Supporting Statement – Part B

Collections of Information Employing Statistical Methods

Overview

This field test will use a probability sample of each Program's eligible participants. Because the purpose of this field test is to validate and select assessment items, composites, and case-mix adjusters for the final version of the instrument, our priority is to include the range of programs and persons likely to be encountered in a national implementation rather than to describe a population of Programs. Therefore, a probability sample of Programs is unnecessary, but a probability sample of eligible persons within each participating Program is desirable. The sample will need to be large enough to allow us to make program-level estimates of the scores on any assessment items and composites. This goal requires that the sample of individuals within each program be representative of the population served by that program. In addition, we plan to embed an experiment in the overall sampling design in order to test the impact of survey mode on participant response.

We are planning to include up to 30 Programs in the field test. CMS will send an RFA to States to identify those that are interested in participating as pilot study sites. Based on responses to the RFA, CMS expects to select ten states with two or more programs within each state. The selection criteria from the CMS RFA include:

- Some states must have two or more programs, serving different populations.
- The participating programs must encompass a variety of disability populations, including, at a minimum, at least one of the following: older adults with age-related disabilities; non-elderly adults with physical disabilities; adults with intellectual/developmental disabilities; and adults with severe and persistent mental illness. Programs serving smaller, diagnosis-specific populations, such as traumatic brain injuries or persons with HIV/AIDS, are also eligible to participate.
- Each Program will provide a list of unique IDs and potential stratification variables for the entire population. The survey vendor will select a stratified random sample of the population. The vendor will send the IDs for the selected sample from each Program, which will provide the needed contact information and administrative data for the selected persons. Sample size recommendations can be found below in the Sampling Plan section.
- States will participate in a mode test, in which some interviews are conducted in person and some via phone.
- States will participate in a response option test, in which some interviews are conducted using the longer 4-point Likert scale response options (never, sometimes, usually, and always) and some using the shorter binary response options (mostly yes and mostly no).

- Data must be collected and submitted using a standardized electronic template to be provided by CMS and must include a unique identifier that can be used to link to other state data sources (i.e. claims and assessment data)
- For enrollees in participating programs, states must be able to provide accurate and current contact info as well as minimal information about service use, names of providers, etc.

1. Sampling methods

Goals

The sampling plan is designed to balance the requirements necessary for meeting three overall goals for this survey effort:

1. **Goal 1:** Obtain a sample of program participants sufficient in size to meet the requirements of the psychometric analysis. The generalizability of the results from the psychometric analysis is obtained by maximizing the variance of the sample with respect to the salient characteristics of the study population (age, disability, experiences, etc.), so as to capture the full range of potential response patterns in the population. The rule of thumb is to obtain a *minimum* of 10 *complete responses* for each item that will be used in the psychometric analysis. At this time, the survey includes 55 assessment items, which translates into a minimum of 550 completed surveys, assuming that each completed survey contains a non-missing response for each substantive item. However, given that *some* substantive items will be legitimately skipped by respondents to whom the subject matter of the item does not apply, this number will typically need to be larger. In addition, some completed surveys may still have some degree of item non-response (when a respondent skips an item that he/she should have answered). Thus, a minimum number of completes of at least 1,000 would be needed to assure that we meet this goal. Assuming an overall response rate of 50%, we would need a sample of at least 2,000 for this goal.
2. **Goal 2:** Obtain a sample that is both large enough and representative enough to: 1) make program-level estimates of assessment scores (items, composites, and global ratings), and 2) rank-order programs based on their performance relative to the average performance across participating programs. If a program is small, achieving this goal might require surveying the program's entire population.
3. **Goal 3:** Identify potential mode effects. We will randomly allocate the initial sample to two data collection modes – telephone (CATI) and face-to-face (CAPI). This task is complicated by the fact that the programs encompass several disability populations whose members may be unable to respond to a telephone interview. Because the CATI mode will likely experience lower response and completion rates than the CAPI mode, we will assign to CATI only the minimum number of sample members needed to make reliable estimates for the CATI mode. This approach will enable us to detect mode effects and minimize overall non-response.
4. **Goal 4:** Identify potential response option effects. Within each mode, (CATI and CAPI) participants will be randomly assigned into one of the two response option formats: half of the sample will be randomly assigned the survey with the 4-point Likert response options

and the other half will be assigned to the binary response options. This approach will enable us to detect variation based on response options and to develop a method of creating a single score.

