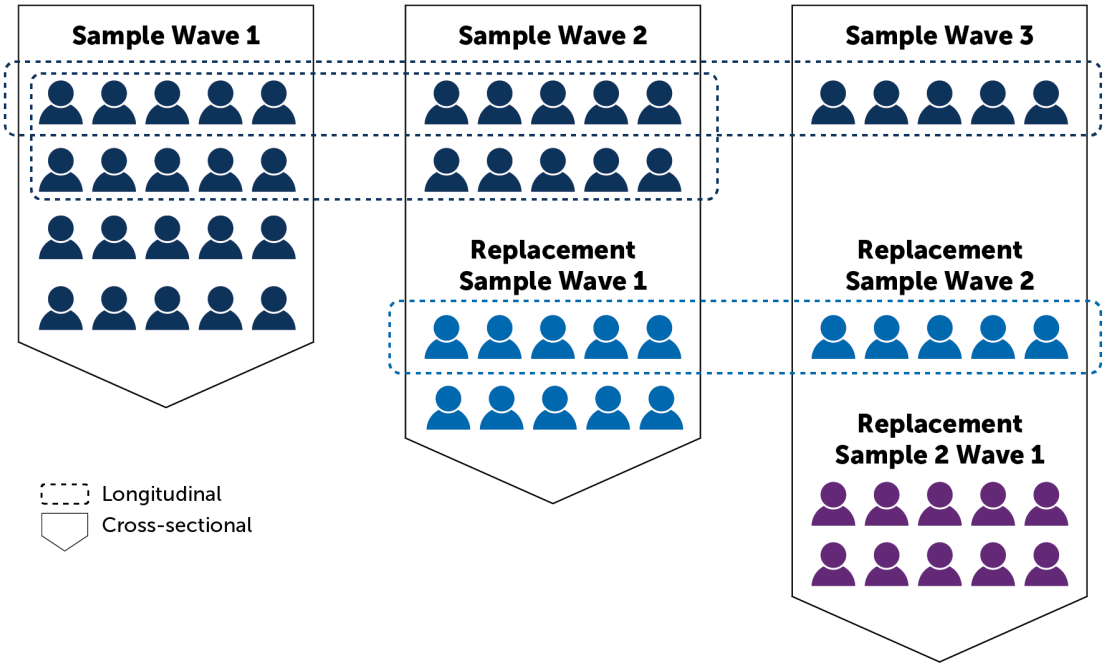
Insights from the field

"Part of what bolstered our longitudinal approach was our ability to achieve a relatively high initial response rate, and we refreshed the sample with new cohorts over time. However, our longitudinal design was also limited because we were unable to conduct a baseline survey before the program was implemented. Having a baseline would have allowed us to see changes within each cohort and make causal inferences about the demonstration."

– Healthy Michigan Voices survey research team

States can also offset the negative impact of attrition by drawing replacement samples of new enrollees, as shown in Exhibit II.1. In addition to maintaining desired sample sizes, replacement samples allow states to collect information about beneficiaries who enroll after the first survey wave(s) and therefore help evaluators to avoid biasing survey data in favor of people enrolled for a long time or those who enroll early in the demonstration.

Exhibit II.1. Longitudinal survey data collection with replacement samples



2. When and how many times should the survey be fielded?

Timing decisions should be informed by the demonstration's logic model, the schedule for implementing the demonstration, the experiences of states that have already implemented similar policies, and states' evaluation resources, and reporting timeframes. States using repeated cross-sections and longitudinal designs must determine how many times to collect survey data and how far apart each survey wave should be. States should allow enough time between survey waves to observe meaningful changes in outcomes.

Evaluators could find it necessary to adjust the planned survey schedule to accommodate changes to demonstration policies or implementation. Documenting the schedule for planned rollout of the demonstration, the changes that take place, and even communications from the state Medicaid agency to beneficiaries can help evaluators plan and adjust the survey schedule over time. Likewise, states should understand the value of keeping evaluators informed about demonstration changes and information communicated to beneficiaries.

Insights from the field

"Implementation and research have to be linked together. We created a timeline of policy changes and communications with beneficiaries so we knew what beneficiaries were seeing and when. We added ourselves to listservs and asked for provider letters. We were careful to avoid fielding the survey right at the time that policy or implementation were changing."

– University of Iowa survey research team

Table II.1 is an example survey schedule based on a five-year demonstration period with four survey waves.

Table II.1. Example survey schedule for a longitudinal survey in a five-year demonstration period

Survey wave	Seeks to collect or measure	Conducted in demonstration months	Add replacement sample?
Baseline	Collects baseline data from which change will be measured over time	0–3 (after demonstration approval and before or shortly after implementation)	No
18-month	Measures change since baseline and short-term impacts, including for former beneficiaries	19–21 (baseline + 18 months)	Yes, to offset attrition
36-month	Measures change since baseline and 18-month waves; also measures intermediate and longer-term impacts, including for former beneficiaries	37–39 (baseline + 36 months)	Yes, although states should assess value relative to cost and complexity, because they will not be able to measure change over time for this group
54-month	Measures change since baseline, 18-, and 36-month waves; measures longer-term impacts, including for former beneficiaries	55–57 (baseline + 54 months)	Yes, for the same reasons shown for 36-month wave, especially if states encounter substantial attrition

3. How long should the field period be?

Allowing more time for data collection can improve response rates by providing the time needed to locate sample members, establish contact, and encourage participation. The field period should be long enough to allow for multiple reminders to nonresponding sample members, ideally in a variety of formats such as letters, postcards, email, text messages, or telephone calls. The field period should also be long enough to pace these reminders so the sample members do not feel badgered by them. In the case of telephone surveys, pacing follow-up over several weeks also increases the likelihood of including respondents who may have interruptions in their telephone service. Evaluators might also want to allow enough time in the field period to move from one mode (like a web survey) to another, more resource-intensive one (like a paper or telephone survey).

To accommodate these activities and minimize the burden on sample members, it is best to allow at least 12 weeks for data collection (as shown in Table II.1). Shorter field periods are

unlikely to achieve high response rates with Medicaid beneficiaries—it can take time and multiple modes of contact to locate and engage this population.

Long field periods of over 12 weeks (in total duration) can be advantageous if they allow evaluators to get responses from more people who are newly enrolled or newly disenrolled because of seasonal employment or demonstration policies. In this case, evaluators could choose to do a "rolling release" of separate cohorts of sample members within the field period (discussed further in Section IV.B). A less complex approach to managing the enrollment churn caused by seasonality and eligibility changes would be to administer the survey at two points in the year.

There are also reasons to limit the field period. Protracted field periods can impact the quality of the data: the further away the survey gets from the ideal point of reference the less it will reflect the time period it is focused on. For example, a 36-month survey interview conducted 42 months after baseline is technically no longer a snapshot of a beneficiary's experience at the 36-month mark. Long field periods are also expensive. Evaluators must staff the survey throughout the field period to monitor the progress of data collection and respond to any challenges that arise. States and their evaluators must balance these considerations against the importance of achieving the desired response rate.

C. Creating a sampling plan

Sampling is a cost-effective way to collect data about a population of interest. Those who are selected for the sample can be used to represent the population as a whole. Common populations of interest in evaluations of section 1115 demonstrations include demonstration beneficiaries, comparison beneficiaries, and former beneficiaries. States can use Medicaid administrative data as the sample frame, or the universe they draw the sample from, for each of these populations. States interested in surveying eligible but never-enrolled beneficiaries must use other sources to construct a sample frame for that population.[9]

In designing the sampling plan, states should consider how to draw cases from the sampling frame to represent the overall population and important subgroups. The sampling plan should also describe the size of these samples and the eligibility criteria for each population of interest. Evaluation designs submitted to CMS should discuss these decisions in detail.

1. How can states and evaluators design a representative sample?

Probability sampling enables the state to draw inferences about the population of interest because each member of the sample is selected at random from the target population and has a known probability of inclusion. Non-probability sampling methods, such as convenience samples, are not recommended for demonstration evaluations; survey results will not represent the population overall and therefore will not support conclusions about a demonstration's effects. For example, a convenience sample consisting of beneficiaries who visit certain health providers

[9] If take-up (the proportion of eligible adults who enrolled) is high, then it could be a difficult task to find respondents for a survey of people who never enrolled. Group interviews (focus groups) can be an alternative strategy for collecting information on barriers to initial enrollment since adequate sample sizes for a survey may be difficult to obtain.

may not represent the views and experiences of beneficiaries who see different providers or who do not have a provider.

Simple random sampling is the most basic type of probability sampling. In a simple random sample, each member of the target population has the same probability of being selected. One way to implement a simple random sample (in the absence of a statistical program that selects it) is to assign every person in the sampling frame a random number, and to draw cases in order of the random number, until the desired sample size is reached. The main drawbacks of this approach are that (1) by chance, the simple random sample may not represent all beneficiaries proportionally with respect to important characteristics like age, sex, race, or area of residence in the state, and (2) achieving adequate representation of small subgroups would require a very large sample.

Stratified random samples allow the state to control the sample size for beneficiaries with certain characteristics. Stratification divides the overall population of interest into separate groups, or strata, defined by one or more characteristics. Each stratum effectively becomes a separate sample. The sample can then be allocated proportionally across strata, or certain strata can be oversampled.

With proportional allocation, stratification allows states to avoid getting a disproportionate distribution of cases by chance. For example, a state stratifying by sex and allocating the sample proportionally would sample men and women separately and include proportional numbers of each. That state would still have to draw a very large sample to achieve adequate representation of small subgroups.

With disproportionate allocation, also known as oversampling, stratification allows states to sample proportionally more cases from some groups than others. Oversampling might be necessary to ensure representation of subgroups that make up small proportions of the overall population (to support robust analyses of smaller subgroups of interest because states expect they may have different outcomes) or that are more difficult to reach in a survey (selecting more cases from hard-to-reach groups). Participation rates vary based on a number of factors including age, sex, income, race, and residence in urban versus rural areas. Stratifying can also hold down costs because it is possible to achieve representation of smaller subgroups without making the overall sample too large.

Insights from the field

"For some demonstration features, relatively small subgroups are affected, so oversampling is essential. For example, some features of the demonstration in our state, like monthly beneficiary account contributions, only affect enrollees with incomes above 100% of the federal poverty level, which is a smaller proportion of the population."

– Healthy Michigan Voices survey research team

Table II.2 shows how each of these sampling strategies would affect the final survey sample.

Table II.2. Illustration of random sampling strategies on distribution of subgroups in a 10% sample of a population of 2,000

Distribution of population characteristics (sex, urban/rural residence, language) by strata in a population of 2,000	Simple random sample of 200[a]	Stratified random sample of 200 with proportional allocation[b]	Stratified random sample of 200 with disproportionate allocation[c]
Female, rural, English-speaking: 220 (11%)	30 (15%)	22 (11%)	25 (12.5%)
Female, rural, Spanish-speaking: 40 (2%)	10 (5%)	4 (2%)	25 (12.5%)
Female, urban, English-speaking: 580 (29%)	46 (23%)	58 (29%)	25 (12.5%)
Female, urban, Spanish-speaking: 160 (8%)	4 (2%)	16 (8%)	25 (12.5%)
Male, rural, English-speaking: 240 (12%)	36 (18%)	24 (12%)	25 (12.5%)
Male, rural, Spanish-speaking: 60 (3%)	2 (1%)	6 (3%)	25 (12.5%)
Male, urban, English-speaking: 560 (28%)	70 (35%)	56 (28%)	25 (12.5%)
Male, urban, Spanish-speaking: 140 (7%)	2 (1%)	14 (7%)	25 (12.5%)

[a] simple random sample takes a sample of 10 percent of the overall population and does not sample within strata. As a result, some strata are under- or overrepresented in the final sample just by chance.

[b] Using a stratified random sample with proportional allocation, the strata are proportionate to subgroups in the population.

[c] Using a stratified random sample with disproportionate allocation, the design uses different fractions of strata, as needed, to ensure adequate representation in the final sample of subgroups that make up small proportions of the overall population.

Clustered or multi-stage sampling can be used in conjunction with any of the above probability sampling methods. Clustered sampling involves selecting a sample of primary sampling units first, then selecting a probability sample of beneficiaries within those units. A primary sampling unit consists of a set of beneficiaries who are clustered in some way, usually by geography. This method is often used when there is a logistical or budgetary reason not to select the sample from the entire population (perhaps because there is no universal sampling frame), or if the survey cannot be conducted across the entire state because it involves in-person data collection. For example, if a new program or policy is implemented only in certain counties or via certain managed care organizations, it might make sense to sample counties or organizations first, then sample beneficiaries within those primary sampling units. Although these methods offer logistical efficiencies, they reduce the precision of the resulting estimates (see discussion in Appendix B).

2. How big should the sample be, and what is the target response rate?

Power calculations can help states ensure they use evaluation resources wisely. Power calculations are done to determine the minimum sample size needed to support statistically sound analyses of demonstration policy effects and detect subgroup variations. Underpowered analyses, based on too-small sample sizes, may fail to detect real policy effects. Overpowered analyses, however, can waste evaluation resources because they collect overly large samples. Appendix B describes the information needed to make power calculations. In most cases, states' independent evaluators will make these calculations, but awareness of the general process could support states' evaluation planning. States should provide power calculations for the overall sample and all subgroups as part of the evaluation designs they submit to CMS.

each survey wave.[10] High response rates are important because they give states a measure of confidence that the final sample represents the population of interest as intended. Low response rates are problematic because they signal a higher likelihood of bias in the survey results. This is especially concerning when respondents differ in meaningful ways from nonrespondents, and the ways they differ are related to demonstration outcomes (Groves 1989; Heffetz and Reeves 2019). States whose surveys have low response rates should exercise caution about concluding that survey results reflect policy effects for the entire demonstration population.

In general, response rates have been declining over time for many surveys. This decline is documented in federal surveys and echoed across the industry (Brick and Williams 2013; Czajka and Beyler 2016; Dillman et al. 2014; Kreuter 2013). Surveys of low-income populations can be particularly challenging to administer successfully given the barriers to survey participation, such as an unstable housing situation or interruptions in telephone access. Section IV.C discusses several strategies states can use to mitigate nonresponse bias and achieve high response rates.

States fielding longitudinal surveys must also consider whether to survey all baseline sample members in each successive survey wave or limit the sample for successive waves to those who responded in the prior wave. This choice has implications for the sample size needed at baseline, the survey costs, and the analytic value of the resulting data. If the sampling strategy limits the sample for each wave to previous respondents, states will need to consider anticipated attrition and response rates for each wave, working backward from the desired number of completed interviews sought in the final wave to determine the right sample size for the first wave. States could reduce the number of cases in the initial sample if they increase the intensity of planned follow-up, which would allow them to achieve a higher response rate with a smaller sample. As noted, it is also possible to use replacement samples to deal with attrition, but states may want to collect repeat observations on some minimum number of people in the sample. In contrast, if states choose to survey all the originally selected sample members in each wave, regardless of whether they responded to a previous wave, a smaller sample is needed upfront. However, this strategy may result in fewer sample cases with repeated observations over time, reducing the analytic value of the survey data.

3. What are the survey eligibility criteria?

Clear eligibility criteria are necessary to determine who the surveyed demonstration and comparison group members will be. Survey eligibility criteria may overlap with, but are not the same as, demonstration eligibility criteria—for example, people who do not speak the language used in the survey would still be eligible for the demonstration, whereas people who have moved out of state would be ineligible for both. Survey eligibility criteria inform the sampling plan, response rates, and administration of each survey wave. States and their evaluators should specify whether certain criteria make a person ineligible for only a particular wave of survey administration or for all subsequent waves of administration.

[10] The American Association for Public Opinion Research (AAPOR) website has information on calculating survey response rates. See https://www.aapor.org/Education-Resources/For-Researchers/Poll-Survey-FAQ/Response-Rates-An-Overview.aspx. (Accessed May 22, 2019.)

Eligibility criteria may exclude people who do not speak the language(s) used to administer the survey or who have died, moved out of state, or aged out of Medicaid eligibility (see example in Box 1). Depending on the survey design and target population, the length of time sample members must be enrolled in the demonstration is another important criterion. Survey questions asking about experience in the demonstration require that beneficiaries must have had enough time to be exposed to demonstration policies. However, requiring many months of continuous enrollment will exclude people who churn in and out of enrolled status.

Box 1. Eligibility criteria for 2016 Healthy Michigan Voices survey of demonstration beneficiaries

- Enrolled in the demonstration for a total of at least 12 months
- Enrolled for 10 of the past 12 months
- Enrollment in managed care in 9 of the past 12 months (i.e., experiencing demonstration policies, which are delivered via managed care organizations)
- Ages 19–64
- Complete Michigan contact information and income level available in administrative data
- Preferred language of English, Arabic, or Spanish

When administering the survey, evaluators can use administrative data and a sample management system to confirm survey eligibility and group assignment. For example, people initially included in the survey group for current demonstration beneficiaries may become disenrolled. Former beneficiaries may be included in the survey, but will likely receive different questions than current enrollees.

4. How can survey weights increase the representativeness of the sample and adjust for nonresponse?

States' independent evaluators should have experience constructing survey weights and applying them in analyses to achieve the goal of unbiased estimates. Weighting is done to ensure that survey data represent the population of interest as closely as possible by placing more weight on some observations than others.[11] Constructing weights is a multi-step process that reflects characteristics of the sample, adjustments for nonresponse, and adjustments to reflect known information about the population.

The base weight accounts for the probability of selection. Evaluators then adjust this weight for different nonresponse patterns. Low response rates do not mean that nonresponse bias must exist, and high response rates do not guarantee an absence of nonresponse bias. However, response rates and nonresponse bias are related: the lower the response rate, the more likely that nonresponse bias will be a problem. This guide suggests several strategies to help evaluators guard against low response rates (see Section IV.C). Despite those efforts, the resulting survey data are usually incomplete, and states should plan to allocate resources for weighting adjustment procedures that account for nonresponse.[12] Nonresponse weighting adjustments are an effective way to account for differences between respondents and nonrespondents, provided that (1) the

[11] For a brief review of survey weights and weighting, see Lavallée and Beaumont (2015).

[12] Note that weights generally account for *unit* nonresponse; that is, when the entire survey is considered incomplete. When a survey is considered complete, but a particular question has not been answered, that is called *item* nonresponse, which is generally dealt with using statistical imputation procedures.

variables that are used to calculate nonresponse adjustments are correlated with whether the sample member responded and with the outcomes of interest, and (2) the variables are available for both respondents and nonrespondents. Analytic tools exist to support the identification of nonresponse patterns, which can then be used for weighting adjustments. For example, R-indicators, or representivity indicators, are numeric indices that show which characteristics are underrepresented or overrepresented among survey respondents when compared to the overall sample.[13] States should seek to partner with independent evaluators who have experience with nonresponse adjustments, and/or use technical assistance provided by CMS. Finally, after adjusting for nonresponse, evaluators may want to make post-stratification adjustments and do weight trimming.[14]

5. What can states do to ensure the sample selection process runs smoothly?

Samples of different populations of interest—such as demonstration beneficiaries, comparison beneficiaries, new enrollees, or other groups—will be most accurate if they are drawn within a few weeks of the field period. Older sample data are likely to have out-of-date group assignments and contact information. States and evaluators should begin planning to produce and transfer sample data well in advance of when the sample draw occurs. States should determine what information will be included in the sample file, who will pull the file, how the data will be securely transferred to the evaluator, and when the files will be needed. States should also plan to draw a test file early on to practice generating the file and to allow evaluators to familiarize themselves with the file layout, data structure, and any coding or quality issues.

