Health Insurance Marketplace Consumer Experience Surveys: Enrollee Satisfaction Survey and Marketplace Survey

Supporting Statement—Part B
Collections of Information Employing Statistical Methods

December 16, 2014

Centers for Medicare & Medicaid Services

TABLE OF CONTENTS

Section **Page**

1. Potential Respondent Universe and Sampling Methods..2

1.1 Marketplace Survey..3
1.1.1 Respondent Universe..3
1.1.2 Sample Size Calculations..5
1.1.3 Marketplace Survey Sampling Methods..15

1.2 QHP Enrollee Survey..16
1.2.1 Respondent Universe..16
1.2.2 Sample Size Calculations..17
1.2.3 QHP Enrollee Survey Sampling Methods..20

2. Information Collection Procedures..23

3. Methods to Maximize Response Rates and Address Non-Response Bias.........................24

4. Tests of Procedures..26

5. Statistical Consultants...27

1. Potential Respondent Universe and Sampling Methods

This supporting statement includes information in support of:

- A Health Insurance Marketplace (HIM) Survey ("Marketplace Survey")
- A Qualified Health Plan Enrollee Experience Survey ("QHP Enrollee Survey")

A description of the surveys and the testing goals related to each survey are provided in Part A of this submission. Testing goals that directly impact the sample size estimates are summarized in Sections 1.1.2 (for the Marketplace Survey) and 1.2.2 (for the QHP Enrollee Survey) of this document.

Both surveys will be administered in four annual rounds; one survey each year in 2014, 2015, 2016, and 2017. The first two of these rounds (2014 and 2015) are described in this submission:

1. A **psychometric test** of each survey in 2014 using a single survey vendor. The goal of each psychometric test is to evaluate the reliability and validity of each survey. This goal includes assessing the measurement properties of the instrument, individual survey items, and reporting composites. It also includes testing the equivalence of these measurement properties across language and mode of administration. Because the QHP Enrollee Survey includes CAHPS 5.0 Health Plan Survey core items and some CAHPS supplemental items sets, another goal for the QHP survey will be to determine the extent to which the measurement properties of the existing CAHPS items hold for the QHP population, which will include persons who have been previously uninsured. Results of psychometric testing will inform revisions to both surveys, including shortening the survey instruments and reducing respondent burden.

2. A **beta test** of the revised survey in 2015. The main goals for the Marketplace Survey beta test are 1) to produce assessment scores in each Marketplace for composites, global ratings, and individual report items from the tested and revised Marketplace survey and use these scores to provide initial feedback to states, and 2) to the extent feasible (given sample size limitations), conduct subgroup analyses to determine if disparities in consumer experiences by race, ethnicity, income, and disability exist within *each* Marketplace to help CMS meet its regulatory oversight requirements. A secondary goal is to rerun the the psychometrics to confirm the psychometric properties of the revised instrument on a larger, state-based sample. The goals for the QHP Enrollee Survey beta test, which will involve data collection by multiple survey vendors hired by the issuers in each State, are to test the vendor system (this is explained in more detail in Part A), verify the psychometric properties of the revised QHP instrument on a larger, national sample, and provide initial feedback to QHPs regarding the data collection, submission and calculation processes before public reporting begins in 2016.

As of December 2014, the psychometric test phase of the Marketplace survey is close to completion and the data collection period for the psychometric test of the QHP Enrollee survey has closed. The beta test round of the QHP Enrollee Survey has been approved, CMS has begun providing materials to approved survey vendors, and survey vendor training is underway. CMS is now seeking expedited clearance for the beta test round of the Marketplace survey. We request approval as soon as possible because sampling for the Marketplace Survey beta test is set to begin in mid-February 2015 and data collection about one month later.

Robust results from the psychometric analysis are obtained by capturing, to the greatest extent possible, the full range of experiences and response patterns in the population. For this reason, the construction of the sampling frame and the sampling methods for the psychometric tests were designed to capture this full range of experiences of the populations within each State or QHP. In addition, there are other considerations related to reliability and validity that guide the sample size estimates and sampling methods, and these considerations are described in the sections that follow. The psychometric test component of this information collection is not designed to provide state-level or QHP-level estimates; only the aggregate results of this analysis will be discussed or disseminated. The second component of this information collection (in 2015) includes a provision for initial feedback to States and QHPs, subject to the limitations associated with response rates (including non-response bias analyses) and data quality findings.