The first goal sets the minimum threshold for total number of completions and the second goal sets the maximum. With respect to the second goal, field testing of previous CAHPS surveys has demonstrated that the optimal number of responses per unit assessed is 300 if the unit is an organization in which the patient is likely to be served by a wide variety of individual health physicians and other professionals; for instance, the required sample size for the CAHPS Health Plan and Hospital Surveys is 300. Because the intent of the HCBS Survey is to assess the full range of services offered by programs comprising a wide variety of providers, we have made an initial assumption that we will require a *minimum* of approximately 300 completions per program to produce reliable program-level scores. Assuming 30 programs, we would need approximately 9,000 completions total.

In addition, we have conducted a power analysis to verify that 300 completions per program would be sufficient to meet our analytic goals. The specifications of this power analysis are described in the sections below.

Response Rate Assumptions

Total required sample size is a function of the desired number of completes divided by the estimated overall response rate – that is, out of all the individuals sampled, what percent actually complete the survey. This rate takes into consideration the percentage of individuals who are unreachable, the percentage of individuals who are deemed ineligible by screening criteria, and the percentage of eligible individuals who actually complete the survey. To avoid confusion in the text below, we refer to this rate as the 'yield rate,' and reserve the term 'response rate' to refer to number of completed interviews with individuals divided by the number of eligible individuals in the sample.[1]

We expect an overall yield of 50%, based on three factors. First, we expect no more than 5 to 10% of the sampled population will be found ineligible for the survey. Historically the aged/disabled portion of the Medicaid population has been quite stable – the eligibility status of these low-income and chronically-impaired individuals does not vary over time. Furthermore, eligibility for Medicaid HCBS programs must be verified at least every 12 months, by statute. A small number of sampled individuals may either die or be institutionalized between sample selection and attempts to contact, but we do not expect this impact to be significant.

A more critical issue may be difficulty in contacting and/or locating sampled individuals. Sampled disabled Medicaid participants may move, or change or lose telephone numbers over time. In addition, our experience with state administrative datasets for these programs suggests they are not necessarily updated on a timely basis. While direct care and case management staff

[1] See p. 44 of AAPOR's "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, Revised 2011" for some different definitions of response rates. http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions2&Template=/CM/ContentDisplay.cfm&ContentID=3156

generally have accurate contact information in order to serve clients, this updated information may not be entered into the automated system. We estimate that, as a result, up to 20% of the eligible population may be "unreachable."

For those who are both eligible and can be contacted, our experience with other consumer surveys of this population, particularly the field test for the Participant Experience Survey, indicates we can expect a response rate of about 75%. The HCBS population can be characterized by social isolation and unemployment. We have found that most individuals welcome and appreciate an opportunity to discuss their experiences and lives with trained interviewers. When all 3 factors are considered, we assume around 90% eligible, with 80% reachable, and a 75% response rate, which gives us a yield of just over 50% (0.90*0.80*0.75 = 0.54). In order to obtain 300 completes per program with a yield rate of 50%, we would need to sample 600 participants from each program, for a total sample size of 18,000 (600 sampled * 30 programs).

Statistical power and sample size requirements

The power of a statistical test is the probability of detecting an effect in the sample data that actually exists in the population, or the inverse of power of the probability of making a Type II error (a false negative). Statistical power is a function of sample size, the type of statistical test that will be employed for estimating the effects (e.g., a two-sample t-test, an F-test in a one-way ANOVA, a correlation, or the beta estimated by a regression procedure), the specified alpha (i.e., probability of making a Type I error), and the effect size one wishes to be able to detect. The effect size (**d**) for any given outcome measure is the observed difference in means for a given comparison (e.g., comparing programs to a benchmark) divided by the variance associated with the outcome – as such, the effect size is a *standardized* indicator of the 'effect' of interest.