D. Importance of mode in collecting survey data

There are four common modes of collecting survey data: self-administered web-based, self-administered paper-based, interviewer-assisted by telephone, and interviewer-assisted in-person.[15] Each mode has inherent advantages and limitations and its own set of best practices for both design and implementation. States must decide which mode to offer to sample members and whether to offer them more than one. Using different modes sequentially can help states achieve higher response rates because respondents can respond in the mode that is most convenient for them, yet they do not have to choose between different modes offered at the same time. Decisions on mode have significant impacts on survey quality, costs, and response rates.[16]

1. What are the key considerations when choosing a survey mode?

Length, or volume of information. The nature and volume of the data states wish to collect should be the principal determinants of survey mode—the decision about which mode(s) to use and the process of instrument development (discussed in Chapter III) are therefore related. Long

[13] See Schouten et al. (2009) for a discussion of R-indicators.

[14] Post-stratification adjustments ensure that marginal weighted totals match internal or external population totals and weight trimming minimizes the variance due to unequal weighting.

[15] Telephone and web surveys that use computer programming to design and deploy survey instruments are also known as computer-assisted telephone interviews (CATI) or computer-assisted web interviews (CAWI). In-person interviews conducted by field staff using these computer-based survey instruments is known as computer-based personal interviews (CAPI); they can be conducted on laptops, tablets, or smartphones.

[16] For more information on survey mode decisions, see Dillman 2009 and 2014 and De Leeuw 2018.

questionnaires are not likely to be successful in self-administered modes. When sample members can visually gauge the questionnaire length by, for example, noticing the thickness of the paper questionnaire, they may toss it out without responding. For web surveys, respondents might grow fatigued partway through the interview and stop before finishing it or click through it without reading carefully. As a result, longer questionnaires benefit when interviewers administer them, because interviewers can help respondents stay engaged and minimize item nonresponse (that is, nonresponse on particular survey questions). States that are using survey data to answer a large number of research questions should use interviewer-assisted modes.

Complexity. Instruments with (1) complex skip patterns or routing paths,[17] (2) items that are asked only of specific subgroups, such as current and former beneficiaries, or (3) items that are asked as follow-ups to particular series of responses are best fielded using computer-assisted formats, whether web-based or telephone-based. These modes can incorporate pre-programmed skip logic, thereby reducing the burden of navigating complex instruments for both respondents and interviewers. Finally, computer-assisted formats are useful when surveys loop through similar scenarios; for example, when they ask about the characteristics of each job held in the past three months. Such questions are difficult to administer successfully with paper instruments.

Languages of administration. When conducting a survey in more than one language, computer-assisted modes can allow interviewers or respondents to convert to another language at the moment the request is made and toggle between languages at any time. In contrast, paper-and-pencil questionnaires can be sent on request or in response to language preference flags in the administrative records. Operationally, paper surveys must assume one language and then rely on respondents to take the extra step of asking for a questionnaire in their preferred language. This can potentially reduce the response rate for non-English speakers. Furthermore, non-English-speaking sample members can face barriers to participation if the survey outreach is not in a language they understand well.

Budget and available resources. All survey modes require skilled labor for survey design, programming, and testing. For paper-based questionnaires, there are additional labor costs associated with formatting each version (such as by language or survey group), printing and mailing, receipting, entering data, and doing quality assurance checks on the completed survey forms.[18] States using interviewer-administered modes will need to allocate resources for training interviewers, paying them to conduct the interviews, and paying travel expenses for in-person interviews. Using more than one mode can also add labor costs associated with merging the data sets from different data collection systems.

Table II.3 summarizes the considerations for each mode.

[17] Skip patterns and routing paths are interchangeable terms. They direct respondents to a new question or set of questions based on their response to an earlier question, automatically skipping non-relevant questions.

[18] Evaluators typically assume "double entry" of all paper questionnaires to ensure the data recorded match what the respondent provided. Two different survey team members record the responses in each questionnaire received. The survey team analyzes responses to locate discrepancies, and makes corrections using the original questionnaire. An alternative to a double entry system is to scan completed questionnaires. However, this process requires special formatting and contracting with a vendor to scan completed forms.

Table II.3. Mode considerations in designing state beneficiary surveys

Mode	Length	Complexity	Group-specific items or modules	Administration in different languages	Unique costs to implement
Self-administered web	Brief; ideally not to exceed 15 minutes	All levels; complex surveys with skip patterns can be programmed to reduce burden	Easily facilitated (for example, to distinguish items for current and former beneficiaries)	Easily facilitated	Website hosting
Self-administered paper	Brief, ideally not to exceed 15 minutes	Low	Create different versions for each subgroup	Create separate versions for each language	Mail handling, postage (outbound and return), and data entry
Interviewer-administered telephone	All lengths; interviewers keep respondents engaged. Best to keep <30 minutes to minimize burden and fatigue.	All levels; complex surveys with skip patterns can be programmed to minimize respondent burden or risks related to interviewer skill	Same as self-administered web	Easily facilitated; interviewers can also toggle from one language to another, as needed, after the interview begins	Labor for conducting outreach and administering interviews
Interviewer-administered in-person	Same as telephone	Same as telephone	Same as self-administered web	Same as telephone	More expensive labor for outreach and administering interviews; travel expenses

2. How can mixed-mode designs overcome barriers to participation and achieve target response rates?

Barriers to participation by mode. For surveying people with low incomes such as Medicaid beneficiaries, the choice of survey mode is critical because different modes can raise barriers to survey participation and therefore impact response rates. For example, a survey designed to be conducted as telephone-only, with interviewer administration, may help mitigate the risk of item nonresponse because interviewers can probe for answers or address respondents' concerns about particular survey questions. However, using this mode alone runs the risk of nonresponse bias, because the respondents could have only intermittent phone service, use pre-paid cell phones (with limited numbers of minutes per month), or have no telephone at all. Likewise, paper questionnaires delivered by mail might not reach sample members who are homeless or in unstable housing situations, or who change their mailing address without updating the postal service. Finally, states could find web surveys appealing given the low cost per unit and the common perception that access to the Internet is widely available through smartphones or in community settings such as public libraries. However, people with low incomes can still have problems completing surveys in this mode (Matulewicz 2017; DiMaggio 2001), and the quality of Internet access can vary by geographic area. For example, although many people have smartphones, they may not have access to wireless connections outside of public spaces. For

these reasons, web surveys of Medicaid beneficiaries may be more successful as part of a mixed-mode design than as the only mode.[19]

Table II.4 summarizes the participation barriers that people with low incomes face for each possible survey mode.

Table II.4. Potential barriers to survey participation for low-income populations, by mode

Mode	Potential barriers to participation
Self-administered web	• Limited digital literacy – trouble logging into or navigating a survey website • Limited literacy • Lack of access to a secure, private place to complete the questionnaire • Concerns about phishing scams • Concerns about data privacy
Self-administered paper	• Transiency – moving often, sometimes without a forwarding address • Housing instability – living with friends or family, impossible to locate using traditional online resources • Limited literacy • Cognitive barriers to navigating skip patterns or following survey instructions • Impaired vision, making it difficult to read the text
Interviewer-administered telephone	• Intermittent phone service • No access to a telephone • Use of pre-paid cell phones, causing person's phone number to change frequently • Calling plans with limited number of minutes per month • Concerns about telephone scams; confusing surveys with telemarketing • Hearing impairments that preclude participation by telephone without use of assistive technologies
Interviewer-administered in-person	• Lack of access to a private place to answer the questions • Transiency – moving often, sometimes without a forwarding address • Housing instability – living with friends or family and cannot be located using traditional online resources • Concerns about personal safety (opening the door to a stranger), scams, confusion with for-profit or other solicitations

Using modes sequentially to make it easier for beneficiaries to respond. States can use different modes of data collection in turn to let sample members choose the one that is most convenient for them. For example, the letter inviting prospective respondents to complete the survey can offer a web alternative, and the evaluator can also mail a paper questionnaire at the start of a second phase of the field period. Follow-up may continue thereafter, with calls placed to those who have not responded to either mode. Staggering the introduction of new modes

[19] Evaluators fielding a survey by web should supply the following information to the state: (1) rates of participation by web in other surveys they have recently conducted with this population; (2) participation rates from a pilot survey; and/or (3) results and participation rates from the first wave of survey data collection. These data points can inform whether the web mode is a cost-effective option for inclusion in a mixed-mode design. It may be that the costs involved in development, testing, and deploying the survey by web are not outweighed by the relatively low costs associated with data collection or the likely number of responses that can be expected. States should also consider the quality of responses collected by web as part of this decision.

within each survey wave makes it possible to use less expensive modes before switching to more expensive ones.[20] This strategy also avoids overwhelming sample members with choices, which may depress response rates (Medway 2012). States and their evaluators should specify when each mode will be made available to sample members, depending on time elapsed in the field period or on other metrics, such as the percentage of surveys that are completed.

Importance of integrated data collection systems for mixed-mode surveys. Advance thought should be given to how data from multiple modes will be integrated. For example, if the survey switches from mail to telephone, evaluators will need to track cases that have returned a completed questionnaire by mail to exclude them from the queue of cases prepared for telephone follow-up. As another example, if survey staff enter data from completed paper surveys in a system that is separate from the telephone (or web) survey software, state evaluators should have a plan in place to integrate the files at the end of the survey.

3. How do different modes affect data quality and facilitate quality assurance checks?

Each mode of collecting survey data has a unique impact on the quality of the data. Table II.5 summarizes quality concerns that can arise when using different modes of survey data collection, and some quality assurance steps states can take to eliminate or mitigate the impact of these issues. Less expensive modes like self-administered paper surveys can result in data that are incomplete (for example, if respondents did not follow the skip patterns or skipped whole pages), inconsistent, or illegible, and therefore not viable for entry and use in analysis. Computer-assisted modes can help improve data quality by ensuring that respondents are automatically routed to the correct questions, and can also include dynamic data quality checks, such as warning messages to a respondent whose answer is not within the allowable range or who leaves an item blank. Interviewer-administered modes also can also support high data quality, as interviewers can work to ensure that respondents answer the questions, probe responses that are not clear, and address respondents' concerns that could have otherwise resulted in nonresponse. Interviewer-administered modes are subject to different quality issues, but states can invest in training and monitoring strategies to overcome them. These investments are nontrivial, but they can have real impacts on the quality of the data.

[20] To make wise use of evaluation resources, states should create a detailed plan for optimizing response for each mode before moving to another, more expensive, mode. For more information on developing such a design, see Groves et al. (2006b).

Table II.5. Proactively addressing data quality concerns in each mode

Mode	Quality concern	Quality assurance steps
Self-administered web	• Relies on the respondent to read questions carefully, navigate the screen, and comprehend survey terminology • Missing data (item nonresponse) • Poor response quality—for example, respondents might select only responses at the top of a list or proceed rapidly through items without reading carefully	• Define terms that might be unfamiliar with hyperlinks to minimize verbiage on screen. • Deploy "soft checks," which flag items that are left blank or responses entered faster than expected. • Randomize sequence of response options when possible (if not administering questionnaire in other modes, and when format of responses does not necessitate rank-order).
Self-administered paper	• Relies on the respondent to read questions carefully, comprehend terminology, navigate skip directions, record answers clearly, and return completed questionnaire • Missing data (item nonresponse) • Lack of adherence to instructions, such as selecting multiple answers to a "select one" question	• Define terms that might be unfamiliar. • Develop guidance on what will constitute a completed case—if specific items or a specified portion of the items are blank, interviewers can call the respondent to retrieve missing data. • Develop guidance (ahead of data entry) for data processing staff on how to edit data in these circumstances.
Interviewer-administered telephone	• Potential social desirability bias—respondents could adjust their answers to satisfy the interviewer, or present themselves in ways they think the interviewer will consider positive or normative • Interviewers could collect poor quality data by not recording responses correctly, not reading questions verbatim, not probing responses when they need to, or pacing the interview too rapidly for respondents to digest the information. • Interviewers can introduce bias by persuading or pressuring respondents to answer in a particular way.	• Address social desirability bias before asking questions on potentially sensitive topics. For example, assure respondents there are "no right or wrong answers." • Word questions to normalize all possible responses. Train staff to read questions on sensitive topics in the same way they read all other questions. • Hire survey staff with the aptitude to do the work well, and train them to develop the skills to carry out the interviewing task successfully. • Have onscreen prompts for probes that are expected to be necessary in many cases. • Monitor calls and provide ongoing feedback and training on adherence to best practices, such as probing, pacing, and reading items verbatim.
Interviewer-administered in-person	• Same issues as telephone interviewing, plus the risk of falsification (i.e., interviewers populating responses or entire interviews to get credit for a completed interview)	• Same strategies as in telephone interviewing. In addition, embed validation checks within a questionnaire for data only known to the respondent, or that could serve as flags for potentially falsified cases during data review. • Conduct validation checks on interviews. Strategies could include a review of the data collected (are responses internally consistent?), review of process data[a] (such as interview duration), inclusion of audio-recordings during the interview that can be activated for specific items, and sending mailings or placing calls to respondents to confirm they completed the interview (Murphy et al. 2016).

[a] Process data, sometimes called "paradata," provide information about the data collection process (Couper 2000). For example, these data could include the mode of survey completion, timestamps linked to when a question was answered, or location of a field interviewer when conducting an in-person interview.

Different responses to the same question in different modes. Research has shown that mode affects how respondents understand what is being asked of them and affects their responses to the questions (De Leeuw 2018). Changing the response based on the mode is part of what is known as "mode bias." For example, in interviewer-administered formats, respondents hear the question and response options, and they are more likely to answer a question with the response option they heard last. In contrast, in self-administered formats, respondents see the question and its response options, and tend to use the response options at the top of a list more often than they pick the ones further down. Respondents using mobile devices to complete web surveys may be less inclined to scroll through numerous response options or read detailed text.

States can try to mitigate this bias when they design the instrument and monitor data collection and processing. Evaluators deploying high quality mixed mode surveys will (1) ensure the questionnaires are designed with equivalence across each mode, (2) estimate mode effects by separating the intended mode selection effects from unintended mode measurement effects, and (3) adjust for unintended differential mode effects when the survey data file is being prepared.[21] States should also analyze their data to gauge how big an issue mode bias could be, both while the survey data are being collected and at the end of each wave.

Sensitivity of survey items. Questions about sensitive topics run the risk of being left blank in self-administered modes. Interviewer-assisted modes give respondents a chance to ask questions, and interviewers can sometimes allay respondents' concerns about a given topic. However, some respondents may answer sensitive questions differently with an interviewer than they would in self-reporting modes, in accordance with their perceived social norms or expectations about the "right" answer. States should factor this into their planning on mode choice and their testing efforts (see Section III.C for further discussion of pre-testing).

[21] For more information on best practices for deploying mixed-mode surveys, see De Leeuw (2018) and Dillman (2014).

This page has been left blank for double-sided copying.

III. DEVELOPING THE INSTRUMENT

This chapter offers guidance on developing a high quality questionnaire, or survey instrument, for each wave of survey administration (Section A), preparing instruments for use (Section B), and testing draft instruments and survey procedures (Section C). A well-designed instrument sets the stage for collecting high quality data across any and all modes. In contrast, a poorly designed instrument can impose unnecessary burden on respondents, yield poor quality data, and result in high rates of nonresponse. All draft survey instruments must be submitted to CMS before implementation. To conserve evaluation resources, states should submit instruments for CMS approval before evaluators begin survey programming for computer-assisted modes.

A. Moving from research questions to survey questionnaires

In this section, we suggest a series of steps to organize the process of developing an instrument using both existing and new survey items.

1. What is the best way to start developing the instrument?

As a first step, states and their evaluators should consider whether it is possible to collect data on outcomes of interest through reliable non-survey sources. Linking survey data to Medicaid administrative data or other administrative data could have several advantages.[22,23] The chief advantage is the potential to reduce the length of the survey instrument, which in turn decreases respondent burden and bolsters response rates (see length/timing guidelines by survey mode in Table II.3, Section II.D). Linking survey and administrative data can also help evaluators triangulate data sources or make the best use of data from each source. For example, Medicaid administrative data might be a better source of information on topics that are difficult for respondents to recall, such as health service utilization in the previous year.[24] Survey data can

Insights from the field

"Gathering accurate income information through surveys is difficult because wage variation year-to-year—or even month-to-month—is high for low-income populations. Wages may change frequently or are seasonal. In addition to measurement error, we saw high item nonresponse in our baseline survey, which is also an issue in national surveys like the BRFSS. Compounding both issues is the fact that income questions take a good deal of time to field. Based on our experience, we strongly recommend triangulating data sources for income by looking for non-Medicaid administrative data like unemployment filings, tax data, and welfare income."

– University of Pennsylvania evaluation team for Kentucky HEALTH

[22] Potential non-Medicaid administrative data sources include data systems for public programs such as Temporary Assistance for Needy Families (TANF) and the Supplemental Nutrition Assistance Program (SNAP), state workforce or tax data, state unemployment insurance filings, and all-payer claims database (APCD) data. These sources might not be available in every state, and if they are available, they could have serious limitations. For example, relatively few states have APCDs, and APCDs vary in (1) whether they use a unique person identifier that can track transitions across insurance coverage types, (2) the quality of the unique person identifier, and (3) how many insurance carriers report to the APCD.

[23] Combining data from different sources can be problematic if the non-survey source has limited variables, uses a different unit of analysis (e.g., household rather than person), is not available for the needed time frame, or is otherwise not linkable to sample members.

[24] Evaluators who plan to link administrative data to survey responses from disenrolled beneficiaries should ensure they maintain a crosswalk of Medicaid IDs to study IDs.

then be used to build on administrative data by providing context or information on beneficiaries' decision making process when they seek care. Finally, evaluators can also consider using linked claims data to form subgroups with different levels of service utilization as part of the survey sampling strategy.

However, administrative data have only limited ability to replace survey data for many measures. For example, state workforce or tax data might not be reliable sources of income information because some workers have low enough incomes that they do not file taxes, or they may have contingent or temporary employment they do not report to states other than to satisfy community engagement requirements. Likewise, survey data may be the best source of information on expected outcomes among former beneficiaries, such as transitions to commercial health insurance. Even if states are able to follow beneficiaries using data sources like state all-payer claims databases, the ability of these databases to follow people across coverage transitions varies by state.