In the following sections, we first address the Marketplace Survey, then the QHP Enrollee Survey. For each survey, we describe the respondent universe, sample size calculations, and sampling methods separately for the psychometric test and the beta test.

1.1 Marketplace Survey

1.1.1 Respondent Universe, Study Population, and Sampling Frame

The *respondent universe* is the theoretical population to which findings apply. For the Marketplace Survey, the respondent universe includes any adult (age 18+) eligible for health insurance coverage offered through the Health Insurance Marketplaces, including those eligible for Medicaid coverage.

The *study population* consists of individuals to whom it is possible to gain access. For the Marketplace Survey, the study population consists of consumers who interacted with the Marketplaces within a specified time frame. This definition includes any adult who has at a minimum provided their contact information, regardless of how far they have gotten in the application and enrollment process. This definition also includes consumers who enter their information themselves through the website, submit a paper application, or have the information entered for them by a telephone or in-person assistor. [1]

Individuals in the study population are classified as one of four types of Marketplace consumers: (1) effectuated QHP enrollees (those who have enrolled in a QHP and paid their first premium), (2) QHP enrollees who have not yet paid their premium (enrolled but not effectuated), (3) those who have accessed a Marketplace, completed and submitted an application, but have not enrolled in a QHP, and (4) those who have accessed a Marketplace and entered contact information, but who have not yet completed the application and thus have not yet selected and enrolled in a QHP.

[1] For consumers applying by phone or in-person, representatives still enter their data in the web site (either Healthcare.gov or an SBM's dedicated state-based web site), and thus we assume that a phone or in-person assisted application can be partially completed, and that a consumer applying by phone or in-person may not yet have enrolled in a QHP. Paper applications are also entered using the web site but could also be incomplete, and some applicants submitting paper applications may not yet have enrolled in a QHP at the time of sampling.

Due to limited resources and the time required for translation of the survey materials, the study population is also limited to those who prefer to speak and read in one or more of three languages: English, Chinese, and Spanish.

All 50 States and the District of Columbia (D.C.)[2] are classified in one of two groups based on how their Marketplaces are organized:

- State-based Marketplaces (SBMs)—include 15 States (including D.C.), all of which are currently running their own State-specific exchange web sites for enrollment. Eventually, the SBMs will transmit their individual-level application and enrollment data to CMS.
- Federally-facilitated Marketplace (FFM)—includes the remaining 36 States where enrollment operates through the federal government's web site, Healthcare.gov. Although enrollment in these 36 States operates through a single website, the FFM comprises 36 distinct markets, since the issuers and their associated sets of plan offerings are unique in each State, and in-person assistance will vary by state. For purposes of the Marketplace Survey, we consider State Partnership Marketplaces (SPMs) and Supported State Based Marketplaces (SSBMs) to be a part of the FFM because both use the FFM for enrollment and eligibility functions.

The *sampling frame* is the list of individuals from which the study sample is taken. For the Marketplace Survey, list-based sampling frames were obtained from CMS databases. The sampling frames included all available records for the four distinct groups of eligible consumers. As described above, these four groups are defined by their applicant status:

1. Potential applicant (PA) – consumers who have completed any step prior to submitting an application, after providing contact information,
2. Potential enrollee (PE) – consumers who have successfully completed and submitted an application that includes their family size and income information,
3. Enrollee (E) – consumers who have selected a QHP from their Marketplace, and
4. Effectuated enrollee (EE) – QHP enrollees who have made their first premium payment to the selected QHP issuer.

1.1.1.1 Psychometric Test

The psychometric test component only included participants in the 36 FFM states. CMS constructed a sampling frame for the 2014 psychometric test using the administrative data from the 36 FFM states contained in the databases in which application and enrollment information is stored. The frame included individuals who provided contact information at any point from October 1, 2013 to March 30, 2014.