We prefer to use effect sizes in power analyses because with a large enough sample, even extremely small differences across groups of interest will be statistically significant, but might be substantively meaningless. Using effect sizes to guide the sampling design ensures that the sample is sufficient to detect differences large enough to be substantively meaningful without being needlessly large (and costly). Sample sizes increase exponentially, as do costs, with the ability to detect ever decreasing effect sizes, so the goal is to obtain a sample size sufficient for detecting differences in performance that are meaningful to the various stakeholders in the research. The other advantage to using effect sizes in a power analysis is that one does not necessarily need to have *a priori* estimates of the means and variances of the salient performance measures for the population of interest. Rather, we can examine a set of means and variances associated with a range of effect sizes. We follow the convention that defines small, medium and large effect sizes as 0.20, 0.50 and 0.80 respectively.[2]

For the proposed research, statistical power is defined as the probability of detecting true difference in program performance as compared to a benchmark (the average performance across all programs), or of detecting true differences in the propensity and ability to respond, as well as differences in response quality, across modes (for the mode experiment) and across response

[2] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd Edition. Lawrence Erlbaum Associates, NJ

options (for the response option experiment).

Determining the desired power involves considering whether the consequences of a false negative (a Type II error) are as serious, or less serious, as the consequences of a false positive (a Type I error), the inverse of which is the alpha level of the statistical test, which is typically set at 5%. If the consequences of a false negative are as great as those of false positive, then we would want our statistical test to have a power of 95%, which corresponds to a 5% probability of making a Type II error (incorrectly failing to reject the null). Typically, however, it is more serious to make a Type I error, and thus power can generally be set lower than 95%. A popular convention is to require that statistical tests have a power of 80%. [2] This is equivalent to saying that the consequences of finding a difference in program quality where there is none (a false positive) is four times as serious as the consequences of a false negative.

Sample Size Estimates for Initial Sample

Table 1 below summarizes the results of a power analysis using some hypothetical mean differences and a variance that would be associated with small, medium, and large effect sizes. The variance used for these calculations is typical of those observed for the various 'recommend' survey questions that appear in several CAHPS surveys.[3] The scenarios described in the table are designed to yield some potential sample size estimates suitable for satisfying the second goal described above – making program-level estimates that will allow us to rank-order programs by performance relative to a benchmark.

[3] For example, for Home Health CAHPS: Would you recommend this agency to your family or friends if they needed home health care? Response options: definitely yes, probably yes, probably no, definitely no.

Table 1. Sample Size Estimates for Three Different Scenarios

	Effect Size	ES	Mean Diff (4-pt scale)	Estimated Average Completes per Program	Total Completes	Yield Rate	Total Sample	Sample per Program
Scenario 1: Unequal variances, unbalanced design; assuming single program variance is lower than pooled variance	Small	0.18	0.10	66	1,645	0.5	3,290	132
	Moderate	0.51	0.28	10	252	0.5	504	20
	Large	0.82	0.45	5	132	0.5	264	11
Scenario 2: Unequal variances, unbalanced design; assuming single program variance is higher than pooled variance. This is the most realistic scenario.	Small	0.18	0.10	1,369	34,230	0.5	68,460	2,738
		0.36	0.20	344	8,610	0.5	17,220	689
	Moderate	0.51	**0.28**	**178**	**4,440**	0.5	**8,880**	**355**
		0.73	0.40	89	2,220	0.5	4,440	178
	Large	0.82	0.45	71	1,770	0.5	3,540	142
Scenario 3: Equal variances (each program variance = pooled variance) and unbalanced design	Small	0.18	0.10	298	7,440	0.5	14,880	595
	Moderate	0.51	0.28	41	1,020	0.5	2,040	82
	Large	0.82	0.45	18	450	0.5	900	36