After states and their evaluators decide which beneficiary outcomes require measurement with survey data, they can begin the questionnaire development process by coming up with a list of the topics the survey should address. Table III.1 is an example list that reflects CMS's evaluation design guidance for eligibility and coverage demonstrations.[25]

[25] The CMS guidance on designing evaluations of eligibility and coverage demonstrations includes suggested research questions, evaluation methods, and data sources for evaluations of community engagement requirements, premiums, non-eligibility periods, and retroactive eligibility waivers. States with eligibility and coverage demonstrations might not have all of these demonstration policies, and therefore would need to cover only some of the topics included in Table III.1.

Table III.1. Survey domains and topics relevant for evaluations of section 1115 eligibility and coverage demonstrations

Current employment, education, and credentials	
Current employment status	Past year: employment status[a]
Average number of hours worked per week[a]	Past year: number of months of continuous employment
Wage at current job(s)[a]	Past year: average number of hours worked per week[a]
Whether job(s) offer health insurance[a]	Highest grade attained
Industry of current job(s)[a]	Degree, credentials, certificates attained[a]
Income and expenses	
Income sources outside of wages (e.g., retirement, disability benefits, TANF, SNAP, money from friends or family)[a]	Child care costs
Amount from each source of income	Transportation costs
	Recent loss of public program eligibility
Health insurance coverage and barriers to coverage	
Coverage by any form of health insurance	Eligibility for employer-sponsored health insurance
Source(s) of current health insurance (commercial insurance through employer or spouse's employer, Medicaid, Marketplace, Tricare, other)	Out-of-pocket medical spending in past year[a]
	Problems paying insurance or medical expenses
	Barriers to enrollment in new coverage
Health status and health care utilization	
Current physical health status	Hospital admissions in the past year
Days with poor mental health (past 30 days)	Preventive care receipt (e.g., flu shot)
Current mental health status	Chronic care receipt for targeted conditions (e.g., asthma, diabetes)
Days with poor physical health (past 30 days)	
ER admission(s) in the past year	Unmet medical needs during state program lockout
Knowledge of program requirements[c] and barriers to compliance	
Knowledge of demonstration requirements such as community engagement or payment requirements; beneficiary account rules and incentives; non-eligibility periods or other consequences for noncompliance	Barriers to compliance with requirements (knowledge, child care, health, transportation, other)
	Barriers to timely renewal
Demographics, living arrangements, and household characteristics	
Age[b]	Race[b]
Sex[b]	Ethnicity[b]
Primary language	Household composition and living arrangements
Work disability	

[a] These topics may require more than one survey item to collect high quality data or to measure the topic with the desired degree of specificity.

[b] Demographic information may be included in administrative data, but may not be reliable in quality. States could choose to include questions about demographics to capture more detail, for example, by asking about more types of ethnicity than Hispanic or Latino, or they could use demographic questions to verify the respondent's identity.

[c] States might want to survey for program awareness in addition to knowledge of program requirements. Responses to items on program knowledge have limited utility if respondents have no awareness of the program or their participation in it. If responses signal no awareness, states can choose to reduce burden by routing respondents out of items that assume knowledge.

ED = Emergency room; TANF = Temporary Assistance for Needy Families; SNAP = Supplemental Nutrition Assistance Program.

Next, evaluators could develop a table that lists the survey topics for each population of interest, using columns to indicate whether questions on the topic will be asked of all groups of interest and for which wave(s) of administration. These topic tables can later serve as checklists to ensure that questionnaires include the right survey questions. Topic tables can also help to

organize the instrument development effort by identifying where all groups should be asked the same questions or where only certain groups will get questions.[26]

Table III.2 is a simplified example table shell for a survey with four rounds of data collection for two study groups—a demonstration group ("Demo") and a comparison or control group ("Ctrl"). This example assumes that it is not necessary to ask the comparison group about its understanding of the program, and that both groups have Medicaid coverage at baseline.

Table III.2. Example topic table to guide instrument design

	Baseline		18-mo		36-mo		54-mo	
Domain	Demo	Ctrl	Demo	Ctrl	Demo	Ctrl	Demo	Ctrl
Current employment, education, and credentials	X	X	X	X	X	X	X	X
Income and expenses	X	X	X	X	X	X	X	X
Health insurance coverage and barriers to coverage			X	X	X	X	X	X
Health status and health care utilization	X	X	X	X	X	X	X	X
Knowledge of program requirements and barriers to compliance			X		X		X	
Demographics, living arrangements, and household characteristics	X	X	X	X	X	X	X	X

2. Which survey topics warrant one survey item, and which should have several?

Next, states and their evaluators should determine the level of depth and precision they want to achieve on each topic. This helps determine the number of questions that need to be in the survey instrument. Examples of relevant survey topics follow.

- **Employment.** It is possible that respondents are currently working at more than one job, or will have held more than one job during the time frame of interest. Therefore, states and their evaluators should decide whether to collect data on the characteristics of all jobs held by the respondent, just the respondent's primary job, or the most recent job (for those who are currently not working). More than one question might be needed to understand the experiences of those who work multiple jobs, and whether those jobs differ in content or on other dimensions. Operationally, asking the same questions about multiple jobs requires loops (repeated sets) of follow-up questions (such as hours, wages, health insurance, and industry) for each job, which can be facilitated with programming for computer-assisted surveys. For questions about recent employment, expanding the timeline for measurement could enable the state to collect observations on more individuals, especially those who are employed sporadically or seasonally. This would, however, make the survey longer and more complex.

[26] For computer-assisted interviewing modes, programmers can route respondents into or out of specific instrument subsections based on their group assignment (for example, whether they are demonstration or comparison group members). States deploying their survey using pencil-and-paper administration should create one instrument for each group of interest, rather than expecting respondents to follow complicated skip instructions.

- **Sources of income, and amount received from each source.** Asking about the respondent's income by source can help evaluators create a more valid and precise measure of total income than they would be able to create by asking respondents to estimate a total amount. Further, some respondents might not interpret certain resources, such as public benefits, as "income." Asking about each source separately, and finding out how much the respondent receives from that source, adds two items to the instrument per source.
- **Demographics and household characteristics.** If available administrative data are of good quality and have detailed enough demographic information to allow analysis, states could keep the survey shorter by excluding demographic questions, but they could still include demographic questions if they wanted more response options. For example, survey instruments can include questions about more ethnicity types than Hispanic or Latino, or non-binary questions about sex to understand the experiences of those who do not identify as male or female, or who identify as transgender. Certain demographic items can also be placed at the beginning of a questionnaire to validate that the survey has reached the intended respondent. This is especially useful in households where the respondent shares the same first and last name as a child.

3. What are the advantages to using existing resources in developing survey items?

Once topics are selected, the states can start to compile survey items that will yield the data needed for each of them. As a first step, states should determine whether existing national surveys or other states' publicly available surveys contain relevant items that can be adopted verbatim or modified.

Using items from existing surveys has several key advantages. For example, this approach can increase confidence in the validity of the resulting data, because those items were carefully constructed by survey experts. Items used in national surveys have also been subjected to rigorous testing and widely used. As another example, if states retain the original item wording, they will be able to compare data for their 1115 demonstration evaluation to data for the state as a whole, to data for other states, and possibly to data at the national level. This can be particularly useful for assessing and interpreting data on outcomes such as rates of employment, health insurance coverage, and self-reported health status. In addition, adopting or modifying existing items conserves evaluation resources because this reduces the number of new items the state and its evaluator must develop and test.

As noted in CMS's evaluation design guidance for eligibility and coverage demonstrations, national household surveys that are potentially useful for evaluations because their sample sizes support state-level estimates include the Integrated Public Use Microdata Sample version of the American Community Survey,[27] the Behavioral Risk Factor Surveillance System,[28] and the

[27] The Integrated Public Use Microdata Sample version of the American Community Survey (ACS) is a research-ready version of the ACS prepared by the Minnesota Population Center at the University of Minnesota. See https://usa.ipums.org/usa/acs.shtml.

[28] The Behavioral Risk Factor Surveillance System was established by the Centers for Disease Control and Prevention and is administered by state health departments. See https://www.cdc.gov/brfss/about/index.htm.

Current Population Survey Annual Social and Economic Supplement.[29] Other national surveys have smaller sample sizes but more items that could be relevant. These include the National Health Interview Survey,[30] the Survey of Income and Program Participation,[31] the American Housing Survey,[32] and the Consumer Assessment of Healthcare Providers and Systems.[33]

To support rigorous evaluations, help states conserve evaluation resources, and increase the comparability of state-based beneficiary survey data, Appendix C to this guide provides a comprehensive bank of survey items compiled from existing national household surveys and surveys used in recent state-based evaluations of section 1115 demonstrations. Each survey item maps to an outcome of interest listed in the CMS evaluation design guidance for eligibility and coverage demonstrations. States can adopt or customize these items for their own evaluations as needed. In addition, Appendix C suggests question wording for topics for which we could not find existing survey items. Evaluators could also reference a public repository of survey items compiled by the Centers for Disease Control and Prevention, known as Q-Bank, which provides information on how each survey item was evaluated.[34]

4. What strategies can inform development of new items or response options?

Some topics of interest for evaluations of 1115 eligibility and coverage demonstrations will require entirely new or significantly modified survey items. For example, states are unlikely to find existing items that assess beneficiaries' understanding of state-specific demonstration requirements. Evaluators can draft new questions based on published research or administrative data, but the risk in this approach is that they could make incorrect assumptions about beneficiaries' perceptions or experience, and as a result include non-relevant response options or exclude potentially relevant ones.

Another way to approach development of new items is to collect qualitative data through individual or group interviews (also called focus groups) to help elicit issues and themes that should be reflected in survey items.

---

[29] Current Population Survey Annual Social and Economic Supplement is administered by the U.S. Census Bureau. Available at https://www.census.gov/programs-surveys/saipe/guidance/model-input-data/cpsasec.html.

[30] The National Health Interview Survey is administered by the National Center for Health Statistics. Available at https://www.cdc.gov/nchs/nhis/about_nhis.htm.

[31] The Survey of Income and Program Participation is administered by the U.S. Census Bureau. Available at https://www.census.gov/programs-surveys/sipp/about/sipp-content-information.html.

[32] The American Housing Survey is administered by the U.S. Census Bureau. Available at https://www.census.gov/programs-surveys/ahs/about.html.

[33] The Consumer Assessment of Healthcare Providers and Systems is administered by the Agency for Healthcare Research and Quality at the U.S. Department of Health and Human Services. There are several versions of this survey; suggested as useful for 1115 demonstration evaluations are the Hospital, Adult Medicaid, and Home and Community Based Services versions. Available at https://www.ahrq.gov/cahps/about-cahps/cahps-program/index.html.

[34] For more information on Q-Bank, see https://wwwn.cdc.gov/QBANK/Home.aspx.

- **One-on-one interviews** allow for flexibility in scheduling, which can reduce the burden on participants. Individual interviews also afford privacy, which can be especially important for sensitive subjects. However, individual interviews are labor- and resource-intensive, which limits the number of participants states can include.
- **Group interviews (focus groups)** foster group discussion, which enables participants to build on the ideas or contributions of others (Jarrett, 1993). Although they can be labor-intensive to schedule, group interviews can also be an efficient way of collecting qualitative information from several people in a single session, thereby compressing the amount of time researchers spend collecting the data. Group interviews can be done in person or online. Evaluators should plan to recruit more participants than they need in case some of them cannot attend.[35] Reminder calls or text messages can also bolster attendance. These strategies are especially important for people with low incomes because they might have trouble finding transportation or getting time off their job to attend.

In either approach to qualitative data collection, respondents do not need to constitute a representative sample of demonstration beneficiaries, although it is possible to stratify the recruited sample to ensure it includes people from subgroups of interest. Table III.3 has some examples of open-ended, exploratory questions states could include in focus groups.

Table III.3. Examples of exploratory interview questions to inform survey development

Topic	Exploratory questions

Knowledge of state program requirements	• When this new state program rolled out, how did you learn about the expectations that came with it?
	  o If web-based information: Tell me about your experience with that site.
	  o If paper-based information: Do you still have the paper? How would you describe it—was it clear or confusing? Why?
	• Tell me about [key demonstration policy]. Is that something you've heard of before today? If yes, what is it about? What are your thoughts about that?

Barriers to complying with state requirements	• Some people find it hard to meet requirements for [fill in demonstration policy].
	  o Have you ever found it hard to meet the requirements? What did that look like for you?
	  o What sorts of resources or help, if any, were available to you in dealing with these problems?
	• Here's an example. [Give an example.] Can you tell me what that would look like in your life, for you or your family? What challenges would you face? Any others?

Qualitative research techniques can also be incorporated into survey instruments by including items that allow open-ended responses. This approach can be useful if the state wants to capture respondents' experiences in their own words or allow responses that are not covered in a pre-set list. Including an open-ended response option (following a pre-set list) can be especially useful if the state is still learning about beneficiaries' lived experience with a given policy. The state can use themes identified through the open-ended response option to inform new pre-set response options in the next wave of the survey.

[35] This is especially important for group interviews with Medicaid beneficiaries or others with low incomes, who might have trouble participating in a group for a host of reasons including an inability to find child care or transportation or get off work. Evaluators could address this by working with providers or other community-based organizations that serve Medicaid beneficiaries to identify and recruit potential participants.

Although they can add value to surveys, open-ended responses should be used sparingly. They place a burden on respondents to think about their answer and then take the time to record it or convey it to an interviewer. The resulting data can be labor-intensive and expensive to process, especially with surveys administered to large numbers of people. Processing open-ended responses can include editing for typos and spelling errors, at a minimum, or it can involve coding for themes.

5. What are best practices for wording a survey item?

States should seek to collaborate with independent evaluators who have experience in developing neutral, readable survey questions for Medicaid beneficiaries. There are many potential pitfalls involved in designing new survey items (Box 2).

Box 2. Common pitfalls in wording new survey items

Confusingly worded questions can produce measures that suggest limited understanding of demonstration policies when in reality beneficiaries did not understand the questions.

Difficult-to-read questions likewise lead to biased data or low response rates on those items. Readability matters—keep item wording, along with text in survey materials, at or below a 5th grade reading level.

Double-barreled questions ask about more than one aspect of a policy but only allow one answer, preventing accurate interpretation of responses.

Leading questions, or questions that lead beneficiaries to avoid admitting ignorance about a policy, produce biased measures of beneficiaries' understanding and preferences.

Questions about sensitive topics such as risky health behaviors may lead beneficiaries to give socially acceptable responses, resulting in social desirability bias.

6. How should items be modified for different modes?

States deploying mixed-mode surveys will also need to modify item wording or transition text as needed for each mode. For example, in an interviewer-administered format, the preamble text to a series of agree/disagree statements would read, "Next I am going to read you a list of statements. For each, please tell me whether you **agree** or **disagree.** There are no right or wrong answers." In a self-administered format on paper, the same preamble changes to, "Below is a list of statements. For each, please check the box to show whether you **agree** or **disagree**. There are no right or wrong answers."

B. Preparing a survey instrument that can collect high quality data

After drafting the survey instruments, states and their evaluators should prepare the instruments for pre-testing and data collection by defining completed and "qualified partial" cases. States must also decide how to translate surveys if conducted in languages other than English, whether to allow proxy interviews, and what programming is necessary to support computer-assisted modes.

1. How should completed and qualified partial cases be defined to facilitate survey administration and data analysis?

After survey instruments are complete, and before data collection starts, states should work with evaluators to stipulate clear criteria on what constitutes (1) a completed questionnaire or

interview (i.e., a "completed case") and (2) a "qualified partial" that would be used in analysis. Partial completes can be designated based on a number of criteria, including what percentage of the questions have been answered, whether critical items have been completed, or a combination of the two. Stipulating that a case counts as completed if some percentage of questions are answered can be excessively complicated to operationalize in surveys with complex routing paths. It is often better to consider whether there are critical variables that, if missing, would have a significant negative impact on planned analyses. Whenever possible, these items should be placed toward the beginning of the survey instrument. During each field period, evaluators should report to states on the number of completed cases and qualified partials. Evaluators can also make use of the qualified partial designation by ceasing to follow up on respondents who reach a qualified partial status—this can conserve resources and avoid placing more burden on sample members. Regardless of how states and their evaluators define completed cases and qualified partials, the definitions should be clearly documented as part of the survey methodology.

2. How can surveys collect high quality, reliable data from non-English speakers?

Making surveys available to non-English speakers can increase response rates by removing a language barrier that could otherwise prevent some sample members from responding. If there are enough non-English speakers in populations of interest, states should consider offering them the survey in their own language.[36]

If the survey is being administered in languages other than English, states must decide whether to use professional or ad hoc translation. Professional translation of each version of the questionnaire ensures reliable survey administration and is the preferred approach to collect high quality data, because states and their evaluators can be confident that questions and responses are comparable across languages (Harkness 1998). Questions are translated in advance and worded the same for all respondents. Evaluators can also try to ensure that only staff who demonstrate proficiency in a given language are allowed to conduct interviews in that language.

In contrast, ad hoc translation means that the questionnaire could be administered in another language, but each person conducting the interview in that language might word the questions slightly differently. This approach does not ensure reliability because interpreters can have different communication styles and language mastery. The resulting data will likely be of poor quality. As an alternative, some sample members who do not speak the languages the survey is administered in could be included in the survey data via proxy response (as described in the next section).

Important considerations for translating survey instruments include readability, formality of language, and availability of existing translated items elsewhere. First, readability is just as important in other languages as it is in English—if states and their evaluators ensure that questions are understandable for people at a fifth-grade reading level in English, then the translation should be at the same reading level. Second, states will need to consider whether

[36] If Medicaid administrative data include information on language preference, this can help states choose an appropriate survey design.

survey items should use formal or informal word choice, especially for words like "you," which can be distinct in many languages. Although formal wording can establish legitimacy and set a professional tone, some respondents could find it off-putting, especially for interviewer-administered surveys. Third, if states elect to use items from existing national or state surveys, they should also use the available translated versions of those items.[37] This is especially important if states plan to compare their findings to state or national data on non-English-speaking populations.

3. When should surveys allow proxy interviews?

Sometimes sample members cannot respond for themselves. In these instances, the questions could be answered by someone else who is knowledgeable about the topics covered in the interview and could reasonably respond on the sample member's behalf. By permitting a proxy to complete the interview, states can include responses on behalf of sample members who would otherwise be unable to participate. This includes individuals with cognitive disabilities or other mental impairments, or those who would not be able to participate in the modes that are offered due to physical limitations. Although allowing proxy interviews raises survey response rates, the data collected might not be as direct a reflection of the sample member's experience as data collected through self-report.