1.1.1.2 Beta Test Phase

The study population is the same as for the psychometric test component, with the exception of the time frame: any adult (age 18+) who, at any point from November 15, 2014 to February 15, 2015, has at a minimum provided their contact information, regardless of how far they have gotten in the application and enrollment process.

[2] For the proposed data collections we classify D.C. as a "State," hence there are reference to "51 States" in this document.

In the time between sampling activities for the psychometric test and the beta test, CMS worked with its contractors to try to resolve the sampling frame limitations associated with State Based Marketplaces. During this time, CMS determined that participation in the 2015 beta test would be optional and CMS and its contracotrs worked with the 14 states that planned to operate a SBM for the 2014–2015 Open Enrollment Period to determine whether SBMs were interested in participating in the beta test. Through these efforts, seven SBM states (CA, CT, HI, KY, MN, RI, WA) have indicated that they will voluntarily participate in the beta test. In total, the Marketplace Survey beta test will be conducted in 44 states, including the 37 states using the FFM and the 7 SBM states who have agreed to participate.

1.1.2 Sample Size Calculations

Sample size is calculated first by determining the minimum number of completed responses needed to meet the goals of the data collection, and second by inflating that number by a large enough factor to account for the estimated rate of survey non-response.

We will follow American Association for Public Opinion Research (AAPOR) guidelines in calculating response rates. The response rate is the result of dividing the number of completed interviews/questionnaires by the number of eligible respondents who were selected to participate. Potential respondents fall into the following categories:

1. Eligible and interview completed (c).

2. Eligible and not interviewed (e).

3. Ineligible (e.g., out of scope; only potential respondents who have explicitly indicated ineligibility are included here) (i).

4. Unable to determine eligibility (u).

According to AAPOR guidelines, the total number of participants selected to be surveyed (n) is the sum of eligible and completed (c), eligible and not interviewed (e), ineligible (i), and unable to determine eligibility (u). That is n = c + e + i + u. By design, our survey sampling frames will only include eligible individuals, with eligibility determined using administrative data from CMS databases. However, among those with unknown eligibility (u), there is likely to be a small proportion (x) who may in fact be ineligible. This proportion (u) will be estimated using the following formula:

$$x = \frac{c+e}{c+e+i}$$

The response rate will then be calculated as:

$$Response\,Rate\,(RR) = \frac{c}{c+e+(x*u)}$$

In the above formula, the denominator includes all original survey units that were identified as being eligible, including units with pending responses with no data received, post office returns because of “undeliverable as addressed,” and any new eligible units added to the survey. The denominator will not include units deemed out-of-scope, or duplicates.

Sometimes only partial interviews will be obtained due to a respondent's breaking off an interview or completing only part of a mailed questionnaire. For the proposed data collections, CMS will follow the CAHPS standard: a questionnaire will be considered complete if responses are available for 50% or more of a selected list of key survey items—the items that all respondents are eligible to answer.

Response Rate Estimates

Total required sample size is a function of the purpose of a given component of this study and the desired number of completes divided by the estimated overall response rate calculated as described above. Historically, response rates for CAHPS surveys span a fairly wide range. The 2012 Commercial Health Plan CAHPS response rates are approximately 30% and Medicaid CAHPS response rates are approximately 27%. In the recent psychometric test of the CAHPS survey for Cancer Care, the response rate was 48%; for the Dental CAHPS psychometric test and early implementation, response rates ranged from as low as 40% to as high as 70% in some population segments. Based on experience with psychometric tests of several different CAHPS instruments, and in light of the relatively low response rates obtained with the Medicaid population (which is similar to the Marketplace population) and with the Commercial Health Plan CAHPS survey (which is similar to our QHP Enrollee survey), CMS assumed that the overall response rate for both surveys would be 30%. Observed response rates from the psychometric test varied by both mode and language.

For the psychometric test, the Spanish and Chinese versions of the survey were conducted via mail only. This decision was made because the small Chinese and Spanish samples proposed did not justify the expense of developing Chinese and Spanish versions of the CATI and Internet programs. We hypothesized that the lack of a phone option for individuals who prefer Spanish and Chinese might reduce the likelihood that those individuals would respond to or complete the survey. Observed response rates in the psychometric test were indeed lower for consumers with a Spanish language preference, but they were higher for those with a Chinese language preference.