ES = the calculated effect size, which is equal to the estimated mean difference of the outcome of interest divided by the standard deviation of that outcome. All ES's shown assume that the standard deviation is 0.55, which is the observed overall standard deviation for the 'recommend' item used in Home Health CAHPS (measured on a 4-point scale). In part, the specification of an 'unbalanced design' is based on the assumption that the number of events sampled from each program (the strata) will vary. Since standard CAHPS reporting of performance scores is based on an analysis that uses a two-sample t-test to compare the mean of one 'unit' (program) to the pooled mean of all the other units (programs), the specification of an unbalanced design is also based on the fact that the N for the pooled mean could be as much as 29 times the size of the N for the single program mean (we have assumed 30 programs, so one unit would generally have approximately 1/29th the sample size of the other 29 units combined). The single program variance in Scenario 1 is based on the lowest observed variance from CAHPS Home Health survey data; scenario 2 uses the highest observed variance from CAHPS Home Health survey data; comparing Scenarios 1 and 2 gives you the most liberal and most conservative estimates, based on our best guesses about the mean differences and variances we will find in the HCBS population.

As shown in the table, Scenario 2 reflects the most realistic set of assumptions, which can be stated as follows:

1. The statistical test of interest is to compare the mean score from each program to the overall mean of the other programs (the benchmark).
2. Such a test constitutes an unbalanced design in that the number of observations for a single program will be approximately 1/29th the number of observations that make up the benchmark score.
3. The variance in the outcome for a single program will be greater than the variance for that same outcome for the other 29 programs combined. Since variance is, mathematically, a function of the number of observations, the pooled variance for 29 programs will very likely be smaller than the variance for a single program.

Based on the goal of detecting a medium effect size using assumptions from Scenario 2 and an assumed 50% yield rate, the power analysis shows that we would need to sample at least 355 participants from each of the 30 programs to be able to detect effects of a moderate size when making program-level estimates in order to rank-order programs by performance relative to a benchmark (the overall mean in this case). This corresponds to a total sample size of 10,650, but does not take into consideration the variability due to the fact that the respondents may be served by a wide variety of individual providers. Since we cannot readily estimate this variability prospectively, we plan to use the more conservative CAHPS guideline of obtaining 300 completes per program, which requires a sample of 600 participants from each program.

Experimental design for mode test

For the mode test, we assumed *a priori* that we would have to randomly assign about 10% of participants from the initial sample of 18,000 to the CATI mode, while the remaining 90% of participants would be assigned to the CAPI mode. This strategy would provide samples of about 1,800 for the CATI mode and around 16,200 for the CAPI mode. Such an allocation will result in an unbalanced design where the CAPI mode will have approximately 9 times as many sample members compared to the CATI mode. We conducted a power analysis to verify that such a distribution would be adequate for the mode test.

The mode test needs to take into consideration the differences across populations in respondents' ability to complete a telephone survey. Table 2 displays the estimated distribution of the *total* population of interest across different *program* populations (e.g., the aged, physically disabled, etc.). Table 2 also characterizes each of these program populations by our estimation of their ability to complete a computer assisted telephone interview (CATI). Note that these assumptions are not based on specific empirical data regarding differences in propensity to respond across these populations, but rather represent assumptions based on our past experience working with and surveying similar populations. As shown in the table, we estimate that about 6% of the population will be able to complete a CATI survey, while around 50% *might* be able to complete such a survey (the 3 *probable* groups combined), and around 41% will *probably not* be able to complete a CATI survey. For the *Money Follows the Person* survey, which included the same population,

84% of the community-based sample was able to respond to the CAPI survey.[4]

Table 2. Population Size Estimates and Hypothesized Ability to Complete CATI

Population	Estimated % of Total HCBS Population	Likelihood of Being Able to Complete CATI?	Estimated # of Sample Members
Intellectual/developmental disability (ID/DD)	41%	Limited	7,434
Aged	11%	Probable	2,052
Aged/physically-disabled	39%	Probable	7,092
Physically disabled	6%	Good	1,134
HIV/AIDS	1%	Probable	252
Mental Health	< 1%	Probable	36

Note: we estimate that children make up about 2% of the total HCBS population, but they will not be included in the study, so they are excluded from our estimates. Percentages do not sum to 100 due to rounding. Estimated numbers in the last column were derived by multiplying 18,000 by the percentage in the second column, but do not match exactly due to rounding of the percentages.