If the state chooses to permit proxy interviews, there should be a specific question included in the instrument that captures whether the interview is being completed by self- or proxy report. Each item should incorporate logic that flags when wording should be adjusted for proxy respondents. These flags are called "text fills," and they help ensure smooth and consistent administration of questions without burdening an interviewer to adjust text as he or she goes. Examples of such fills include: "What is [your/(his/her)] …" or "For how long [have you/has (Sample member's name)] …" Such text fills also help keep a proxy respondent focused on the task of answering the questions on someone else's behalf, not about themselves.

Items that are subjective in nature cannot be answered by proxies. Only the sample member can report on his or her lived experience, knowledge of a policy, or barriers to compliance. Subjective items should have clearly designated specifications for programmers that route proxy respondents out of a given question or out of an entire module. Self-administered paper instruments can include similar instructions.

4. How can computer-assisted formats optimize survey instruments?

If states and their evaluators decide to use computer-assisted interviewing to deploy the survey, it is critical that the survey instrument provided to programmers contains all of the necessary information to inform programming and sets the stage for the collection of high quality data. Examples of this information follow.

- **Item universe.** Evaluators should give programmers instructions about which respondents should receive each item. If a particular item is relevant only to a certain subgroup, the

[37] The American Community Survey, the American Housing Survey, the Behavioral Risk Factor Surveillance System, the Consumer Assessment of Healthcare Providers, and the National Health Interview Survey all have Spanish versions available online.

instrument should specify how to flag a respondent as being in the relevant group. For example, a sample member's information can be pre-populated at the onset of the survey, or the flag can be determined by a response to a previous question.

- **Fills** give the programmer a word or phrase that should populate in the questionnaire based on predetermined criteria. For example, the software could be programmed to populate "he/she" for proxy interviews so a sample member's name does not have to be repeated over many items. Fills can also be generated based on respondents' answers, for example by populating follow-up questions with the name of a job or employer the respondent gave earlier in the survey.
- **Response option values.** Each response option should be linked to an alpha or numeric value, such as "1" for yes and "0" for no, or a "D" for don't know. Having these labels clearly marked helps ensure programmers link response options to descriptors that will later be used for statistical analysis.
- **Text field character limits.** Where an instrument contains open-ended questions, programming specifications should include a limit for the number of characters to allow in the response. For web surveys, respondents will infer how much detail is expected of their response to the question based on the size of the text box. (Smaller boxes imply that brief responses are being sought.)
- **Checks to ensure data quality.** Computer-assisted interviews provide an opportunity to integrate quality checks to prevent collection of poor quality data. Allowable ranges should also be clearly specified. Examples could include a maximum value of seven days for responses to a question about number of days worked per week. Programming specifications can include soft or hard checks, or both. Soft checks alert interviewers or respondents when a response does not fall within the allowable range of values or when an item of critical importance has been left blank, but they allow a respondent to proceed to the next item even if the issue is not rectified. All items flagged as critical to the analysis should include soft checks when the item has been left blank. Hard checks prevent the respondent from continuing until the issue is rectified, and should be used sparingly—if at all—to respect respondents' right to skip questions they do not want to answer. Hard checks can also frustrate respondents and lead to early termination of the survey.

5. Why is it important to test the programmed instrument, and what does testing entail?

Rigorous testing of programmed instruments is critical to ensure data quality, and must be done before evaluators pre-test the instrument with actual respondents (discussed in the next section). Testing computer-based instruments against the programming specifications ensures that questions are: (1) being asked of the right subgroups, (2) presented on the screen in a way that matches the specifications, (3) presenting response options in the intended format (for example, "check one" and "check all that apply"), and (4) following the intended skip patterns or routing paths. For CATI surveys, testing will focus exclusively on the software being used to deploy the interviews. For web surveys, testing should also simulate the experiences of different types of respondents, who can access the survey from different kinds of devices and different Internet browsers. Testing should therefore include practice cases on different devices (such as

smartphones, tablets, and desktop computers), different web browsers (such as Firefox, Explorer, or Chrome), and assistive technologies used by people with disabilities.[38]

C. Testing survey instruments and procedures

Once the instrument is drafted, pre-testing provides an important opportunity to improve it (Presser et al. 2004; Beatty and Willis 2007). This section discusses the benefits of pre-testing, the logistics involved, and typical uses of pre-test data, concluding with a brief discussion of pilot testing, which helps evaluators test survey protocols and processes.

1. Why is pre-testing important?

Pre-testing gives evaluators an opportunity to:

- **Collect feedback from the study population.** Pre-testing allows evaluators to practice administering the interview, or observe self-administration (by web or on paper), with people who would be typical members of the study population. In conducting pre-test interviews, evaluators can assess whether respondents understand each item as evaluators intend it to be understood and whether any questions are particularly sensitive or difficult to answer. Evaluators can also assess how participants complete the questionnaire—for example, whether they read instructions, follow skip patterns correctly, or leave certain items blank.
- **Check sequencing and flow of items.** Pre-testing can help evaluators assess earlier decisions about item sequencing. This is especially important because items placed earlier in the survey can affect the responses to later questions. Assessing sequencing can also conserve programming resources if evaluators pre-test a pencil-and-paper questionnaire before developing programming specifications.
- **Collect or confirm burden estimates.** Pre-testing enables states to collect data on the amount of time necessary for respondents to complete the survey instrument in any mode. If the survey takes substantially longer to complete than anticipated, and the state is not willing to eliminate any questions, the pre-test findings give states a chance to refine the evaluator's budget and data collection plans, including the time and staff needed to administer that wave of the survey.
- **Verify that programmed instruments work as intended.** Pre-testing can provide a final check to ensure that programming and software for computer-assisted modes are working as intended. (If a computer-assisted mode is planned, two rounds of pre-testing may be beneficial—one to improve the instrument before programming and one to assess how well programmed instruments work.)

2. What are the logistics involved in a pre-test?

At least one round of pre-testing should be conducted in the mode(s) the survey will be administered in. If the survey is designed to be administered by an interviewer, it should be tested with an interviewer asking questions and recording responses. If it is designed to be self-

[38] For more information about testing web surveys for accessibility, see Matulewicz (2008).

administered, with responses recorded on paper, evaluators should watch respondents fill out the paper instrument in a face-to-face setting.

When conducting pretest interviews face-to-face, interviewers can also make note of body language or facial expressions, which can reveal which items touch on sensitive subjects or confuse respondents.[39] For example, people who find an item difficult to answer might widen their eyes, take and expel a deep breath of air, or place their hand to their head. Although interviewers can observe some of these reactions when conducting a pre-test interview by telephone, they have more opportunities for observation when they can see the respondent.

After the pre-test interview, researchers should debrief respondents on their experience overall and their understanding of or reaction to particular items. It is helpful to develop a debriefing interview guide in advance. This can help evaluators capture feedback on specific items, such as newly worded questions, in a systematic way.

3. How can evaluators use pre-test findings?

Evaluators should review pre-test results and decide which findings merit changes to the draft instrument. Changes can include reducing the item count, reordering items, revising item wording (when states do not plan to compare the results to results from the source survey), adding or dropping response categories, improving skip pattern logic, adding transition text, or adding text before sensitive items to (1) let respondents know why the information is being requested (how the data will be used) and (2) reiterate that the data will be kept private (that is, the evaluation contractor will maintain confidentiality). When a specific item is found to be particularly sensitive, states might also want to revisit the value of the item relative to the survey overall, especially if they are conducting the survey in self-administered formats, because respondents may break off the interview entirely if they think the questions are too personal.

4. What is a pilot test, and how can it help improve the design of the survey?

In contrast with a pre-test, which tests the survey instrument itself, a pilot test assesses the survey process overall, starting with invitations to respond and ending with calculation of participation rates and analysis of data quality across modes. Pilot tests are a kind of dress rehearsal ahead of the field period, and the data are typically not included in the final survey data file. They can provide insights that evaluators can use to modify survey-related mailings, scripts used by interviewers, and interviewer training materials. They can reveal greater participation in some modes over others, suggesting ways to refine survey plans and reallocate resources. Pilot tests, like pre-tests, also enable evaluators to assess an instrument's quality. Evaluators can use data generated during the pilot test to check for rates of missing values and adherence to skip patterns. Evaluators would also check the survey responses to check for social desirability bias or measurement issues across modes.

Not all states will have the resources to conduct a pilot test in addition to pre-test interviews. As an alternative, states can consider using a "soft launch." In this approach, evaluators begin the

[39] Face-to-face pretesting can be more challenging given travel requirements. Because Medicaid beneficiaries often face transportation challenges, evaluators may need to recruit a larger number of participants for in-person interviews than they would for telephone interviews to generate the same number of completed pretest interviews.

field period with a small cohort of cases to ensure no major issues arise with the questionnaire or data collection systems. For example, if a given wave of data collection plans to release 20,000 cases, the soft launch might begin with 200 cases. Unlike a pilot test, survey responses collected during the soft launch are often included in the file for analysis as long as there are no significant quality issues that warrant changes to the instrument.

IV. FIELDING THE SURVEY

Successful fielding of a beneficiary survey requires careful planning and skilled survey management. The goal of this chapter is to help states and their evaluators incorporate strategies to collect high quality data that can support planned analyses. These strategies include managing the sampled cases (Section A); operationalizing complex sample releases across the survey field period (Section B); minimizing sample attrition and unit nonresponse (Section C); and monitoring for quality (Section D).

A. Importance of sample management

Sample management is critical to ensure appropriate contact, routing, and follow-up with each sample member, as well as accurate record-keeping.

1. How should evaluators manage the sample from the initial draw to the final wave of data collection?

States should ensure that independent evaluators have experience with the following sample management functions, which are facilitated by using sample management software:

- **Organizing case records,** including (1) identifying beneficiaries selected for the survey from state administrative records, along with data about them that are relevant to the sampling plan, such as sex, race/ethnicity, or location of residence; (2) activating the survey for the right sample members at the right time (also called "releasing sample," as discussed later in this chapter); and (2) linking each sample member to the appropriate group (such as comparison or demonstration), which could receive a unique set of questions.
- **Supporting efficient data collection**, including (1) flagging language preferences for survey outreach and administration; (2) updating contact information, including mailing address, telephone number(s), or email; (3) recording contact history and outcome of locating efforts; (4) tracking participation status for each survey wave, including variable(s) such as response versus nonresponse; mode; and language; (5) and sending follow-up messages or reminder mailings to nonresponders as scheduled.
- **Facilitating appropriate actions for cases that complete an interview, become ineligible, or withdraw from the study.** This includes (1) tracking sample members who become ineligible, withdraw, or change group assignment (such as from current beneficiary to former beneficiary); (2) ensuring that sample members who are deemed ineligible or have withdrawn from the survey are not contacted in subsequent waves; (3) ending follow-up for sample members who complete the survey.

2. How does tracking survey response help evaluators measure progress toward survey goals?

Once a sample is released for data collection, the survey team will manage the sample and track progress toward survey goals by assigning a disposition code to each sample member. The disposition can, and often will, change over time as the field period progresses. The way that survey researchers use disposition codes can vary; for example, some will simply assign the status "complete," whereas others include codes for different types of completes (such as by mode or by proxy). These codes act as survey monitoring data to help the survey team assess

overall progress and flag low response rates in particular subgroups. This is especially important to ensure balanced response rates for demonstration and comparison groups. Setting up tracking systems helps evaluators catch problems early and enables them to mitigate potential nonresponse bias before the end of the field period, possibly avoiding the need to conduct formal nonresponse bias analysis later.

At the end of each wave of data collection, evaluators should produce a record of the final status for each sample member. Final statuses fall into the following categories: (1) completed interviews, (2) eligible nonparticipants; and (3) confirmed ineligibles or sampled individuals who did not provide enough information to confirm their eligibility.[40] Using these data, states can compute rates of cooperation and refusal as well as survey response rates. States conducting longitudinal surveys can combine these data with subgroup analysis to refine data collection plans for subsequent waves. For example, nonresponding sample members in one wave could be offered a bigger incentive or receive telephone follow-up sooner than other sample members do in the next wave.

As noted in section II.C, states should follow guidance from the American Association for Public Opinion Research (AAPOR) on computing response rates, and should document which AAPOR response rate was used when presenting the final response rate in any resulting reports.[41] AAPOR's formal guidance notes that "although response rate information alone is not sufficient for determining how much nonresponse error exists in a survey, or even whether it exists, calculating the rates is a critical first step to understanding the presence of this component of potential survey error. By knowing the disposition of every element drawn in a survey sample, researchers can assess whether their sample might contain nonresponse error and the potential reasons for that error."[42]

3. How does the tracking mode of completed surveys support sample management and prepare for future waves of the survey?

Once states decide to deploy more than one mode, they should develop plans to track respondents' mode of participation for each wave of the survey. This makes it possible to analyze the data by mode, allowing evaluators to determine whether there is bias in relation to mode of participation or, viewed differently, whether subgroups of interest are more likely to respond via a particular mode. For example, evaluators could assess whether employment rates look different for those who were interviewed by telephone compared to those who self-administered the survey. This would tell them whether beneficiaries who successfully transitioned to employment (or did not) are more effectively reached by one mode. In addition, in longitudinal surveys, mode choice from one wave can be used to inform planning for later waves. If, for example, response by web was minimal in Wave 1, evaluators might shorten the time between the start of Phase 1 (completed on the web) and the start of Phase 2 (mailed on paper).

[40] For additional guidance on defining final disposition codes, see https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf.

[41] AAPOR provides a tool known as the "response rate calculator" (version 4.0) which is free and available to the public. The tool is posted here: https://www.aapor.org/Standards-Ethics/Standard-Definitions-(1).aspx.

[42] See https://www.aapor.org/Standards-Ethics/Standard-Definitions-(1).aspx.

B. Survey field periods with more than one release of sample

The overall length of the field period should reflect whether the state plans to release more than one set of sampled cases. This approach has several advantages, but it also increases the complexity of the sample management task.

There are several reasons why evaluators might want to release more than one set of sampled cases. As noted in Section II.B, a rolling release for a survey of newly enrolled beneficiaries makes it possible to include people with seasonal enrollment, as well as those who churn on and off the enrollment rolls during the course of a year. Additional releases of sample could be needed to achieve the target number of completed cases needed for analysis, either because of sample attrition or because initial response to the survey was lower than anticipated. Releasing more than one set of sampled cases can also help evaluators manage the number of surveys they are fielding in a given mode at one time if there are limits on evaluators' capacity.

The decision to release more than one set of sampled cases should be approached with caution, as it can have a negative impact on the overall response rate for the survey because more sampled cases are contributing to the denominator, and not all will result in completed cases. States and their evaluators should also pay careful attention to response by survey strata, or subgroup of the population, to ensure the additional cases help fill gaps opened by lagging response rates or attrition. In addition, evaluators will have to decide whether to include in the final data file cases in which sample members need more than one session to complete an interview or return a questionnaire after follow-up for that cohort has ended.

Table IV.1 shows a survey with two sample cohorts. Evaluators release the sample for Cohort 1 and conduct all of the follow-up across the first 12 weeks of the overall field period. In Week 9, evaluators release Cohort 2. The field period for Cohort 2 runs for 12 weeks as well, but takes place in Weeks 9 through 20 of the overall field period. Each cohort progresses through the phases of data collection in the same way—for example, each receives invitation letters in Week 1 of its cohort-specific field period, a reminder postcard in Week 2.5, and reminder calls in Weeks 3 through 12).

Table IV.1. Illustration of a 20-week field period with two sample cohorts

Overall field period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Week of data collection: Sample Cohort 1	1	2	3	4	5	6	7	8	9	10	11	12								
Week of data collection: Sample Cohort 2									1	2	3	4	5	6	7	8	9	10	11	12

Note: Shaded cells are months in which there is no data collection for the sample cohort.

Regardless of how long the field period is, evaluators should develop specific criteria for when they will stop contacting particular sample members. For example, each sample member might receive up to 10 attempts by telephone before evaluators move to alternate means of contacting (or locating) these sample members. Categorizing cases as having exhausted all

outreach can also be useful in informing the number of cases needed for a supplemental release of sample.

C. Strategies to minimize nonresponse and sample attrition

As noted in Section II.C, high response rates are difficult to achieve in surveys of Medicaid beneficiaries, but states can use a number of strategies to achieve their target rates and minimize attrition between waves of longitudinal surveys. These strategies include responding to sample members' questions or concerns about the survey and having consistent survey branding, advance notification mailings, effective nonresponse follow-up, interim contacts, and strategies for locating hard-to-reach sample members.

1. How can evaluators promote participation by sharing information on the survey?

Answering sample members' questions. Some sample members might be reluctant to participate at first, but willing to consider it if evaluators can answer their questions or assuage their concerns. Evaluators could establish a toll-free telephone number, a survey email address, and/or a website to help them. Information about how sample members can learn about the survey should be displayed clearly on all survey mailings. Evaluators should choose telephone numbers or an email address that will ensure the questions reach people who are knowledgeable about both the study and the demonstration. This is important because beneficiaries might be unaware of the demonstration or even their enrollment status until they are asked to participate in a survey. Questions about enrollment status should be directed to the Medicaid agency. If the survey is fielded in more than one language, and evaluators provide a telephone number for questions, bilingual staff should be available to take calls in the survey's available languages. Finally, websites are useful because they can provide information and lend legitimacy to the survey.[43]

Insights from the field

"It has been important to have a human being available to clarify terms. Respondents may not know the difference between terms like premiums and co-payments, or may not understand demonstration terms used by the state. Sometimes the survey staff were the first explainers of demonstration policies. Clarifying terminology would have been much harder with a paper survey."

– University of Iowa survey research team

For any of these strategies, evaluators should prepare staff (or website text) to address several common questions or concerns from sample members and respondents. These include questions about the purpose of the study, whether participation is voluntary, whether the decision to participate affects Medicaid (or other) benefits, what incentives are available for taking part, how long the survey will take, how and where to get started, how answers will be used, and how and whether respondent information will be kept private. As evaluators prepare responses, they should take care to apply the same good practices they applied in the questionnaire itself, such as using plain language, avoiding technical jargon, and keeping the reading level as close to fifth grade as possible.

Messaging through community-based organizations. Community-based organizations that serve Medicaid beneficiaries can be powerful allies in disseminating information about the

[43] An example of an informational website can be found at https://2020census.gov/en.

survey and assuring sample members of its legitimacy. They can publicize the survey in newsletters or other notifications they share with the populations they serve. Evaluators can also give these organizations one-page handouts explaining the survey.

2. How can branding and consistent messaging encourage responses?

Survey branding. A professional look and feel helps demonstrate the legitimacy of the survey to potential respondents, who are often inundated with solicitations for products or services by mail and email.[44] The name, logo, and color scheme help sample members recognize the survey as legitimate and distinguish survey-related mailings from junk mail or bills. For longitudinal surveys, it is important to keep the look and feel of survey messaging the same across all survey waves because this helps sample members recognize the survey and respond to the inquiry. In turn, this helps to mitigate sample attrition.