Another problem with using a mail-only mode with non-English consumers is the potential bias stemming from situations where non-English speakers have lower literacy. We reasoned that such individuals would be better able to complete a phone survey compared to a mail or web survey. Thus, limiting data collection to the mail mode only might have excluded some of the more vulnerable populations from the psychometric test data collection – non-English speakers who have trouble reading or writing, even in their native language. For the beta test, all sampled consumers will be surveyed using the mail with phone follow-up mode and thus we can compare phone responders to mail responders to help determine if phone responders in Chinese and Spanish differ from mail responders in those languages.

For the beta test, the response rate assumptions have been modified based on the results from the psychometric testing phase. Response rates achieved during the psychometric testing of the Marketplace Survey were around 24% on average, but, as mentioned above, varied by both mode and language. Mail with phone follow-up obtained the best response rate (32%), which is why CMS will be using that mode with all three languages during the beta test. Since, in the psychometric test, this mode included only those consumers in the English sample, we do not know at this time what response rates will be achieved using this mode among the consumers with a Chinese or Spanish language preference.

During the psychometric test, the response rates for the Chinese, English, and Spanish samples were 36 percent, 32 percent, and 25 percent respectively. The use of a data collection method that includes phone follow-up may increase the response rates for the Spanish and Chinese consumers. Among the English language sample, the response rate for the mail-only mode was only 20%, and thus the phone follow-up appears to produce a 12 percentage point increase in the response rate (an increase of 60%). Comparable increases among the Chinese and Spanish samples during the beta test would result in response rates of 57 percent and 40 percent respectively; however, it is unlikely that actual response rates will be this high. Thus, for the beta test, we will use the response rates observed from the psychometric test data collection in the beta test sample size calculations. After the beta test, additional analyses will be conducted to determine whether any additional changes are needed to the data collection methods for future versions of this survey.

1.1.2.1 Marketplace Psychometric Test Sample Size Estimates

Our sample size estimates for the Marketplace Survey psychometric test reflected the sample sizes necessary for fully evaluating reliability and validity of the instrument.

Reliability testing included the evaluation of:

- Internal consistency reliability (ICR) of proposed composites (as indicated by Cronbach's alpha)
- Equivalence reliability, which tests the consistency of measures across mode and language
- Unit-level reliability, which tests the extent to which a measure score differentiates signal (i.e., differences in scores across reporting entities, such as Marketplaces or QHPs) from noise (i.e., random measurement error); also referred to as inter-unit reliability (IUR)

Face validity (the survey questions are representative of the concepts they are supposed to reflect) was established via the formative research—the review of existing instruments, focus groups, input from a technical expert panel and other stakeholders, and the cognitive testing (described in Section 4 below).

Construct validity was assessed using confirmatory factor analysis (CFA) and multi-trait analysis. The CFA tested the fit of the data to the factor structure, generated factor loadings, and performed statistical tests of those loadings. The multi-trait analysis compared the correlations of items with their composite total (correcting for overlap[3]) to the correlations of those items with competing composites, and estimates an indicator (scaling success) of discriminant validity.

[3] Howard KI, Forehand GG. A method for correcting item-total correlations for the effect of relevant item inclusion. *Educ Psychol Meas*. 1962; 22 (4), 731-735.

In CAHPS, there are two statistics used to assess unit-level reliability.[4] One is a measure of IUR based on the F-statistic from an analysis of variance (ANOVA). The IUR is equal to (F-1)/F, which is a summary measure of the between-unit variance minus the within-unit variance over the between-unit variance.[5] The other measure is the intra-class correlation (ICC), which is also calculated using statistics produced by an ANOVA. The ICC in this context is the between-unit variance minus the within-unit variance over the total variance adjusted for the average number of respondents per reporting unit.[6] The IUR provides the reliability based on the sample size associated with the data, while the ICC indicates the reliability of a measure for a single respondent. The reliability coefficient can take any value from 0.0 to 1.0, where 1.0 signifies a measure for which every respondent reports an experience identical to every other respondent evaluating the same unit. Scales with reliability coefficients above 0.70 provide adequate precision for use in statistical analysis of unit-level comparisons,[7] though it has been argued that measures with reliability coefficients of at least 0.90 are optimal.[8]