Assuming the initial sample reflects the population distribution shown in Table 2, we would expect to have approximately 1,130 sample members in the group with a *good* likelihood of being able to complete CATI, almost 9,400 who *probably* will be able to complete CATI, and just over 7,400 whose likelihood of being able to complete CATI will be *limited*, or who will be less likely to be able to complete CATI (intellectual/developmental disability population).

Even though we have a fair amount of confidence that some members of the various populations may *not* be able to complete a CATI survey, we still need to test the impact of survey mode within each of these groups in order to test whether our predictions are supported by empirical evidence. Thus, the goal is to assign the minimum number of participants needed to provide sufficient power to each design cell shown in Table 3, while reserving the remaining sample for the CAPI mode so as to avoid losing a large portion of the sample to ability-based non-response (e.g., failure to complete a survey due to impairment) – a strategy that constitutes an unbalanced design with the CAPI mode sample being substantially larger than the CATI mode sample. Table 3 displays the logic of this design along with our assumptions about how the variance might differ across the 6 design groups. Recall from Table 1 that 0.55 is the observed standard deviation for the 'recommend' item from the field test of CAHPS Home Health survey, a survey of a population that is somewhat similar to the HCBS population.

[4] Simon, S.E. & Hodges, M. R. (2011). Money follows the person: change in participant experience during the first year of community living. *Reports from the Field*, 6. Mathematica Policy Research.

Table 3. Mode Test Design Grid

Likelihood of Being Able to Complete CATI	CATI	Estimated StDev	CAPI	Estimated StDev	Average StDev across Modes
Good	min # needed per group to test mode	0.55	same as min # needed for CATI + remainder of sample from this group	0.55	0.55
Probable	min # needed per group to test mode	0.75	same as min # needed for CATI + remainder of sample from this group	0.60	0.68
Limited	min # needed per group to test mode	1.0	same as min # needed for CATI + remainder of sample from this group	0.75	0.88

Note: 0.55 is the observed standard deviation for the 'recommend' item (range 1 to 4) from the field test data from CAHPS for Home Health Care. The largest agency-level standard deviation for that item was 1.2, and the average among the 10 agencies that displayed the greatest variance of responses to this item was 0.74. The estimated variances (standard deviations) shown in this table are based on the patterns observed in those field test data.

The design assumes that, among those with a *good* likelihood of being able to complete a CATI survey, the variance across modes will be essentially the same. This variance was set equal to the observed overall standard deviation from HH CAHPS, similar to what was done in Table 1 above. However, we assume that the overall variance will increase as the likelihood of being able to complete a survey decreases, and that this variance will increase more for CATI. In other words, those respondents less able to complete a survey interview will exhibit greater variation of responses to substantive questions, as well as a lower propensity to respond, and this pattern will be even more pronounced when the survey is administered over the telephone as opposed to face-to-face, since the former mode will be more difficult for those in the *probable* and *limited* groupings.

Because of the potential difference in response variance across the ability levels, we decided that it would be important to make separate mode effect estimates within each ability level, which requires a separate power analysis for each of the ability groupings. Because of the different variances across ability levels, the mean difference associated with a medium effect size increases as ability decreases (see Table 4). We are not interested in detecting small effect sizes, because they would be of little practical significance, so the power analysis uses successively larger mean differences (mean differences that constitute medium ESs) as inputs within each ability level. If we were trying to detect the same mean difference within each level (0.28, for example), we would need much bigger samples for those populations that we expect to have the most difficulty with CATI. As shown in Table 4, the unbalanced design results in the need for larger sample sizes as the hypothesized likelihood of completing a CATI survey decreases.