Branding begins with the name or title, which can be the same for each wave of data collection or modified to include the wave or timing of the survey. For example, states could use titles like "Health in Our State: Wave 2" or "Health in Our State: 18-Month Survey." States could create logos specifically for the survey, although logos of the state department of health or Medicaid agency can also be used to establish credibility. If states use a logo, it should be displayed on study mailing materials, such as outer envelopes, to increase the likelihood that the mail will be opened. States can also use logos in study emails, making them into a hyperlink that routes email recipients to the state website or to the introduction page of a web-based survey.

Notification letters. People selected for the survey are more likely to respond if they receive an advance letter (or postcard) (De Leeuw et al. 2007). This mailing should be kept brief, and should include the purpose of the study, what is being asked of the sample member, the sponsoring agency, any risks and benefits to participation, and a closing statement of thanks. Most importantly, this letter should include a request for an action on the part of the recipient, such as logging into a web survey or calling a telephone center to complete or schedule an interview. Action-focused letters can reduce survey costs because they reduce the number of sample members who need reminders later (Johnson et al. 2017). This is a best practice even for web surveys; advance notification letters let sample members know that an email invitation is coming. This helps them distinguish the email invitation from unwanted solicitations or spam (Dillman et al. 2014).

3. What kinds of incentives work best?

Monetary incentives can be a powerful motivator that encourages response, especially for Medicaid beneficiaries, who typically have low incomes. Incentives increase response rates without compromising data quality (Singer and Kulka 2000). In particular, prepaid incentives, such as a $2 or $5 bill in the invitation letter, can bolster response, as sample members feel compelled by social exchange (Church 1993; Singer et al. 2000). These incentives can be expensive, but can offset even more expensive follow-up later in the field period, and can be more effective than post-paid incentives because they thank potential respondents in advance and deliver instant gratification for their willingness to complete the questionnaire or interview. In

[44] For more information on the importance of professional-looking materials, and on survey industry best practices for mail, web, and mixed mode surveys generally, see Dillman et al. (2014).

contrast, several weeks may elapse between the time when respondents complete their interviews and the time when they receive a post-paid incentive. States using post-paid incentives should use gift cards or cash instead of checks, because some sample members might not have access to banking services. States can also consider using differential incentives to motivate sample members to respond in particular modes (such as web). For example, the invitation mailing could inform sample members that they can receive a $15 gift card for completing the survey online or receive $10 for completing it by telephone (Mooney et al. 2012).

4. How should evaluators approach nonresponse follow-up?

Nonresponse follow-up strategies often include sending several types of reminders on a staggered basis across the field period. Varying the type of reminder can be effective because different kinds of nonresponse follow-up will work for different segments of the population. States should therefore plan to include a variety of strategies that fit within the survey budget and timeline. Common strategies include reminder letters, postcards (which convey a message without the barrier of the envelope), telephone calls, emails, and text messages.[45] In-person visits to nonresponding sample members' home addresses can also be used, though this follow up strategy may be expensive.

Different follow-up methods have different advantages. For example, the advantage of postal mail is that envelopes can be marked with "address service requested." This could allow the evaluator to receive updated address information from the postal service if the sample member is no longer living at the address of record. The advantage of telephone or in-person reminders from trained data collection staff is that nonrespondents can interface with a person who can immediately address their concerns and convert potential refusals.

In general, nonresponse follow-up is most effective when staggered or paced across the field period. Because some members of this population have intermittent telephone service, re-contacting sample members at telephone numbers previously deemed non-working can bear fruit, as some will come back in service when bills are paid or when service is replenished through public benefit programs. Evaluators should also allow time for multiple contacts to a given telephone number or address at different times of day and days of the week. Regardless of mode, it is important for evaluators to track the number of contacts made for each sampled case. It may be helpful to distinguish easier-to-reach respondents from those who were more difficult to reach to assess whether their responses or outcomes differ.

Nonresponse follow-up should be planned in advance. To ensure that the follow-up plan is carried out as intended, evaluators can create a table that shows, by week of the field period, what type of nonresponse follow-up will be used. This table can also help evaluators identify (1) whether there is a wide enough variety of reminders and (2) whether the spacing between reminders is long enough to allow sample members to respond before they get another reminder.

[45] States should ask respondents for their explicit permission to contact them via text message, acknowledging that such correspondence may result in data usage or other charges. Text messages should only be sent to sample members who have explicitly consented to receive them. Accordingly, text messaging would not be a viable follow-up strategy in the initial wave of a survey.

5. How can interim contacts keep sample members engaged in longitudinal surveys?

States planning longitudinal surveys should incorporate interim mailings into the overall communication plan. Interim mailings play a critical role in maintaining connections with sample members between waves of the survey, and therefore help minimize attrition. Interim contacts can take the same format as nonresponse reminders, such as text messages, postcards, or letters. If several months will elapse between survey waves (for example, one round is conducted at 18 months and the next at 36 months), states should plan on using at least one interim mailing between each wave. If a state is sending only one mailing, it should be placed at the midpoint of the time period between the two waves. If the state is sending more than one mailing, they can be spaced out evenly across this time period.

In addition to helping sample members stay connected to the study, interim mailings help ensure the survey is using the best available contact information for the sample members. These mailings give the sample members a chance to respond with updated contact information, or can allow evaluators to seek updated addresses from the postal service. Survey staff can then update the sample management database with new addresses. These efforts help to ensure a strong start to the field period for each successive wave.

6. What strategies can states use to locate sample members when data from administrative records are no longer current?

Using different locating strategies can help evaluators collect responses from hard-to-reach sample members, and can also reduce attrition between waves of longitudinal surveys. Locating strategies vary in terms of cost and impact. Experience with locating is an important qualification for the organization contracted to administer state surveys—if states' independent evaluators do not have this experience, evaluators should plan to contract with a separate firm that specializes in survey data collection.

Examples of effective locating strategies include:

- **Cross-checking sample members' addresses in batch format through the National Change of Address database and/or other address databases.** Address databases can provide updated, accurate mailing addresses to use for survey mailings if sample members moved after the state initially collected the contact information stored in Medicaid administrative records. These services are typically provided by vendors. Investing in this strategy helps states avoid sending advance letters to addresses that are no longer current, which can cause delays in contacting the sample member and wasted postal costs. Ensuring that advance letters are sent to the correct address also helps interviewers avoid placing cold calls to sample members who have not received their advance letters.
- **Individual-level searches using online tools such as web-based search engines, social media sites, or the Social Security Death index**. This approach is far more expensive than batch searching because it requires much more labor. It should be saved until other locating strategies, such as mailings or telephone follow-up, have been exhausted. This service is typically provided by the evaluator or the evaluator's survey contractor.
- **Requesting additional contacts for sample members in each wave of survey administration.** States using a longitudinal survey design should ask respondents for the name and contact information of a person who will always be able to reach them in the

future if they move or get a new telephone number. These secondary contacts can be added to the sample members' records in the sample management database and called upon, as needed, to help locate the sample member.

Results from any or all of these strategies should be continuously recorded in the study's sample management system.

7. How can states and evaluators support voluntary participation in the survey?

One of the hallmarks of high quality research is a commitment to the rights of study participants, including voluntary participation. For example, sample members should be allowed to skip questions they do not wish to answer, across any or all survey modes. Even though evaluators should strive to maximize response rates, sample members also have the right to decline to participate in a specific wave of the survey or in the survey overall. Survey invitation letters should include language about the voluntary nature of the study, along with the risks and benefits of participation.

Evaluators should also design a specific process for handling requests to withdraw from the survey, as well as clear, easy-to-follow instructions for making those requests. Documenting this process during the survey planning phase will support evaluators in carrying it out effectively, and could be helpful if the survey is reviewed by the state's Institutional Review Board, which seeks to protect human research subjects. The same process should be used for all individuals across the evaluation period. Evaluators should also establish a way to ensure that sample members who decline to participate in the survey overall are removed from the sample for later waves of the survey. When administering the survey, evaluators should use sample management software to track cases that withdraw—this informs attrition rates and ensures those cases receive no further outreach.

D. Monitoring the data collection process for quality

Monitoring for risks to data quality can help evaluators conserve study resources that would otherwise be spent addressing negative outcomes, such as missing or data poor quality data.

1. How can evaluators monitor for differential nonresponse by subgroup?

In addition to monitoring response rates for the survey overall, it is important to ensure response rates are similar for different subgroups (Schouten 2009). To monitor response by subgroup, states should identify the variables that reflect key characteristics of the sample—including demographic characteristics and membership in demonstration versus comparison groups—and ensure those are in the sample management database. These variables can also help determine the composition of replacement samples for later rounds of data collection or to make statistical adjustments for nonresponse at the end of each round of data collection (Groves 2006a; Schouten 2009).

2. How can evaluators monitor the quality of survey responses?

The quality of survey data can be affected by many issues, including item nonresponse; respondents speeding through a survey instrument without giving careful consideration to the questions; skip logic that is not followed as intended, resulting in missing data or undue burden

on respondents (Conrad et al. 2017); the introduction of bias in respondents' answers because interviewers are not following best practices; or even interviewer falsification of responses (Murphy et al. 2016). Staff monitoring data quality might spot problems early in the field period and fix instruments or give interviewers more training to address them. Independent evaluators should also check for overall rates of item nonresponse and nonresponse to items deemed critical to the analysis, as well as for potential variation in response to survey items by mode (for mixed mode surveys).

Because the type of response (self-report or proxy) impacts data quality, evaluators should track how respondents were helped in answering survey questions, including whether they completed the survey with a proxy or had some type of assistance, such as ad hoc translation, having the questions read aloud, or other help. Evaluators can analyze survey responses by response type to get a deeper understanding of whether responses differ significantly on outcomes of interest. Finally, survey staff should monitor rates of proxy interviews for the interviewers overall and by individual interviewer. If a particular interviewer has a higher-than-average rate of proxy interviews, supervisors can give feedback or additional training to ensure the interviewer follows criteria for proxy use and improve the interviewer's skills in promoting self-reporting.

This page has been left blank for double-sided copying.

V. CONCLUSIONS

Beneficiary surveys serve an important purpose in evaluations of section 1115 demonstrations, providing states and CMS with data that cannot be collected in other ways. In sharing their experiences, current and former beneficiaries help states deepen their understanding of the implementation and outcomes of Medicaid eligibility and coverage policies. Survey respondents' perspectives help demonstration states assess whether these policies have achieved their intended goals, and whether they result in unintended consequences for states' more vulnerable residents.

This guide gave an overview of considerations and best practices in designing and administering beneficiary surveys, but it does not substitute for the detailed planning effort states will need to undertake to conduct a high quality survey. CMS encourages states to collaborate with independent evaluators who have expertise in survey methodology and the organizational capacity to successfully administer surveys. States can use this guide to help them select evaluation partners who can support the functions described here—potential evaluators should be able to give states a clear plan for how they will design the instrument(s), prepare to field the survey, develop and implement a survey plan that falls within the study's timeline and budget, and monitor survey production statistics and data quality. Evaluators should also be able to give examples of studies they have conducted with Medicaid beneficiaries or similar populations in the recent past. Once states have selected an independent evaluator, the evaluator can help states provide the necessary plans and documentation to CMS.

Finally, CMS is available to provide technical assistance to demonstration states considering beneficiary surveys as part of their evaluation design, or those working through a particular survey-related challenge. Supporting states' survey efforts is consistent with CMS's goal to encourage rigorous evaluations of section 1115 Medicaid demonstrations to generate high quality evidence that informs Medicaid policy.

This page has been left blank for double-sided copying.

REFERENCES

Agency for Healthcare Research and Quality. CAHPS Health Plan Survey. Rockville, MD: U.S. Department of Health and Human Services, 2018. Available at http://www.ahrq.gov/cahps/surveys-guidance/hp/index.html. Accessed May 24, 2019.

Agency for Healthcare Research and Quality. CAHPS Home and Community-Based Services Survey. Rockville, MD: U.S. Department of Health and Human Services, 2018. Available at http://www.ahrq.gov/cahps/surveys-guidance/hcbs/index.html. Accessed May 24, 2019

Agency for Healthcare Research and Quality. CAHPS Hospital Survey. Rockville, MD: U.S. Department of Health and Human Services, 2018. Available at http://www.ahrq.gov/cahps/surveys-guidance/hospital/index.html. Accessed May 24, 2019

Agency for Healthcare Research and Quality. Supplemental Items for the CAHPS Health Plan Adult Survey 5.0: Utilization. Rockville, MD: U.S. Department of Health and Human Services, 2018. Available at http://www.ahrq.gov/cahps/surveys-guidance/item-sets/hp/suppl-utilization-hp-adult.html. Accessed May 24, 2019

Beatty, Paul C., and Gordon B. Willis. "Research Synthesis: The Practice of Cognitive Interviewing." *Public Opinion Quarterly*, vol.71, no. 2, January 2007, pp. 287–311.

Bentler, Suzanne, Peter Damiano, Elizabeth Momany, Brooke McInroy, Erin Robinson, and Mark Pooley. *Evaluation of the Iowa Health and Wellness Plan: Member Experiences in the First Year*. Iowa City, IA: Public Policy Center, The University of Iowa, 2015.

Brick, J. Michael, and Douglas Williams. "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys." *Annals of the American Academy of Political and Social Science,* vol. 645, January 2013, pp. 36–59.

Centers for Disease Control and Prevention. "Behavioral Risk Factor Surveillance System Survey Questionnaire." Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2016.

Centers for Disease Control and Prevention. "Behavioral Risk Factor Surveillance System Survey Questionnaire." Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2018.

Centers for Medicare & Medicaid Services. "Healthy Indiana Plan 2.0 Enrollee Survey." OMB Control Number: 0938-1300, expiration date 11/30/2019. Developed by Social & Scientific Systems, Inc. under contract to CMS.

Centers for Medicare & Medicaid Services. "Healthy Indiana Plan 2.0 Disenrollee Survey." OMB Control Number: 0938-1300, expiration date 11/30/2019. Developed by Social & Scientific Systems, Inc. under contract to CMS.

Centers for Medicare & Medicaid Services. Montana Health and Economic Livelihood Partnership Plan Beneficiary Survey: Enrollees. OMB Control Number: 0938-1332, expiration date 06/30/2020. Developed by Social & Scientific Systems, Inc. under contract to CMS.

Centers for Medicare & Medicaid Services. Montana Health and Economic Livelihood Partnership Plan Beneficiary Survey: Disenrollees. OMB Control Number: 0938-1332, expiration date 06/30/2020. Developed by Social & Scientific Systems, Inc. under contract to CMS.

Church, A. H. "Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis." *Public Opinion Quarterly,* vol. 57, no. 1, 1993, pp. 62–79.

Conrad, Frederick, Roger Tourangeau, Mick Couper, and Chan Zhang. "Reducing Speeding in Web Surveys by Providing Immediate Feedback." *Survey Research Methods*, vol. 11, no. 1, 2017, pp. 45–46.

Couper, M. P. "Web Surveys: A Review of Issues and Approaches." *Public Opinion Quarterly,* vol. 64, 2000, pp. 464–494.

Czajka, John L., and Amy Beyler. "Declining Response Rates in Federal Surveys: Trends and Implications." Final report submitted to the U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation. Washington, DC: Mathematica Policy Research, June 15, 2016.

De Leeuw, Edith. "Mixed Mode: Past, Present, and Future." *Survey Research Methods,* vol. 12, no. 2, 2018, pp. 75–89.

Dillman, Don A., Jolene D. Smyth, and Lea Melani Christian (Eds.) (c2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* Hoboken, NJ: John Wiley & Sons Inc.

Dillman, D. A., G. Phelps, R. Tortora, K. Swift, J. Kohrell, and J. Berck, et al. "Response Rate And Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR) and the Internet." *Social Science Research,* vol. 38, no.1, March 2009, pp. 1–18.

DiMaggio Paul, and Eszter Hargittai. "From the 'Digital Divide' to 'Digital Inequality': Studying Internet Use As Penetration Increases." Working Papers 47, Princeton University, Woodrow Wilson School of Public and International Affairs, Center for Arts and Cultural Policy Studies.

Dorr Goold, Susan, Jeffrey Kullgren, Sarah Clark, and Christina Mrukowicz. "Healthy Michigan Voices Beneficiary Survey Interim Report." Ann Arbor, MI: Institute for Healthcare Policy and Innovation, University of Michigan, 2016.

Groves, R. M. (2006a). "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly,* vol. 70, no. 5, 2006, pp. 646–675.

Groves, R. M., F. J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. *Survey Methodology* (second ed.). Hoboken, New Jersey: John Wiley and Sons, 2009.

Groves, R. M., and S.G. Heeringa, S. G. (2006b). "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society: Series A (Statistics in Society),* vol.169, no. 3, 2006, pp. 439–457.

Groves, Robert M. *Survey Error and Survey Costs*. Hoboken, NJ: John Wiley and Sons, 1989, 2004.

Harkness, J., and A. Schoua-Glusberg. "Questionnaires in Translation." *ZUMA-Nachtrichten Spezial,* vol. 3, 1998, pp. 87–126.

Heffetz, Ori, and Daniel B. Reeves. "Difficulty of Reaching Respondents and Nonresponse Bias: Evidence from Large Government Surveys." *The Review of Economic Statistics,* vol. 101, no 1, March 2019, pp. 176–191. doi:10.1162/rest_a_00748.

Jarrett, R. L. "Focus Group Interviewing with Low-Income Minority Populations: A Research Experience." In D. L. Morgan (ed.), *Successful Focus Groups: Advancing the State of the Art*, Newbury Park, CA: Sage Publications, 1993, p. 184.

Johnson, Amy, Ryan Callahan, Jesse Chandler, and Jason Markesich. "Using Behavioral Science to Improve Survey Response: An Experiment with the National Beneficiary Survey." In Focus Brief. Washington, DC: Mathematica, December 2017.

Kreuter, Frauke. "Facing the Nonresponse Challenge." *Annals of the American Academy of Political and Social Science*, vol. 645, January 2013, pp. 23–35.

Lavallée, P., and J.F. Beaumont. "Why We Should Put Some Weight on Weights." *Survey Insights: Methods from the Field, Weighting: Practical Issues and "How to" Approach.* Invited article, 2015. Available at https://surveyinsights.org/?p=6255. Accessed May 24, 2019.

Lynn, Peter (ed). *Methods in Longitudinal Surveys.* New York: Wiley, 2009.

Matulewicz, Holly H., and Jeff Coburn. "Universal Design for Web Surveys: Practical Guidelines." *Survey Practice,* November 2008.

Medway, Rebecca L., and Jenna Fulton. "When More Gets You Less: A Meta-Analysis of the Effect of Concurrent Web Options on Mail Survey Response Rates." *Public Opinion Quarterly*, vol. 76, no. 4, January 2012, pp. 733–746. Available at https://doi.org/10.1093/poq/nfs047.

Mooney, Geraldine, Cheryl De Saw, Xiaojing Lin, Andrew Hurwitz, and Flora Lan. "Influencing Mode Choice in a Multi-Mode Survey." Presented at the American Association for Public Opinion Research conference, Orlando, May 2012.