Since unit-level reliability is partly a function of sample size, the IUR allows for the calculation of the number of respondents needed per reporting unit to obtain a particular level of reliability (similar to a power analysis) in *future* data collections, and thus it is especially important with respect to future respondent burden.[9] For the psychometric test, it was not necessary to obtain an IUR of at least 0.70 for the final recommended measures. However, to be useful for making sample size recommendations for future rounds of data collection, past experience demonstrates that it is best to have data from all accountable units when the universe of accountable units is finite (as with the FFM states); where the universe of accountable units is theoretically not finite (as with QHPs), it is best to have data from at least 30 accountable units selected across the full range of unit performance (i.e., from the poorest performing units to the best performing units).

Our sample size recommendations were based on our estimate of the minimum number of responses per equivalence group (i.e., mode and language groups) needed at the national level to conduct the psychometric analyses described above. This estimate is described in more detail immediately below. In order to evaluate unit-level reliability, which requires that we have a consistent number of completed surveys from each state, we propose distributing the total national sample evenly across the 36 states in the FFM. This strategy is presented in more detail in Section 1.1.3.1, which describes our proposed sampling methods for the Marketplace survey psychometric test.

[4] For a discussion of the methods used to calculate the reliability of CAHPS measures, see pp. 62-63 in the document "Instructions for Analyzing Data from CAHPS® Surveys: Using the CAHPS Analysis Program Version 4.1," Document No. 2015, updated on 04/02/2012; available here: Instructions for Analyzing Data from CAHPS Surveys: Using the CAHPS Analysis Program Version 4.1. Much of the text in this section is based on information provided in that document.

[5] Winer BJ. Statistical principles in experimental design. New York: McGraw-Hill, 1970; also Zaslavsky AM, Buntin MJB. Using survey measures to assess risk selection among Medicare Managed care plans. *Inquiry*, 6/2002, 39(2), 138-151.

[6] Hays RD, Revicki D. Reliability and validity (including responsiveness). In P. Fayers & R. Hays (eds.). *Assessing quality of life in clinical trials: Methods and practices,* 2nd ed. Oxford: Oxford University Press, 2005, 41-53.

[7] Nunnally, JC (1978). *Psychometric theory* (2nd edition). New York: McGraw-Hill Book Company.

[8] Zaslavsky AM, Statistical issues in reporting quality data: small samples and casemix variation, *Int J Qual Health Care*, 2001;13(6):481-488.

[9] For a discussion of reliability and its relationship to sample size, see the document, "Fielding the CAHPS Clinician & Group Surveys: Sampling Guidelines and Protocols (Document No. 1033)," available here: Fielding the CAHPS Clinician & Group Surveys: Sampling Guidelines and Protocols (Document No. 1033).

Sample Size

Factor analyses, multi-trait analyses, and the estimates of equivalence and internal consistency reliability were conducted separately for each survey administration mode using all complete responses from eligible sample members across the whole FFM. The generalizability of results from psychometric analyses depends on capturing the full range of covariance among consumer experiencesin the population. Standard psychometric practice is to obtain a *minimum* of 10 *complete responses* for each assessment item used in the psychometric analysis (including substantive questions combined into composites, but not screeners, ‘About You’ items, or questions designed to determine survey eligibility). This recommendation is grounded in sound measurement theory[10] and practice in the statistical analysis of multivariate data (including factor analyses).[11]

The Marketplace survey included 30 assessment items, which translates into a *minimum* of 300 completed surveys nationwide if each completed survey contained a non-missing response for each substantive item. However, we expected that *some* substantive items would be legitimately skipped by respondents to whom the subject matter of the item did not apply; therefore, this number needed to be larger. In addition, we expected that some completed surveys would have some degree of item non-response (when a respondent skips an item that he/she should have answered). Thus, we proposed a target of 15 complete responses at minimum for each assessment item. This translated into a minimum number of completes of 450 (15*30) per group if psychometric analyses conducted separately for each group. For surveys conducted in English, there were five mode experiment groups (telephone-only, mail with telephone follow-up, mail-only with 1st class mail follow-up, mail-only with FedEx follow-up mailing, and web-only), and thus we needed a minimum of 2,250 completed surveys to conduct psychometric testing for each of the five modes. In addition, we aimed for 450 completed surveys each for both the Spanish and Chinese surveys to conduct psychometric analyses separately for each language (only one administration mode was planned for Spanish and Chinese). We expected that this approach would result in an overall total of 3,150 completed surveys (2,250 in English, 450 in Spanish, and 450 in Chinese).