Table 4. Sample Size Estimates for Mode Test

Likelihood of Being Able to Complete CATI	ES	Mean Diff	Estimated Number of Completes Needed for Each Mode	Yield Rate	Total Sample per Mode
Good	0.51	0.28	350	0.5	700
Probable	0.51	0.35	410	0.5	820
Limited	0.51	0.45	430	0.5	860

ES calculation based on average estimated standard deviation within each ability group: 0.55 for *good*, 0.68 for *probable*, and 0.88 for *limited* (see Table 3). Mean differences listed are those that will yield a medium ES based on the variance assumptions in each ability group. Estimates based on a two-sample t-test that tests the difference in CATI vs. CAPI means within each ability group, and an unbalanced design (about 9 times more sample allocated to CAPI).

Since the total of 700 for the CATI mode within the *good* likelihood group is more than half of the total portion of the sample we expect to have in this group (see Table 5), we may have to simply randomly allocate half of this group to each mode. For now, we have left the allocation as is; when we have concrete data on the population distributions, we can revise the allocation plan.

Table 5. Allocation of Initial Sample for Mode Test

Likelihood of Being Able to Complete CATI	CATI	CAPI	Total
Good	700	434	1,134
Probable	820	8,576	9,396
Limited	860	6,610	7,470
Total	**2,380**	**15,620**	**18,000**

Experimental design for response option study

We will conduct an experiment to study the effects of using two different response option formats. The two formats include a four point Likert response scale (Never, Sometimes, Usually, Always) and a binary response set (Mostly yes, Mostly no). Previous research with a similar population found that some people will be unable to respond using a Likert scale but may be able to use a binary option that is less cognitively burdensome (Kane et al., 2004). Within each mode (CATI and CAPI) participants will be randomly assigned into one of our two response option formats. Based on our mode experiment sample sizes we had 2,380 assigned to CATI and 15,620 assigned to CAPI. Within each of these samples half will be randomly assigned the survey with Likert response options and the other half will be assigned to the binary response options. Respondents who are assigned the Likert response options but who are unable to answer the questions after

three question attempts with those response options are then switched to the binary response options. Switching them to the binary response options allows us to keep as much data as possible for the field test. All participants who are assigned the binary response options will be asked three questions using the Likert response options at the end of their interview to see if they can use this response format. Respondents who are unable to respond to the Likert scale format, no matter which response options they are originally assigned, will not be included in the response option experiment when calibrating survey responses from the two response formats onto the same metric, but they will still be used for other analyses (i.e., rank-ordering programs by performance relative to a benchmark, and the experiment of mode test).

Table 6 shows the survey questions administered to four types of participants, given their random assignment and their ability to respond using the Likert response options. All participants, regardless of which response options they are assigned, are asked the same three questions in the Likert scale format that are located early on in the survey, which we herein define as the "anchor questions." For the sample that will be answered the binary response set, they will answer these items after the survey is completed.

Table 6. Survey Questions Administered Based on Random Assignment and Ability to Respond to Likert Scale

Ability to Respond to Likert Scale	Random Assignment	
	Likert Scale Response Options	**Binary Response Options**
Able	***Subsample 1:*** All Questions in Likert Scale Format	***Subsample 2:*** 3 Anchor Questions in Likert Scale Format + All Questions in Binary Format
Unable	***Subsample 3:*** 3 Anchor Questions in Likert Scale Format + All Questions in Binary Format	***Subsample 4:*** 3 Anchor Questions in Likert Scale Format + All Questions in Binary Format

For the response option study, we will conduct analyses to answer the following three research questions:

1. Who can respond using the Likert scale format?
2. What is the best approach to match responses across the two response formats, so that data from different response options can be combined?
3. To pool the two response formats, can the same transformation approach be used for both CATI and CAPI mode?

To answer Research Question 1, we will calculate the proportion of those who were unable to answer all three anchor questions using the Likert scale format, regardless of which response format they are originally assigned (i.e., the proportion of subsample 3 and 4 in Table 6). We will then test if the ability of responding to the Likert scale format is influenced by a variety of

participant characteristics such as age, gender, race, etc. In other words, we want to know if participants who can respond using the Likert scale are different from those who are unable to use that response format, in terms of their personal characteristics.