Murphy, Joe, Paul Biemer, Chris Stringer, Rita Thissen, Orin Day, and Y. Patrick Hsieh. "Interviewer Falsification: Current and Best Practices for Prevention, Detection, and Mitigation." *Statistical Journal of the IAOS*, vol. 32, no. 3, pp. 313–326.

National Center for Health Statistics. "National Health Interview Survey Questionnaires." Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2018. Available at https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm. Accessed May 24, 2019.

Presser, Stanley, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Jennifer M. Rothgeb, and Eleanor Singer. "Methods for Testing and Evaluating Survey Questions." *Public Opinion Quarterly*, vol. 68, no. 1, March 2004, pp. 109–130.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. "IPUMS USA: Version 9.0." Minneapolis, MN: IPUMS, 2019. Available at https://doi.org/10.18128/D010.V9.0. Accessed May 24, 2019.

Schouten, B., F. Cobben, and J. Bethlehem. "Indicators for the Representativeness of Survey Response." *Survey Methodology*, vol. 35, no.1, 2009, pp. 101–113.

Singer, Eleanor, and Richard A. Kulka. "Paying Respondents for Survey Participation." Ann Arbor, MI: The University of Michigan, Institute for Social Research, Survey Research Center, 2000.

Tourangeau, Roger, Frederick G. Conrad, and Mick P. Couper. *The Science of Web Surveys*. Oxford: Oxford University Press, 2013.

U.S. Census Bureau. "Current Population Survey, January 2018: Displaced Worker, Employee Tenure, and Occupation Mobility Supplement Technical Documentation." Washington, DC: U.S. Census Bureau, 2018. Available at https://www2.census.gov/programs-surveys/cps/techdocs/cpsjan18.pdf. Accessed May 24, 2019.

U.S. Census Bureau. "Current Population Survey, March 2018: Annual Social and Economic (ASEC) Supplement." Washington, DC: U.S. Census Bureau, 2018. Available at https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar18.pdf. Accessed May 24, 2019.

U.S. Census Bureau. "Survey of Income and Program Participation: 2014 Panel Complete Technical Documentation." Washington, DC: U.S. Census Bureau, 2014. Available at https://thedataweb.rm.census.gov/pub/sipp/2014/2014SIPP_Metadata_AllSections.pdf. Accessed May 24, 2019.

U.S. Census Bureau. "Survey of Income and Program Participation: 2008 Panel Complete Technical Documentation." Washington, DC: U.S. Census Bureau, 2008. Available at https://www.census.gov/programs-surveys/sipp/tech-documentation/complete-technical-documentation/complete-documents-2008.html. Accessed May 24, 2019.

APPENDIX A: SUMMARY OF SURVEY DESIGNS AND RESPONSE RATES FOR RECENT SURVEYS USED IN SECTION 1115 EVALUATIONS

Survey/State	Populations	Wave of survey, year	Total initial sample[a]	Number of completed cases[b]	Incentive	Mode(s)	Length of instrument or interview	Duration of field period	Languages	Nonresponse follow-up
Longitudinal surveys										
Kentucky	Individuals enrolled in Medicaid as of February 1, 2018 and eligible for KY HEALTH	Baseline, 2018	54,720	9,464	$1 pre-paid, $25 post pay cash	Web, telephone	Mean = 32 minutes	19 weeks, April 2018 to September 2018	English, Spanish	Mailed advance letter with $1, mailed reminder letter, telephone follow-up
Healthy Michigan Voices survey of HMP enrollees	≥12 months HMP enrollment (first enrolled 4/14 to 10/15)	Initial (not baseline), 2016	9,350	4,090	$25 gift card, post pay	Telephone	12 to 96 minutes, median = 22 minutes	January 2016 to November 2016	English, Arabic, or Spanish	Mailed introductory packet + up to 2 reminder mailings; minimum of 3 calls[c]
Healthy Michigan Voices Follow-Up	Individuals who completed the 2016 HMV Enrollee Survey and consented to be contacted for follow-up	Follow-up, 2017	3,957 (number that consented to be recontacted at time of first survey)	3,104	$25 gift card, post pay	Telephone	8 to 63 minutes, median = 18 minutes	March 2017 to January 2018	English, Arabic, or Spanish	Mailed introductory packet + up to 2 reminder mailings; up to 3 email and/or text reminders if info provided at initial survey; minimum of 3 calls[c]
Healthy Michigan Voices Follow-Up 2	Individuals who completed the 2017 HMV Follow-Up Survey and consented to be contacted for follow-up	Follow-up, 2018	3,070 (number that consented to be recontacted at time of first follow-up survey)	2,608	$25 gift card, post pay	Telephone	9 to 93 minutes, median = 17 minutes	June 2018 to January 2019	English, Arabic, or Spanish	Mailed introductory packet + up to 2 reminder mailings; up to 3 email and/or text reminders if info provided at initial survey; minimum of 3 calls[c]
Michigan survey of individuals no longer enrolled in HMP (NLE)	≥10 months of HMP enrollment in 12-month period followed by ≥6 months not enrolled in HMP or another Medicaid program	Initial, 2016-2017	4,750	1,123	$25 gift card, post pay	Telephone	n.a.	October 2016 to March 2017	English, Arabic, or Spanish	Mailed introductory packet + up to 2 reminder mailings; minimum of 3 calls[c]

Table A.1 (*continued*)

Survey/State	Populations	Wave of survey, year	Total initial sample[a]	Number of completed cases[b]	Incentive	Mode(s)	Length of instrument or interview	Duration of field period	Languages	Nonresponse follow-up
Michigan survey of individuals no longer enrolled in HMP (NLE Follow-Up)	Individuals who completed the initial survey of individuals no longer enrolled in HMP and consented to be contacted for follow-up	Follow-up, 2018	1,077 (number that consented to be recontacted at time of first survey)	735	$25 gift card, post pay	Telephone	6 to 30 minutes, median = 11 minutes	April 2018 to July 2018	English, Arabic, or Spanish	Mailed introductory packet + up to 2 reminder mailings; up to 3 email and/or text reminders if info provided at initial survey; minimum of 3 calls[c]
Michigan survey of new HMP enrollees	Initial HMP enrollment 5 months prior to sampling month	New enrollee, 2017	1,750	607	$25 gift card, post pay	Telephone	8 to 40 minutes, median = 15 minutes	June 2017 to December 2017	English, Arabic, or Spanish	Mailed introductory packet + up to 2 reminder mailings; minimum of 3 calls[c]
Cross-sectional surveys										
Healthy Michigan Voices survey of HMP enrollees (HMV cohort 2)	≥12 months HMP enrollment (first enrolled 1/16 or later)	New sample, 2018	8,500	2,602	$25 gift card, post pay	Telephone	14 to 82 minutes, median = 23 minutes	January 2018 to January 2019	English, Arabic, or Spanish	Mailed introductory packet + up to 2 reminder mailings; minimum of 3 calls[c]
Iowa Wellness Plan survey	Enrollees in plan ≥6 months	2014	4,050	1,101	$2 pre-paid + response-dependent gift card lottery	Paper (with web option)	107 questions, 30 minutes	12 weeks	English only	Postcard reminder, mailed second survey
Iowa Marketplace Choice survey	Enrollees in plan ≥6 months	2014	2,700	691	$2 pre-paid + response-dependent gift card lottery	Paper (with web option)	107 questions, 30 minutes	12 weeks	English only	Postcard reminder, mailed second survey
Iowa Dental Wellness Plan survey	Newly eligible DWP enrollees (enrolled 6 months)	2014	4,800	1,260	$2 pre-paid + response-dependent gift card lottery	Paper (with web option)	89 questions, 20 minutes	12 weeks	English only	Postcard reminder, mailed second survey

[a] Total initial sample is the total number sampled, including those who turn out to be ineligible for the survey, have a nonworking phone number or incorrect address, or refuse.

[b] We present the number of cases (surveys) completed rather than response rates due to variation in methods for calculating response rates. Evaluators should follow guidance from the American Association for Public Opinion Research on calculating response rates and should report the calculation used.

[c] Calls were made on at least 2 different days/times; if calls were answered by automated recording, survey team recontacted at beginning of next month, when phone minutes typically replenish.

DWP = Dental Wellness Plan; HMP = Healthy Michigan Plan (Michigan's section 1115 demonstration); HMV = Healthy Michigan Voices survey; n.a. = information not available.

APPENDIX B: CONDUCTING POWER CALCULATIONS

This appendix provides a basic description of how evaluators make power calculations. This description is intended to give states an understanding of the process and to support conversations between states and their evaluators. It is not intended to be a step-by-step guide. Power calculations specific to the state's planned beneficiary survey(s) should be included in the evaluation design submitted to CMS.

Evaluators use power calculations to determine the minimum sample size needed to support statistically sound analyses of demonstration policies and detect subgroup effects given their study design. A sample's power can be calculated in two ways: (1) calculating the sample size needed[46] for a given minimum detectable difference (MDD) or confidence interval (CI) half width, or (2) calculating an MDD or CI half width for a given sample size. The MDD is the difference between the means of two population groups that evaluators would like to be able to detect as significantly different from zero, such as the difference between the means for demonstration and comparison groups. The CI half width is the value that is added to or subtracted from a point estimate, such as a sample mean, when calculating a confidence interval; this is sometimes called a "margin of error." The first method—calculating the sample size needed for a particular MDD or CI half width—is preferable for states planning a beneficiary survey because evaluators typically start with decisions on the size of the effects they want to be able to detect as significant, and then proceed with designing the sampling plan. Power calculations require the following information:

1. **The assumed Type I error.** The Type I error is defined as the probability of incorrectly rejecting a null hypothesis if it is true (that is, finding a false positive). For example, if states are evaluating a Medicaid demonstration that is designed to increase employment levels for demonstration beneficiaries compared to other Medicaid beneficiaries, a Type I error is the probability of concluding that the demonstration increased employment when it in fact did not. In the context of a confidence interval, the Type I error is the probability that the calculated interval does not contain the true mean. The Type I error is often set to 5 percent.
2. **The assumed power.** The power is the probability of finding a significant result when there is truly a difference of a certain size; more precisely, correctly rejecting a false null hypothesis.[47] In the example above, the power is the probability that we conclude that a demonstration increased employment levels when it actually did increase them by a certain amount. The assumed power is often set to 80 percent. Note that power is not pertinent to confidence intervals.

[46] Note that "sample size" here means the number of completed surveys. In the design phase, when planning the number of sample cases that must be released, this number must be inflated to account for the expected response rate to determine the number of cases to sample, where the response rate is the proportion of sample cases that result in a completed interview.

[47] In other words, avoiding a Type II error, which refers to not rejecting the null hypothesis when the alternative hypothesis is true.

3. **The population variance**

   a. For an outcome measure that is a proportion or percentage, such as the proportion of former Medicaid beneficiaries who report that they are enrolled in commercial coverage, the variance is calculated as a function of the proportion with the attribute of interest (for example, the proportion of those with commercial coverage). The most conservative value that can be used (the one that results in the largest possible variance) is 0.5, but if the proportion can be reasonably projected, then it is better to use the projected value.

   b. For a continuous outcome, such as the number of hours worked per week, there is no maximum variance. Evaluators can either use a value for the variance that can be expected from previous studies, or, if calculating an MDD or CI half width, evaluators can express the result in terms of the number of standard deviations (essentially setting the variance equal to 1).

4. **The design effect.** A design effect is an adjustment factor that accounts for increased variance due to the structure of the sample design when it deviates from a simple random sample. The "effective sample size" is a measure of the equivalent simple random sample size in terms of precision and is calculated as the sample size divided by the design effect. There are two types of design effects to consider: one related to unequal weighting and the other related to clustering (when applicable). Evaluators can calculate the effective sample size by taking the sample size and dividing it by the product of these two design effects. The larger the design effect, the smaller the effective sample size.

   a. The design effect caused by unequal weighting incorporates (1) unequal weights that are a result of beneficiary subgroup sample sizes that are not proportional, or to a sample selection procedure that is not an equal probability sample;[48] (2) unequal weights that are a result of post-data collection adjustments, including nonresponse adjustments (see Section II.C); and (3) unequal weights that are a result of other adjustments, including, for example, weights that ensure treatment and control groups are equally calibrated. Evaluators can sometimes estimate one or more of the design effects from unequal weighting at the design phase by using information from a previous similar study, but some design effect components can be estimated using direct calculations.

   b. If clustered sampling will be used (for example, first sampling geographic areas or managed care organizations, and then sampling beneficiaries from within that primary sample), the design effect measures the loss of precision caused by this type of design. To calculate the design effect as a result of clustering, evaluators need to know or estimate the intracluster correlation coefficient (ICC), which is a measure of the homogeneity of a particular outcome variable within clusters, and ranges from 0 (no homogeneity within clusters compared to between clusters) to 1 (perfect homogeneity within clusters).[49] Ideally, the ICC is estimated from previous similar studies. If no previous data are available, evaluators can choose from within this range (0 to 1), where the estimate should be closer to 0 if evaluators do not expect observations to be more

[48] Each individual in the sample population must have a quantifiable probability of being selected. This allows evaluators to calculate sampling weights. In an equal probability sample, the different units within each stratum (or subgroup) in the population have equal probabilities of being chosen to be part of the sample.

[49] The ICC is the proportion of the total variance explained by the variance between clusters.

similar within clusters than between clusters, and higher if observations within clusters are expected to be more similar.

5. **Desired margin of error or effect size, or, alternatively, desired sample size.** As noted, starting with a CI half width allows evaluators to calculate the needed sample size (number of completed surveys). Similarly, to calculate the sample size needed to detect an underlying difference between two groups, evaluators should specify the difference in the population means between the two groups being compared (the MDD). It is also possible to start with the desired sample size and then calculate the CI half-width of a confidence interval. Likewise, starting with the desired sample size for each of the two groups being compared allows evaluators to calculate the MDD.

Evaluators will use all of these factors to calculate a sample size given an MDD for a comparison of two means or two proportions from equal-sized groups. This approach allows evaluators to derive the number of cases needed for the treatment and control (or demonstration and comparison) groups, assuming the same number of cases in each group.

States and their evaluators should also be aware that this general description of sampling and power calculations will need to be adapted for certain analysis types. For example, for difference-in-difference analyses, it will be necessary to incorporate four sample sizes in the MDD formula, two for the treatment group (before and after), and two for the comparison group (before and after).

This page has been left blank for double-sided copying.

APPENDIX C: EXAMPLE ITEM WORDING FOR MEASURES LISTED IN THE CMS EVALUATION DESIGN GUIDANCE FOR SECTION 1115 ELIGIBILITY AND COVERAGE POLICIES

Example survey items that collect demographic data

Appendix Table C.1 provides example survey items that collect demographic data on survey respondents. These items can generate control variables and inform subgroup analysis of the survey data if evaluators are not using administrative data as the source of demographic information. Evaluators must weigh instrument length against comparability in deciding how many demographic questions to include in beneficiary surveys. Evaluators electing to use the example items in the table below will need to modify item wording to accommodate the mode (web, paper, telephone) and type of administration (self-administered or proxy) of their survey. Response options are shown in parentheses, and the item number or name and source survey are shown in brackets. Hyperlinks route readers to the relevant questionnaire.

Table C.1. Example survey items that collect demographic data

Measure	Example item(s), [source]	Notes
Age	• What is your age? (Specify age in years) [AGE, BRFSS 2018] • What is your age? (18 to 24; 25 to 34; 35 to 44; 45 to 54; 55 to 64; 65 to 74; 75 or older) [33, CAHPS Adult Medicaid]	A continuous response format may be preferable for analysis, and the data can always be converted to categories. However, the continuous response format may have higher non-response than a categorical response format, especially if the survey is self-administered.
Sex / gender	• Format 1: What is your sex? Format 2: What was your sex at birth? Is it… (Male; Female; Don't know/Not sure) [SEX1, BRFSS 2018] • Which of the following best represents how you think of yourself? (Lesbian or Gay; Straight, that is, not gay; Bisexual; Something else; I don't know the answer; Refused) [SOFEMALE, BRFSS 2018] • Do you consider yourself to be transgender? (Yes – Transgender, male-to-female; Yes – Transgender, female-to-male; Yes – Transgender, gender nonconforming; No) [TRNSGNDR, BRFSS 2018]	BRFSS offers two formats of questions on sex. The BRFSS items on sexual orientation and transgender identification can help identify subgroups who may experience different outcomes.
Primary spoken language	• What language do you mainly speak at home? (English; Spanish; Chinese; Russian; Vietnamese; Portuguese; Some other language – Specify) [32, CAHPS Hospital Survey]	States may want to modify response options to include other high-prevalence languages in their region
Work-disability status	• Because of a physical, mental, or emotional condition lasting 6 months or more, does <this person> have any difficulty working at a job or business? (Yes; No) [DISABWRK, IPUMS ACS 2007]	
Race	• What is <Person X's> race? *Mark (X) one or more boxes*. (White; Black or African American; American Indian or Alaska Native – Print the name of enrolled or principal tribe; Asian Indian; Japanese; Chinese; Korean; Filipino; Vietnamese; Other Asian – Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on; Native Hawaiian; Guamanian or Chamorro; Samoan; Other Pacific Islander – Print race, for example, Fijian, Tongan, and so on; Some other race – Print race) [RACE, IPUMS ACS 2017]	
Ethnicity	• Is <Person X> of Hispanic, Latino, or Spanish origin? (No, not of Hispanic, Latino or Spanish Origin; Yes, Mexican, Mexican American, or Chicano; Yes, Puerto Rican; Yes, Cuban; Yes, another Hispanic, Latino, or Spanish Origin – Specify) [HISPN, IPUMS ACS 2017]	
Number of children in household	• How many children less than 18 years of age live in your household? (Specify number) [CHILDREN, BRFSS 2018]	
Marital status	• What is <this person's> marital status? (Now married; Widowed; Divorced; Separated; Never married) [MARST, IPUMS ACS 2017]	

Table C.1 (*continued*)

Measure	Example item(s), [source]	Notes
Living arrangement	• Type of living quarter for the residence. (House, apartment, flat; Unit in rooming house, hotel, motel, etc.) [TPRVLVQRT, SIPP 2014] • Are <...> living quarters owned, rented, or occupied without payment of rent? (Owned or being bought by someone in the household; Rented; Occupied without payment of rent) [ETENURE, SIPP 2014] • How many adults live at your home, including you? (1 – Just the respondent; 2 to 3; 4 or more; Don't know; Refused; Unclear response) [94, CAHPS Home and Community Based Services Survey] • Do you live with any family members? (Yes; No; Don't know; Refused; Unclear response) [95, CAHPS Home and Community Based Services Survey] • Do you live with people who are not family or are not related to you? (Yes; No; Don't know; Refused; Unclear response) [96, CAHPS Home and Community Based Services Survey]	States may be interested in collecting data on living arrangements if assessing transience is of interest, or if household composition is needed to estimate federal poverty level through survey data (evaluators should first assess whether FPL exists as a data element in Medicaid administrative data)

BRFSS = Behavioral Risk Factor Surveillance System; CAHPS = Consumer Assessment of Healthcare Providers and Systems; IPUMS ACS = American Community Survey; SIPP = Survey of Income and Program Participation.