Exhibit B1 shows the expected distribution of the English language completes across the five experimental mode design groups, plus the required number in Chinese and Spanish.

Equal numbers of consumers who indicated that English was their preferred language were randomly sampled within each State and then randomly assigned across the five mode groups so as to obtain 450 completed surveys in each of the five experimental groups.

Exhibit B1. Sample Sizes and Completed Survey Counts for the Marketplace Psychometric Test

Mode†	Target Number of Completed Surveys	Total Number to Sample
Exp 1. Phone only	450	1,500

[10] Nunnally JC & Bernstein IH (1994). *Psychometric theory (3rd Edition).* New York: McGraw-Hill, Inc.

[11] Stevens J (1992). *Applied multivariate statistics for the social sciences (2nd Edition).* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Mode†	Target Number of Completed Surveys	Total Number to Sample
Exp 2. Mail with phone	450	1,500
Exp 3. Mail only with third survey mailed Fed Ex	450	1,500
Exp 4a. Web only – email and pre-notification letter	225	750
Exp 4b. Web only – email only	225	750
Exp 5. Mail only	450	1,500
Total English	**2,250**	**7,500**
Non-English		
Spanish (mail only)	450	1,500
Chinese (mail only)	450	1,500
Overall Total	**3,150**	**10,500**

† Mode experiments will be conducted in English only. All modes other than the mail-only mode (Exp. 5) will be available only to respondents whose language preference is English.

Limitations

The psychometric test component of this information collection was not designed to provide state-level estimates.

CMS was not able to evaluate psychometric properties of the instrument among the 15 SBMs. While this is a serious weakness, it was unavoidable during the psychometric test phase.

1.1.2.2 Marketplace and Beta Test Phase Sample Size Estimates

This component of the implementation will involve the initial fielding of the revised survey in 44 states, including the 37 states that utilize the FFM for enrollment and eligibility along with 7 SBMs that have voluntarily agreed to participate in the beta test.

As described above, to have a sufficient number of responses for *analysis* and *reporting* based on surveys where respondents may interact with a number of different individuals or systems, such as with a health plan or a clinician group, CAHPS generally recommends obtaining completed questionnaires from 300 respondents per reporting entity.[12] These estimates are based on

[12] For health plans, CAHPS recommends a target of 300 completed suveys per plan with a minimum of 100 for reporting. See p. 5 in the document "Fielding the CAHPS® Health Plan Survey 4.0: Commercial Version," Document No. 13b, available here: Fielding the CAHPS® Health Plan Survey 4.0: Commercial Version, and p. 65 in the document "Instructions for Analyzing Data from CAHPS® Surveys: Using the CAHPS Analysis Program Version 4.1," Document No. 2015, updated on 04/02/2012; available here: Instructions for Analyzing Data from CAHPS® Surveys: Using the CAHPS Analysis Program Version 4.1. For clinician groups, CAHPS recommends 300 completed surveys per group. See p. 7 in the document, "Fielding the CAHPS Clinician & Group Surveys: Sampling Guidelines and Protocols," Document No. 1033, updated on 09/01/2011; available here: Fielding the CAHPS Clinician & Group Surveys: Sampling Guidelines and Protocols.

analyses conducted on existing CAHPS data to determine the number of completed responses needed to provide power sufficient to detect differences between one reporting entity (e.g., a health plan) and the mean of all other reporting entities in a given sample. These differences are the basis of the standard CAHPS "star rating," which identifies reporting entities as being below average, average, or above average.