To answer Research Question 2, we will use a z-transformation method to pool the responses from the Likert and binary response sets so we can analyze the combined data and compare the distributions of our outcome variables across response option formats and across test modes. Z-scores transform raw scores from different response formats into a standard metric so they can be compared on the same distribution. The z-score represents the number of standard deviations that the raw score is from the mean of the distribution. Therefore, the distribution of z scores always has a mean of 0 and a standard deviation of 1. To start, we will use all participants who are able to respond to both Likert response options and binary response options (i.e., subsample 1 and 2 in Table 6) and apply a z-transformation for each item, separately for the Likert and binary response options. We will then examine where the binary responses "Mostly yes" and "Mostly no" correspond to the Likert response scale by converting the z-scores into the metric of the Likert scale. This will give us an indication of where people using the binary response would be on the Likert response scale. Finally, we will adopt the same recoding scheme to transform the binary responses into their corresponding Likert scale responses for those who are unable to respond using the Likert scale format (i.e., subsample 3 and 4 in Table 6). The transformed responses in the metric of Likert scale will allow us to combine data from the two response option formats and conduct our analyses using these transformed scores without loss of data.

To answer Research Question 3, we will include all participants who are able to respond to both Likert and binary response options (i.e., subsample 1 and 2 in Table 6) and use their transformed scores in the metric of Likert scale to test for mean differences between the two response formats, within each test mode (CATI or CAPI), at the domain level, and then test the equality of the coefficients to determine if there is a statistically significant interaction effect between mode and response option format. A significant interaction indicates that the relationship between the binary and the Likert scale responses varies by the test mode. In other words, the responses that 'mostly yes' and 'mostly no' correspond to in the Likert metric will differ across the CATI and CAPI modes. Therefore, we will need to conduct the z transformation separately for those assigned to CATI and CAPI.

We conduct power analysis to ensure an adequate sample size for detecting a small-sized effect of response format within each test mode (CATI and CAPI). Assuming an effect size of 0.2 and a balanced design across the two response options (i.e., equal number of participants randomly assigned to the two response options and equal variance of experiences across the two groups), a total of 393 completes is needed for each response option group. With a yield rate of 0.5 and assuming at least 70% of the participants will be able to respond using the Likert scale format, a total of 1123 people is needed for each test mode and each response format. The percentage of 70% is somewhat based on data presented in Table 2, but there is no concrete evidence the percentage may vary for this field test. Thus, the following allocation of initial sample (see Table 7) will allow us to have a sufficient sample to test the performance of the z-score transformation across different test mode.

Table 7. Allocation of Initial Sample for Response Format Test

Assignment of Response Option Format	CATI	CAPI	Total
Likert Scale Format	1,190	7,810	9,000
Binary Format	1,190	7,810	9,000
Total	**2,380**	**15,620**	**18,000**

2. Information Collection Procedures

Data will be collected by a survey vendor. We will use two different modes: Face-to-face and telephone, which will include these steps:

Face-to-face

- Send cover letter and consent information. We will send a letter to potential respondents alerting them to the survey.
- Initial contact: Two weeks after the cover letter is sent, we will make up to 10 calls to contact the person and set up an interview. Four calls will be made during daytime on weekdays, three times during nighttime on weekdays, and three times during the day on weekends. Per the CAHPS guidelines, we will spread these 10 calls over different weeks.
- Reminder: Interviewer will contact the respondent 1 to 3 days ahead of time to remind him or her of the interview.
- Interview: Interviewer will meet the respondent at his or her home to conduct the interview; written consent will be obtained prior to the interview

Telephone arm

- Send cover letter and consent information. We will send a letter to potential respondents alerting them to the survey.
- Interview: Two weeks after the cover letter is sent, we will make up to 10 calls to contact the person and complete the survey on the telephone. Verbal consent will be obtained prior to the interview. Four calls will be made during daytime on weekdays, three times during nighttime on weekdays, and three times during the day on weekends. Per the CAHPS guidelines, we will spread these 10 calls over different weeks.