Example survey items for evaluations of community engagement requirements

Appendix Table C.2 provides example survey items that correspond to outcomes for which a beneficiary survey is a recommended data source in the CMS evaluation design guidance for community engagement requirements (available at https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html). Each row in the table provides the number of the relevant research question and measure in the guidance, along with example items drawn from questionnaires that have been fielded in existing surveys. Where no federal survey item exists, the table provides example items from state-based surveys used in evaluations of 1115 demonstrations with eligibility and coverage policies. Where no established federal or state survey item exists, the table provides suggested wording.

States electing to use the example items in the table will need to modify item wording to accommodate the mode (web, paper, telephone) and type of administration (self-administered or proxy) of their survey. In these examples, parentheses contain response options, and the item number or name and source survey are shown in brackets. Hyperlinks route readers to the relevant questionnaire.

Table C.2. Example survey items for evaluations of community engagement requirements

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQs 1.1, 1.1a, and 1.1b	Probability of being employed	Are you currently…? (Employed for wages; Self-employed; Out of work for 1 year or more; Out of work for less than 1 year; a Homemaker; a Student; Retired; Unable to work) [EMPLOY1, BRFSS 2018] Alternatively: A. Last week, did <this person> work for pay at a job (or business)? (Yes; No – Did not work or retired) B. Last week, did <this person> do ANY work for pay, even for as little as one hour? (Yes; No) [EMPSTAT, IPUMS ACS 2017]	BRFSS structures this item as a single response option. Respondents who feel they fit into more than one category are directed to select the category that best describes them. Using the BRFSS item wording and structure allows states to compare their survey data with trends in their state overall. However, responses would not capture the number of people who report being a student and also being employed. If this detail is important, states should modify this question to a "select all that apply" format or consider asking a separate item to ascertain whether a respondent is currently in school.
	Number of hours worked per week	• How many hours did <this person> work LAST WEEK at all jobs? (Specify hours by subtracting any time off and adding overtime or extra hours worked) [HRSWORK1, IPUMS ACS 1990] Alternatively, repeat the following items for all named jobs: • Average number of hours worked per week at job 1 during the reference month. (Specify number of hours) [TJB1_MWKHRS, SIPP 2014] • Average number of hours worked per week at job 2 during the reference month. (Specify number of hours) [TJB2_MWKHRS, SIPP 2014]	These items should be asked only of respondents who report being employed. States should carefully weigh the time period of reference for items asking about job(s). Asking about jobs worked in the prior week poses less recall burden, though it may not capture employment experiences for respondents with sporadic or seasonal employment. States seeking a comprehensive view of current employment should collect data for multiple jobs separately. This can be done by asking whether respondent worked more than one job during time period of reference (e.g., last week) and then asking for detail on each. If not looping through all current jobs, simplify the task by asking respondents to report on the job where they work most hours. This approach also streamlines complexity for paper questionnaires. Asking respondents to sum hours across all jobs, regardless of time period of reference, can be burdensome and lead to inaccurate reporting if they work multiple jobs and/ or schedules that vary from week to week.

Table C.2 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 1.1c	Hourly wages	For <name's/your> <MAIN> job now, what is the easiest way for you to report <his/her/your> total earnings BEFORE taxes or other deductions: hourly, weekly, annual, or on some other basis? (Hourly; Weekly; Bi-weekly; Twice monthly; Monthly; Annually; Other – Specify; Don't know; Refused) [SCE2, CPS Displaced Worker, Employee Tenure, and Occupational Mobility Supplement 2018] If hourly: <EXCLUDING overtime pay, tips and commissions> What is <name's/your> hourly rate of pay on <this/(your/his/her) MAIN> job? (Enter dollar amount; Don't know; Refused) [SCE4, CPS Displaced Worker, Employee Tenure, and Occupational Mobility Supplement 2018] If weekly, other, or DK: <Including overtime pay, tips and commissions> What are <name's/your> usual weekly earnings on <this job/(your/his/her) MAIN job,> before taxes or other deductions? (Enter $ amount; Don't know; Refused) [SCE12, CPS Displaced Worker, Employee Tenure, and Occupational Mobility Supplement 2018] If bi-weekly: <Including overtime pay, tips and commissions,> What are <name's/your> usual bi-weekly earnings on <this job/(your/his/her) MAIN job>, before taxes or other deductions? (Enter $ amount; Don't know; Refused) [SCE12, CPS Displaced Worker, Employee Tenure, and Occupational Mobility Supplement 2018] If monthly or twice monthly: <INCLUDING overtime pay, tips and commissions> What are <name's/your> usual monthly earnings on <this job/(your/his/her) MAIN job>, before taxes or other deductions? (Enter $ amount; Don't know; Refused) [SCE13, CPS Displaced Worker, Employee Tenure, and Occupational Mobility Supplement 2018] If annually: (<Including overtime pay, tips and commissions> What are <name's/your> usual annual earnings on <this/(your/his/her) MAIN> job, before taxes or other deductions? (Enter $ amount; Don't know; Refused [SCE14, CPS Displaced Worker, Employee Tenure, and Occupational Mobility Supplement 2018]	People may be paid by the hour, but they may also be paid by the day or by the number of things they make, do, or sell. To improve data accuracy, it is ideal to collect wages in whatever format the respondent indicates poses least burden for them. States can then process the information as part of the data file preparation and/or analysis.
	Number of hours worked per week	See example item for RQ 1.1 in Table C.2: • [HRSWORK1, IPUMS ACS 1990]	Items suggested for RQ 1.1c should be asked only of respondents who report being employed
	Industry	• What kind of business or industry do you work in? For example, hospital, elementary school, clothing manufacturing, restaurant. (Specify) [TYPEINDS, BRFSS 2018]	If not collecting for all jobs, focus on primary job (where respondent works most hours)

Table C.2 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 1.1c (continued)	Availability of employer-sponsored insurance	Does <employer name> offer a health insurance plan to any of its employees? (Yes; No) [OFFER, CPS Annual Social and Economic Supplement 2018]	If asking only about a primary job, focus respondent on that job alone, not on whether insurance is offered by any employer. Also note that the emphasis of this item is on whether the insurance is offered, not whether the respondent uses it. Respondents who do not use the insurance may initially say "no" because they did not use it. Accordingly, it may be useful to add a soft check for "no" responses to confirm that the employer does not offer insurance.
RQ 1.2	Highest grade attained, degrees/credentials attained	What is the highest level of school ... completed or the highest degree received by December of <reference year>? (Less than 1st grade; 1st, 2nd, 3rd, or 4th grade; 5th or 6th grade; 7th or 8th grade; 9th grade; 10th grade; 11th grade; 12th grade, no diploma; High school graduate - diploma or GED or equivalent; Some college credit, but less than 1 year - regular Jr. coll. coll. univ.; 1 or more years of college, no degree - regular Jr. coll. coll. univ.; Associate's degree – 2 year college; Bachelor's degree - for example: BA, AB, BS; Master's degree - for example: MA, MS, MBA, MSW; Professional School degree - for example: MD (doctor), DDS (dentist), JD (lawyer); Doctorate degree - for example: Ph.D., Ed.D.) [EEDUC, SIPP 2014]	
	Certifications attained	Has <...> earned an educational certificate at a college, university, community college, or trade school? (Yes; No) [ECERT, SIPP 2014] Has <...> earned a professional certification or license? (Yes; No) [EPROCERT, SIPP 2014]	

Table C.2 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 2.1	Income	INCOME IN THE PAST 12 MONTHS. Mark (X) the "Yes" box for each type of income this person received, and give your best estimate of the TOTAL AMOUNT during the PAST 12 MONTHS. NOTE: The "past 12 months" is the period from today's date one year ago up through today.) Mark (X) the "No" box to show types of income NOT received. If net income was a loss, mark the "Loss" box to the right of the dollar amount. For income received jointly, report the appropriate share for each person -- or, if that's not possible, report the whole amount for only one person and mark the "No" box for the other person. A. Wages, salary, commissions, bonuses, or tips from all jobs. *Report amount before deductions for taxes, bonds, dues, or other items*. (Yes – specify $ amount; No) [INCWAGE, IPUMS ACS 2017] B. Self-employment income from own nonfarm businesses or farm businesses, including proprietorships and partnerships. *Report NET income after business expenses*. (Yes – specify $ amount; No; Loss) [INCBUS00, IPUMS ACS 2017] C. Interest, dividends, net rental income, royalty income, or income from estates and trusts. *Report even small amounts credited to an account.* (Yes – specify $ amount; No; Loss) [INCINVST, IPUMS ACS 2017] D. Social Security or Railroad Retirement. (Yes – specify $ amount; No) [INCSS, IPUMS ACS 2017] E. Supplemental Security Income (SSI) (Yes – specify $ amount; No) [INCSUPP, IPUMS ACS 2017] F. Any public assistance or welfare payments from the state or local welfare office. (Yes – specify $ amount; No) [INCWELFR, IPUMS ACS 2017] G. Retirement, survivor, or disability pensions. *Do NOT include Social Security.* (Yes – specify $ amount; No) [INCRETIR, IPUMS ACS 2017,] H. Any such other sources of income received regularly such as Veterans' (VA) payments, unemployment compensation, child support, or alimony. *Do NOT include lump sum payments such as money from an inheritance or the sale of a home.* (Yes – specify $ amount; No) [INCOTHER, IPUMS ACS 2017]	The first set of example items (#47A-H in the ACS survey instrument) ask respondents about a comprehensive list of income sources, which can help to create a more precise measure of total income. States should retain this wording if comparing demonstration results to the results for the overall state or to other states. However, states should also weigh the utility of a precise measure against respondent burden of reporting each of these 8 items separately. Alternatively, states can ask a single question about income and list the sources to include in the response, as shown in the second option (see next page); however, respondents may not be easily able to compute a total across a number of sources. As a result, the income reported may not be as accurate as when using the first approach. Finally, states can change the wording of these ACS items to ask about monthly income to supply data needed for RQ 2.1b. In general, any items that reference the past 12 months cast a wide net for inclusion of income, but this also increases recall burden for respondents. This could negatively impact accuracy.

Table C.2 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 2.1 (continued)	Income (continued)	Alternatively: • What was this person's total income during the PAST 12 MONTHS? Please include income from all sources, such as wages, salary, commissions, bonuses, or tips from all jobs; income from self-employment; Interest, dividends, net rental income, royalty income, or income from estates and trusts; Social Security or Railroad Retirement; Supplemental Security Income (SSI); public assistance or welfare payments; Retirement, survivor, or disability pensions; and any such other sources of income received regularly such as Veterans' (VA) payments, unemployment compensation, child support, or alimony [INCTOT, IPUMS ACS 2017 **REVISED**]	
RQ 2.1a and 2.1c	Child care costs	Did <reference parent> or <reference parent's> family pay for child care arrangements? (Yes; No) [EPAY, SIPP 2014] IF YES: How much <reference parent> or <reference parent's> family paid for child care in a typical week in the <reference month>? (Specify) [TPAYWK, SIPP 2014]	
	Transportation costs	• How much are <…> commuting expenses? (Specify $ Amount) [TJB6_PVOTHRC, SIPP 2014, **REVISED**]	States should consider revising the listed SIPP item, which asks about "work commuting," to strike the word "work," as respondents could incur costs relating to transportation to community engagement activities other than work. In the SIPP, the question about commuting expenses is only asked of those who, in a prior item, reported riding in a car/van pool, bus, rail, other public transportation, walked or biked to work, reported some other form of transportation to work. If collapsing all commuting-related expenses into a single item, states should consider adding a prompt to encourage respondents to include costs for parking, tolls, public transportation, and mileage for private automobile.

Table C.2 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 2.1a and 2.1c (continued)	Changes to income from loss of public program eligibility	• At any time during <last year>, did you receive <XX benefit>? (Yes, No, Refused, Don't know) [FSNAP, NHIS Family 2017 **REVISED**] • At any time during <last year>, did you have a change in <XX benefit>? **[NEW]** • If yes: How much was <XX benefit> before the change in <last year>? **[NEW]** • How much was <XX benefit> after the change in <last year>? [If using CATI, add confirmation text as follows: "To confirm, the change in your benefit was $[fill] in <last year>."] **[NEW]**	States and their evaluators can estimate changes to income for respondents who report receiving public benefits at some point in the past year. Surveys should ask about disability benefits, TANF, SNAP, housing support, and any other relevant public benefits. Surveys should avoid using the word "income," as respondents may not conceptualize the income from these sources in the same way as income from jobs (as shown in example item).
RQ 2.1b	Reported change in income by month; Probability of earning above 100 percent FPL, by month; Average monthly income since enrollment (or implementation of requirements)	• What was your total income from all sources in a typical month last year? Please include income from all sources, such as wages, salary, commissions, bonuses, or tips from all jobs; income from self-employment; Interest, dividends, net rental income, royalty income, or income from estates and trusts; Social Security or Railroad Retirement; Supplemental Security Income (SSI); public assistance or welfare payments; Retirement, survivor, or disability pensions; and any such other sources of income received regularly such as Veterans' (VA) payments, unemployment compensation, child support, or alimony. (Specify $ amount) [**NEW**] • Were there any months when you made *less* than the typical month? **[NEW]** • If yes: About how many months? **[NEW]** • If yes: How much was your income in those months? (Specify $ amount) **[NEW]** • Were there any months when you made *more* than the typical month? **[NEW]** • If yes: About how many months? **[NEW]** • If yes: How much was your income in those months? (Specify $ amount) **[NEW]** Alternatively, states could ask for average income for each calendar month in the prior year. For example, January (Specify $ amount), February (Specify $ amount) and so on.	Family size and income are needed to calculate FPL. States can use survey questions on number of children and marital status (see Table C.1) to estimate family size, in combination with survey questions on income. In general, measuring income by month is burdensome for respondents and may be subject to significant recall error.

Table C.2 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 3.1	Reported enrollment in commercial coverage, including ESI and Marketplace plans	Are you covered by any kind of health insurance or some other kind of health care plan? (Yes; No; Refused; Don't know) [FHICOV, NHIS 2017 – Family - **REVISED**] *If yes…* What kind of health insurance or health care coverage <do you/does ALIAS> have? INCLUDE those that pay for only one type of service (nursing home care, accidents, or dental care). EXCLUDE private plans that only provide extra cash while hospitalized. Enter all that apply. (Private health insurance; Medicare; Medi-gap; Medicaid; SCHIP – CHIP/Children's Health Insurance Program; Military health care – Tricare/VA/CHAMP-VA; Indian Health Service; State-sponsored health plan; Other government plan; Single service plan – dental, vising, prescription; No coverage of any type; Refused; Don't know) [HIKIND, NHIS 2017 – Family] *If private insurance…* Which one of these categories best describes how this plan was obtained? (Through employer; Through union; Through workplace, but don't know if employer or union; Through workplace, self-employed or professional association; Purchased directly; Through Healthcare.gov or the Affordable Care Act, also known as Obamacare; Through a state/local government or community program; Other – specify; Refused; Don't know) [PLNWRK, NHIS 2017 - Family]	This item should be asked of all respondents, regardless of current employment status, because people may be covered by a Marketplace plan or other commercial plan. Beginning with a "gatekeeper" question that asks about any health insurance coverage will reduce respondent burden, as people with no coverage will not need to say "no" for each type of insurance in later items.
RQ 3.1a	Reported offer of ESI (including whether the firm offers ESI and whether the individual is eligible for ESI)	See example item for RQ 1.1c in Table C.2. [OFFER, CPS Annual Social and Economic Supplement 2018] Could <name/you> be in this plan if <you/he/she> wanted to? (Yes; No) [COULD, CPS Annual Social and Economic Supplement 2018] Why <aren't/isn't> <you/he/she> in this plan? (Covered by another plan: Traded health insurance for higher pay; Too expensive; Don't need health insurance; Have a pre-existing condition; Haven't worked for this employer long enough to be covered; Contract or temporary employees not allowed in the plan; Other - Specify) [WNTAKE, CPS Annual Social and Economic Supplement 2018]	These items should be asked only of respondents who report they are employed and also that they are not participating in ESI.
RQs 3.1b and 3.1c	Reported enrollment in ESI	See example item for RQ 3.1 in Table C.2. [PLNWRK, NHIS 2017 - Family]	

Table C.2 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 3.1d	Reported out-of-pocket medical spending in the last year	The next question is about money that <you have/your family has> spent out of pocket on medical care. We do NOT want you to count health insurance premiums, over-the-counter drugs, or costs that you will be reimbursed for. In the PAST 12 MONTHS, about how much did <you/your family> spend for medical care and dental care? (Zero; Less than $500; $500-$1,999; $2,000-$2,999; $3,000-$4,999; $5,000 or more; Refused; Don't Know) [HCSPFYR, NHIS 2017 – Family]	The example item references the past 12 months, which collects costs over the desired period but also requires that the respondent compute any costs that varied over time. This time period also increases recall burden for respondents, which could have a negative impact on accuracy. States could revise this question to ask only about the past 6 months, or the most recent month.
	Reported problems paying medical bills	In the past 12 months did <you/anyone in the family> have problems paying or were unable to pay any medical bills? Include bills for doctors, dentists, hospitals, therapists, medication, equipment, nursing home, or home care. (Yes; No; Refused; Don't know) [MEDBILL, NHIS 2017 - Family] • <Do you/Does anyone in your family> currently have any medical bills that you are unable to pay at all? (Yes; No; Refused; Don't know) [MEDBNOP, NHIS 2017 - Family]	
RQ 3.1e	Reported enrollment in Marketplace plans	See example item for RQ 3.1 in Table C.2: [PLNWRK, NHIS 2017 - Family] Alternatively: Was the plan obtained through Healthcare.gov or the <Health insurance Marketplace/Health Insurance Marketplace, such as STATE NAME>? (Yes; No; Refused; Don't know) [PLNEXCHG, NHIS 2017 - Family]	
RQ 3.2	Health insurance coverage	See example items for RQ 3.1 in Table C.2: [FHICOV, NHIS 2017 – Family - **REVISED**]	
RQ 3.2a	Reported barriers to enrollment in new coverage	• Which one or more of these reasons describe why <name/s> <was/were> not covered? (Too expensive, can't afford health insurance; No health insurance offered by employer of self, spouse, or parent; Not working at a job long enough to qualify; Job layoff, job loss, or any reason related to unemployment; Not eligible because working part time or temporary job; Can't obtain insurance because of poor health, illness, age, or a pre-existing condition; Dissatisfied with previous insurance or don't believe in insurance; Have been healthy, not much sickness in the family, haven't needed health insurance; Able to go to VA or military hospital for medical care; Covered by some other health plan, such as Medicaid; No longer covered by parents' policy; Other – Specify) [HINONE, SIPP 2008]	States may want to limit this question to those who report not having coverage

Table C.2 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 4.1, 4.1a, and 4.1b	Reported physical health status	• Would you say that in general your health is: (Excellent: Very Good; Good; Fair; Poor) [GENHLTH, BRFSS 2018] • Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (Number of days between 1 and 30; None; Don't know/not sure; Refused) [PHYSHLTH, BRFSS 2018]	
	Reported mental health status	• Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? (Number of days between 1 and 30; None; Don't know/not sure; Refused) [MENTHLTH, BRFSS 2018]	If this item follows the question on general health, interviewers can use voice inflection to emphasize "mental health" to help respondents differentiate this item from the prior item. In self-administered formats, states can underline the term "mental health" to achieve the same objective.
	Reported emergency room admission in past year	• In the last 12 months, how many times did you go to an emergency room to get care for yourself? (None; 1 time; 2 times; 3 times; 4 times; 5 to 9 times; 10 or more times) [P-UT1, CAHPS Health Plan Adult Survey - Utilization]	
	Reported hospital admission in past year	Have you been hospitalized overnight in the past 12 months? Do not include an overnight stay in the emergency room. (Yes; No; Refused; Don't know) [FHOSPYR, NHIS DRAFT 2018 - Family **REVISED**].	