We have assumed that interactions with a Marketplace or a QHP will be analogous to this heterogeneous experience, which implies that 300 completed responses per Marketplace would be sufficient for standard CAHPS analysis and reporting activities. However, regulatory oversight requires CMS to determine if disparities in consumer experiences by race, ethnicity, income, and disability exist *within* each state. Such subgroup analyses would involve, for example, comparing the experiences of a small group in a given State, such as Hispanics, to a large group in that State, such as non-Hispanics. Additionally, under the Americans with Disabilites Act, CMS is required to ensure that individuals with a disability have equal access to government services and the Marketplace Survey provides an opportunity for CMS to ensure that it is fufilling this requirement within the Health Insurance Marketplaces. To meet these oversight requirements, a greater number of complete responses are needed from each Marketplace.

To accommodate this objective in the beta test, CMS proposes sampling to obtain 1,200 completed surveys in each state. Our analyses of the psychometric test data show that, at the low end, the prevalence of most of the subgroups of interest to CMS is between 1 percent and 9 percent in most, but not all, states. Our analyses show that 1,200 completed surveys would enable CMS to detect subgroup disparities with corresponding effect sizes ranging from high-medium to low-medium (effect sizes between 0.63 and 0.32) for any subgroup comprising between 2 percent (n ≈ 20) and 7 percent (n ≈ 80) of a given state's Marketplace Survey-eligible population.

Based on the psychometric test results we found higher response rates for Chinese (36%) and Spanish (25%) compared to English (20%) when using the same mail-only mode. Our survey vendor, Ipsos, has found similar results for other health surveys. For this component, CMS will not sample based on language, but will administer Spanish and Chinese versions to sampled consumers who indicated a Spanish or Chinese language preference on their marketplace applications. If no preference is given, the sampled person will be part of the English sample, If language preference is indicated as some other language, the person will be excluded from the frame.

Data from the psychometric test indicate that, for the 36 FFM states included in the psychometric test, consumers with a Spanish language preference comprise around 5% of the total survey-eligible population in those states, while consumers with a Chinese language preference comprise around 0.20% of the total survey-eligible population. Exhibit B2 shows the sample sizes and anticipated number of completes by language for the 44 beta test states *on average*. This approach will yield a total sample of 167,273 consumers resulting in a total of 52,800 completed surveys nationwide.

Exhibit B2. Distribution of Marketplace Surveys by Language for Beta Test Component

Inputs	Overall	English	Spanish	Chinese
Share of population	100%	95%	5%	0.20%
Estimated RR	31.57%	32.00%	25.00%	36.00%
Number sampled	167,273	156,420	10,560	293
Number completes	52,800	50,054	2,640	106

These estimates, however, are based on a pool of states that exclude those states currently expected to be in our beta test—such as CA, HI, and WA—where the proportion of consumers with a preference for Chinese may be much higher than we have observed in the pool of FFM states. The same can be said for Spanish with respect to CA, and perhaps other SBM states as well. When the beta test sampling frame is constructed, CMS will update, if necessary, the population share estimates shown in Exhibit B2.

Number of Chinese and Spanish Surveys Needed for Psychometrics.

Due to a lower than expected survey response rate (RR) among the Spanish language segment of the sample (we had assumed 30% but the actual RR was 25%), combined with low item-level RR for some sections of the survey, CMS was not able to obtain enough completed surveys in Spanish (n=318) and Chinese (n=540) during the psychometric test to fully evaluate the equivalence of measurement properties across language for the full set of composites. Even the higher survey response rate among Chinese was not enough to counter-balance the low item-level response rates to some survey questions. Our sample size estimates for the psychometric test were based on an assumption that item-level response rates would average around 67% (see section 1.1.2.1); in practice, item-level response rates across all languages were as low as 28% on average for some items.[13]

This rate varied substantially by language. For example, for the ‘Seeking Information In-Person’ section of the survey, the item-level response rates were approximately 20%, 50%, and 33% for English, Spanish, and Chinese respectively, which corresponds to a weighted average of 28% across all three languages (English respondents make up the majority of respondents, so their item-level response rate contributes disproportionately to the average). Thus, even though we had obtained 540 surveys completed in Chinese, only around 170 of those consumers responded to each of the questions about seeking information in-person (q37 to q45 on the psychometric test version of the survey).