Based on the sampling goals, power analysis, and design assumptions outlined above, we plan to implement a 2-stage sampling methodology:

- Stage 1: We will draw a Stratified Random Sample of 600 persons from each program within each state – the combination of state and program will constitute the strata. We have

assumed that a sample *size* sufficient for goal 2 will automatically include enough sampled individuals for goals 1 and 3 as well. This will result in a total sample of 18,000

- Stage 2: At this stage, we will divide the entire sample of 18,000 people into three subgroups (good, probable, and limited) based on the population of which they are a member. For example, all sampled members in program populations with intellectual or developmental disabilities would be categorized in the *limited* group, while all those in a program serving the aged would be categorized in the *probable* group. Within each subgroup, we will randomly assign the required number of sample members to the CATI mode, and the rest to the CAPI mode. Thus, for the mode test, we will have an unbalanced design with around 2,400 participants from the CATI mode and approximately 15,600 participants from the CAPI mode. This design results in about 13% of the total sample being allocated to CATI, and 87% to CAPI.

Drawing the Sample

The SURVEYSELECT procedure in SAS provides a variety of methods for selecting probability-based random samples. The procedure can select a simple random sample or a sample according to a complex multistage sample design that includes stratification, clustering, and unequal probabilities of selection. With probability sampling, each unit in the survey population has a known, positive probability of selection. This property of probability sampling avoids selection bias and enables you to use statistical theory to make valid inferences from the sample to the survey population.

The SURVEYSELECT procedure can perform stratified sampling, selecting samples independently within the specified strata, or non-overlapping subgroups of the survey population. Stratification controls the distribution of the sample size in the strata. It is widely used in practice towards meeting a variety of survey objectives. For example, with stratification you can ensure adequate sample sizes for subgroups of interest, including small subgroups, or you can use stratification towards improving the precision of the overall estimates. When you are using a systematic or sequential selection method, the SURVEYSELECT procedure also can sort by control variables within strata for the additional control of implicit stratification. The procedure automatically calculates selection probabilities and sampling weights.

3. Methods to Maximize Response Rates

Every effort will be made to maximize the response rate, while retaining the voluntary nature of the effort. We will contact eligible participants and explain what the survey is about, who is doing it and why, and provide contact information for questions. The research team headed by CMS and its contractor will work closely with the individual states to minimize burden on the programs and on the individual HCBS recipients.

Non-response Analysis

It is important to consider if responders differ from non-responders anytime that the response rate is lower than 80%; response bias could be present if there is evidence that the responding population differed in important ways from the non-responding population. In the likely event that the response rate will be less than 80%, we will conduct a response bias analysis on the subset of individuals that had been subject to the full protocol. We will compare responders to non-responders by available frame characteristics, including gender, age, main reason for receiving HCBS, living situation (e.g. congregate setting, own home or with family) and receipt of select services (e.g. employment supports, personal care).

5. Tests of Procedures

The survey development team conducted formative research to ascertain dimensions of care important to consumers. Using that data, as well as an environmental scan of existing surveys, the survey development team developed a draft survey. The draft survey underwent three rounds of cognitive testing in English, including an experiment to determine the best method to ask questions for respondents with intellectual impairments. The team also conducted one round of cognitive testing in Spanish. Due to the iterative nature of the testing, tests calling for identical questions were conducted with 9 or fewer respondents only. These respondents included individuals with a range of physical, mental, and cognitive disabilities. The results of each round were used to make iterative changes to survey items and response patterns.

6. Statistical Consultants

Provide the name and telephone number of individuals consulted on statistical aspects of the design and the name of the Program unit, contractor(s), grantee(s), or other person(s) who will actually collect and/or analyze the information for the Program.

The project is led by Julie Seibert, PhD, of Truven Health Analytics at (919) 475-6225.

This sampling and statistical plan was prepared and reviewed by staff of CMS and by the American Institutes of Research. The primary statistical design was provided by Chris Evensen, MS, of the American Institutes for Research at (919) 918-2310, Steven Garfinkel, PhD, of the American Institutes for Research at (919) 918-2306, Manshu Yang, PhD, of the American Institutes for Research at (919) 918-2312 and HarmoniJoie Noel, PhD, of the American Institutes for Research at (202) 403-5779.