Table C.2 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 6	Barriers to compliance	There are no established items for this measure (see Notes). Potential item wording is as follows: • At any time in the last 12 months, did you have any problems completing your community engagement hours? (Yes; No; Not sure/Don't know) [**NEW**] • (If yes) What problems did you have? (I didn't have transportation; I couldn't find child care; I couldn't afford child care; I was sick; Someone in my family got sick; I couldn't get enough hours at my job; I lost my job; I didn't take enough credit hours at school; I couldn't get enough volunteer hours; I met the requirement but couldn't report my hours) [**NEW**]	Barriers to compliance will vary depending on the policy design, state implementation, state-specific economic factors, and the state-specific availability of supports. States and their evaluators should therefore develop items and response options using information collected through individual or group interviews with current and/or former beneficiaries. States should consider asking respondents about their knowledge of the requirement (RQ 8) before asking about barriers to compliance (RQ 6). States conducting multi-modal surveys should change the format of "check all that apply" responses to ask each item as a "yes/no" format (such as a grid). It is ideal to place items about barriers to compliance after items capturing program knowledge. Program knowledge questions help prime the respondent to think about the program, which then positions them to think carefully about barriers they face in complying with it.
RQ 7	Number and proportion of beneficiaries reporting receipt of supports provided or arranged by Medicaid agency or included in referrals to non-Medicaid agencies or resources	There are no established items for this measure (see Notes). Potential item wording is as follows: • At any time in the last 12 months, did you receive a referral to any support services to help you complete your community engagement hours? Transportation; Child care; Language or translation services; Job placement; Career or education counseling (Yes; No; Don't know/Not sure) [**NEW**] • At any time in the last 12 months, did you use any support services to help you complete your community engagement hours? Transportation; Child care; Language or translation services; Job placement; Career or education counseling (Yes; No; Don't know/Not sure) [**NEW**]	Supports may be provided or arranged by Medicaid agencies or partnering organizations. States and their evaluators should develop items and response options based on state-specific implementation of supports. Surveys should include at least one item on receipt of supports even if the Medicaid agency refers beneficiaries to other organizations.

Table C.2 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes

RQ 8	Scaled measures of enrollee knowledge of requirements and consequences of noncompliance	Surveys used in evaluations of 1115 demonstrations with eligibility and coverage policies have used various approaches to item wording to gather data on beneficiary knowledge. One approach is to ask beneficiaries what they think the requirement is and offer several response options, only one of which is correct for the individual respondent. An example of this approach is:
		• What do you think would happen, if anything, if a person's contribution(s) is not made on time? (Nothing will change; Their HIP 2.0 coverage will end; They would automatically get moved to HIP Basic; Not sure/Don't know) [35, HIP 2.0 Enrollee Web Survey]
		Another approach is to ask a series of questions with yes, no, and don't know response options. An example is:
		• For these next statements about the Healthy Michigan Plan coverage and costs: if you think the statement is correct, say "yes." If you think it is incorrect, say "no." If you don't know, say "don't know." [47-55, 2016 Healthy Michigan Voices Survey; all response options are Yes; No; Don't know]
		  ○ My Healthy Michigan Plan covers routine dental visits
		  ○ My Healthy Michigan Plan covers eyeglasses
		  ○ My Healthy Michigan Plan covers counseling for mental or emotional problems
		  ○ Only generic medicines are covered by my Healthy Michigan Plan
		  ○ Contributions are what I am charged every month for Healthy Michigan Plan coverage even if I do not use any health care;
		  ○ There is no limit or maximum on the amount I might have to pay in copays or contributions
		  ○ I could be dropped from the Healthy Michigan Plan for not paying my bill
		  ○ I may get a reduction in the amount I might have to pay if I complete a health risk assessment
		  ○ Some kinds of visits, tests, and medicines have no copays
		Regardless of approach to item wording, states should include similar items in surveys of both current and former beneficiaries to facilitate comparisons. An example item for a former beneficiary that is similar to the HIP 2.0 survey item above is:
		• While you were in HIP 2.0, what did you think would happen, if anything, if a person's contribution was not made on time? (Nothing would change; Their HIP 2.0 coverage would end; They would automatically get moved to HIP Basic; Not sure/Don't know) [24, HIP 2.0 Disenrollee Survey]
			Beneficiary knowledge will vary depending on the policy design and state implementation. States and their evaluators should develop items and response options based on the policy design and information collected through individual or group interviews with current and/or former beneficiaries. Items that assess beneficiary understanding should avoid wording that makes respondents feel they are being quizzed or tested.
			States may find it helpful to include some items that are intentionally phrased in ways that do not reflect the policy or would not be considered "true." The Healthy Michigan Plan example at left has this feature. (There are, in fact, limits on contributions, and beneficiaries in Michigan cannot be dropped for failing to pay a bill.)
			Interviewer training should emphasize that recording a "don't know" response to knowledge questions is important, and that "don't know" is not equivalent to missing data.

AHS = American Housing Survey; BRFSS = Behavioral Risk Factor Surveillance System; CPS = Current Population Survey; ER = emergency room; ESI = employer-sponsored insurance; FPL = federal poverty level; IPUMS ACS = Integrated Public Use Microdata Series, American Community Survey version; NHIS = National Health Interview Survey; SIPP = Survey of Income and Program Participation.

Example survey items for evaluations of premiums and beneficiary account payments

Appendix Table C.3 provides example survey items that correspond to outcomes for which a beneficiary survey is a recommended data source in the CMS evaluation design guidance for premiums and beneficiary account payments (available at https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html). Each row in the table provides the number of the relevant research question and measure in the guidance, along with example items drawn from state-based surveys used in evaluations of 1115 demonstrations with eligibility and coverage policies. States electing to use the example items in the table will need to modify item wording to accommodate the mode (web, paper, telephone) and type of administration (self-administered or proxy) of their survey. In these examples, parentheses contain response options, and the item number or name and source survey are shown in brackets. Hyperlinks route readers to the relevant questionnaire.

Table C.3. Example survey items for evaluations of premiums and beneficiary account payments

Research question	Measure needed for research question	Example item(s), [source]
RQ 1.1	Reported knowledge of monthly payment requirements and consequences of nonpayment	See notes and example items for RQ 8 in Table C.2. Additional example items relevant to understanding of monthly payment requirements and consequences of nonpayment are as follows: • How much is your monthly HELP premium? ($0 to $9; $10 to $19; $20-$29; $30-$39; $40-$49; $50 and above; Not sure/Don't know) [15, Montana HELP Plan Enrollee Survey] • If you do not get a physical exam this year, will your coverage plan require you to pay a monthly premium next year? (Yes; No) [81, 2014 Survey of Iowa Wellness Plan Enrollees, REVISED]
RQ 2.2b	Reported knowledge of beneficiary account rules and incentives	See notes and example items for RQ 8 in Table C.2. Additional example items relevant to knowledge of beneficiary account rules and incentives are as follows: • Do you know how much is in your POWER account today? (Yes, I know exactly how much; Yes, I have a pretty good idea; No, I do not really know at all) [27, HIP 2.0 Enrollee Web Survey] • For each of the following statements about POWER accounts, please tell us whether you agree, disagree, or are not sure. The State of Indiana contributes to POWER accounts HIP 2.0 contribution/s go to POWER accounts POWER accounts help people pay for the health care services they need POWER accounts help people understand the cost of their health care services POWER accounts make people feel comfortable about paying for their health care services (Agree; Disagree; Not sure) [38, HIP 2.0 Enrollee Web Survey] • How easy or hard is it to understand what happened to any leftover money in a POWER account at the end of the year? (Very easy; Somewhat easy; Neither easy nor hard; Somewhat hard; Very hard) [39, HIP 2.0 Enrollee Web Survey] • If someone gets all or some of their recommended preventive services, would some of the remaining money in a POWER account get rolled over to next year? (Yes; No; Not sure/Don't know) [41, HIP 2.0 Enrollee Web Survey] • For the following statements, tell me if you strongly agree, agree, are neutral, disagree, or strongly disagree: I carefully review each MI Health Account statement to see how much I owe; The MI Health Account statements help me be more aware of the cost of health care; Information I saw in a MI Health Account statement led me to change some of my decisions about health care [45, 2016 Healthy Michigan Voices Survey]

HELP = Montana's Health and Economic Livelihood Partnership (Montana's section 1115 demonstration); HIP = Healthy Indiana Plan (Indiana's section 1115 demonstration); POWER = Personal Wellness and Responsibility Account (beneficiary health account in Indiana's section 1115 demonstration)

Example survey items for evaluations of non-eligibility periods

Appendix Table C.4 provides example survey items that correspond to outcomes for which a beneficiary survey is a recommended data source in the CMS evaluation design guidance for non-eligibility periods (available at https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html). Each row in the table provides the number of the relevant research question and measure in the guidance, along with example items drawn from questionnaires that have been fielded in existing surveys. Where no federal survey item exists, the table provides example items from state-based surveys used in evaluations of 1115 demonstrations with eligibility and coverage policies. Where no established federal or state survey item exists, the table provides suggested wording.

States electing to use the example items in the table will need to modify item wording to accommodate the mode (web, paper, telephone) and type of administration (self-administered or proxy) of their survey. In these examples, parentheses contain response options, and the item number or name and source survey are shown in brackets. Hyperlinks route readers to the relevant questionnaire.

Table C.4. Example survey items for evaluations of non-eligibility periods

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 1.1a	Reported knowledge of program requirements and how to comply with them	See notes and example items for RQ 8 in Table C.2.	
RQ 1.1b	Reported knowledge of non-eligibility period consequence for noncompliance with program requirements	See notes and example items for RQ 8 in Table C.2, particularly item 35 from the HIP 2.0 Enrollee Web Survey. An additional example item relevant to understanding of non-eligibility periods is as follows: • How long do you think a person would need to wait to re-enroll in HIP 2.0? (No wait time; 3 months; 6 months; 12 months; Not sure/Don't know) [36, HIP 2.0 Enrollee Web Survey]	Interviewers should not be in a position of answering clarifying questions about programs or compliance. States should supply interviewers with resources or information they can provide to respondents who ask questions about the program. Examples of such resources could include a program website, a toll-free telephone number for program-related questions, or a program brochure.
RQ 1.2	Reported barriers to complying with program requirements	See notes and example items for RQ 6 in Table C.2	
RQ 2.1b	Reported knowledge of pathways for re-enrollment in Medicaid after non-eligibility period	Pathways to re-enrollment in Medicaid will vary by state, although most states require a new application. Potential item wording is as follows: • Do you think you will need to submit a new Medicaid application to re-enroll? (Yes; No; Don't know/Not sure) **[NEW]** States should modify this question or add others to ask about other pathways to re-enrollment. An additional example from an existing survey (see response option B) is: • *For each of the following statements, please tell us whether you thought it was part of your HELP plan:* A. Payment of any unpaid premiums within 90 days would have allowed me to keep my HELP coverage. B. Payment of any unpaid premiums after 90 days would have allowed me to re-enroll in HELP within 12 months of my HELP start date. C. Any unpaid premium balance may be collected from my future state income tax refunds (Part of your HELP plan; Not part of your HELP plan; Not sure) [21, Montana HELP Plan Disenrollee Survey]	

Table C.4 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 2.1c	Reported knowledge of options for accessing low-cost health care during non-eligibility periods	There are no established survey items for this measure. Options for accessing low-cost health care may vary by state. States and their evaluators can develop items and response options based on state context and information collected through individual or group interviews with current and/or former beneficiaries. Potential item wording for a survey of former beneficiaries is as follows: • After you were no longer enrolled in [*program name*], did you have any place you could go to get affordable health care services when needed? (Yes; No; Don't know/Not sure) [**NEW**]	
RQ 3.1	Change in physical and mental health status	See example item for RQ 4.1 in Table C.2: [GENHLTH, BRFSS 2018]; [MENTHLTH, BRFSS 2018]; [PHYSHLTH, BRFSS 2018]	
RQ 3.1b	Whether beneficiaries experienced unmet medical need due to cost during non-eligibility period	• DURING THE PAST 12 MONTHS, was there any time when you needed any of the following, but didn't get it because you couldn't afford it? ...Prescription medicines. ...Mental health care or counseling. ...Dental care (including check-ups). ...Eyeglasses ...To see a specialist. ...Follow-up care. (Yes; No; Refused; Don't know) [AHCAFY_1-6, NHIS 2017 - Adult] • DURING THE PAST 12 MONTHS, were any of the following true for you? …You skipped medication doses to save money. …You took less medicine to save money. …You delayed filling a prescription to save money. …You asked your doctor for a lower cost medication to save money. …You bought prescription drugs from another country to save money. …You used alternative therapies to save money. (Yes; No; Refused; Don't know) [ARX12_1-6, NHIS 2017 - Adult]	NHIS and BRFSS items ask about any unmet needs in the past 12 months. States could compare answers for those who did and did not experience a non-eligibility period.

Table C.4 (*continued*)

Research question	Measure needed for research question	Example item(s), [source]	Notes
RQ 3.1b (continued)	Whether beneficiaries experienced unmet medical need due to cost during non-eligibility period (continued)	• Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? (Yes; No; Don't know/Not sure) [MEDCOST, BRFSS 2018] • Not including over-the-counter medications, was there a time in the past 12 months when you did not take your medication as prescribed because of cost? (Yes; No; No medication was prescribed; Don't know/not sure; Refused) [Module 4, Question 3; BRFSS 2016] Alternatively, states could emulate the wording used for Montana, which is more targeted to those experiencing a non-eligibility period, and amends the BRFSS item wording to ask about any health care. • After you were no longer enrolled in HELP, was there any time you needed health care but did not get it because of cost? (Yes; No; Not sure/Don't know) [7, Montana HELP Plan Disenrollee Survey] • After you were no longer enrolled in HELP, what types of health care were you unable to get because of cost? A. A visit to the doctor when I was sick B. Preventive care (such as blood pressure check, flu shot, family planning services, prenatal services, cholesterol or cancer screenings) C. A follow-up visit to get tests or care recommended by my doctor D. Dental care E. Vision (eye) care F. Prescription drugs G. Emergency room care (Yes; No; N/A) [8, Montana HELP Plan Disenrollee Survey]	

BRFSS = Behavioral Risk Factor Surveillance System; HELP = Montana's Health and Economic Livelihood Partnership (Montana's section 1115 demonstration); HIP = Healthy Indiana Plan (Indiana's section 1115 demonstration); NHIS = National Health Interview Survey

Example survey items for evaluations of retroactive eligibility waivers

Appendix Table C.5 provides example survey items that correspond to outcomes for which a beneficiary survey is a recommended data source in the CMS evaluation design guidance for retroactive eligibility waivers (available at https://www.medicaid.gov/medicaid/section-1115-demo/evaluation-reports/evaluation-designs-and-reports/index.html). Each row in the table provides the number of the relevant research question and measure in the guidance, along with example items drawn from questionnaires that have been fielded in existing surveys. Where no established survey item exists, the table provides suggested wording. States electing to use the example items in the table will need to modify item wording to accommodate the mode (web, paper, telephone) and type of administration (self-administered or proxy) of their survey. In these examples, parentheses contain response options, and the item number or name and source survey are shown in brackets. Hyperlinks route readers to the relevant questionnaire.

Table C.5. Example survey items for evaluations of retroactive eligibility waivers

Research question	Measure needed for research question	Example item(s), [source]
RQ 1.2a	Reported knowledge of Medicaid policy on coverage during enrollment gaps	There are no established survey items for this measure. States and their evaluators can develop items and response options based on state policy. Potential item wording is as follows: • If you don't send in your renewal application on time and your current Medicaid insurance ends, which of the following will happen? Choose all that apply. (Nothing will change; I will have to send in a new application; Medicaid will pay my medical bills if they get my application within 3 months after my coverage ends; Medicaid will not cover my bills until I am approved for new coverage) [**NEW**]
RQ 1.2b	Reported barriers to timely renewal	There are no established items for this measure. States and their evaluators should develop items and response options based on state policy, and may also use information collected through individual or group interviews with current and/or former beneficiaries. Potential item wording applicable to beneficiaries who have renewed coverage (in their second or later enrollment span in the demonstration) is as follows: • Did you have any problems filing your renewal application on time? (Yes; No; Not sure/Don't know) [**NEW**] • (If yes) What problems did you have? (Possible response options are: I didn't know it was time to renew until after my coverage ended; I couldn't file the renewal application online; I couldn't file the renewal application in person; I couldn't find the renewal application; I had a hard time filling out the renewal application; I filed my renewal application, but it wasn't accepted) [**NEW**]
RQ 2.1	Reported excellent or very good health status (physical and/or mental overall)	See example items for RQ 4.1 in Table C.2: [GENHLTH, BRFSS 2018]; [PHYSHLTH, BRFSS 2018]; [MENTHLTH, BRFSS 2018]
	Reported prior-year utilization (any overnight hospital stay)	See example item for RQ 4.1 in Table C.2: [FHOSPYR, NHIS DRAFT 2018 - Family]
	Reported prior-year utilization (any ER visit)	See example item for RQ 4.1 in Table C.2: [P-UT1, CAHPS Health Plan Adult Survey - Utilization]
RQ 3.1	Change in physical and mental health status	See example items for RQ 4.1 in Table C.2: [GENHLTH, BRFSS 2018], [MENTHLTH, BRFSS 2018]; [PHYSHLTH, BRFSS 2018]

BRFSS = Behavioral Risk Factor Surveillance System; CAHPS = Consumer Assessment of Healthcare Providers and Systems; NHIS = National Health Interview Survey.

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
Progress Together

mathematica-mpr.com