The beta test version of the survey has 26 assessment items. We would thus need 260 completed surveys (10 per item) to conduct the analysis of the equivalence of measurement properties

[13] The item-level response rate indicates the proportion of respondents who provided a usable response to a given item or composite. Composite-level response rates can be higher than item item-level response rates because a composite is scored for any respondent who provided a response to at least one item in the composite. In section 1.1.2.1, we indicated that we needed a minimum of 10 complete responses for each assessment item that was to be used in the psychometric analysis. We inflated this number to 15 based on assuming an average item-level response rate of 67% to account for missing data (10/0.67 = 15). Responses can be missing for two reasons: 1) the respondent skipped the item because it was not applicable to them based on their response to a screener question (legitimate skip), and 2) the respondent did not provide a response to the question, even though it applied to them (non-response). The observed rate of legitimate skips in the psychometric test was higher than we had assumed it would be.

across the three languages. However, since composite-level and item-level response rates are all lower than 100%, we need to inflate the number of completed survey enough to ensure 260 completes for the items with the lowest response rate. As described above, the ‘Seeking Info In-Person’ items yielded an overall response rate of 28%, though the Chinese respondents had a slightly higher rate of 33%. Applying this item-level response rate to the minimum number needed (260), we would need approximately 788 completed surveys each in Spanish and Chinese to conduct this analysis (260/0.33 = 787.9).

As shown in Exhibit B2, the beta test sampling design will produce a sufficient number of Spanish-language completes (n = 2,640) to conduct the remaining psychometric analyses. However, this design will not produce the 788 Chinese completes needed for this work, and thus consumers with a Chinese langage preference will need to be oversampled. The oversampling approach is described below in Section 1.1.3.2.

As described in this section, CMS has determined the sample size based on CAHPS recommendations related to the ranking of entities and incorporating the specific demands of oversight and QI outlined above. Thus, the sections below:

1. Describe the precision of point estimates associated with various sample sizes, and
2. Describe, in the context of detecting differences between a single State and the mean of all 44 States (i.e., assigning star ratings), the effect sizes associated with various sample sizes.

1.1.2.2.1 Precision of Point Estimates

State-level and national-level estimates both rely on the precision of point estimates for the survey measures (composites, overall ratings, and single item measures). Precision is defined in terms of the margin of error, which is also known as the “half-width” of the confidence interval (typically a 95% confidence interval). The margin of error for a 95% confidence interval (CI) is equal to the standard error of the point estimate multiplied by 1.96 (the margin of error for a 68% CI would be equal to one standard error; the margin of error for a 99% CI would be equal to 2.58 standard errors). Thus, the margin of error is used to construct the CI around the point estimate and describes the range within which we can be confident the true score lies.

We estimated confidence interval precision using PROC POWER in SAS. This approach is analogous to a traditional power analysis, with the margin of error (“CI Half-Width” in SAS) taking the place of *effect size* and the half-width probability (“Prob (Width)” in SAS) taking the place of *power*. Using estimates of a range of variances and standard errors observed from some existing CAHPS surveys (e.g., the psychometric test of the draft CAHPS survey for Cancer Care, the NCQA National Distribution of 2009 Adult Medicaid CAHPS Plan-Level Results, and the 2013 Medicare Part C Report Card results) as inputs, we estimated the sample sizes associated with different levels of precision. Note that CMS has decided on a target number of completes based on standard CAHPS recommendations in combination with the oversight requirements for scoring small subgroups in each State. Thus, this analysis is designed to illustrate the level of precision that can be obtained with those samples under several scenarios.

We used a conditional probability approach (that is, the probability of achieving the desired precision is calculated conditionally given that the true mean is captured by the interval), which is a more conservative approach than the unconditional probability approach. To anchor the margins of error and variance estimates (expressed as standard deviations) to a meaningful CAHPS scale, we have transformed observed scores for the three different types of measures from the existing CAHPS results mentioned above into a 100-pt scale. This transformation expresses the inputs to the power analysis in a scale that is comparable across different types of measures.

To express measures on a 100-pt scale, composites and single item measures are transformed from their original 3-pt or 4-pt scales using a simple linear transformation based on expressing the observed score as a percentage of the distance from the floor to the ceiling of a scale:

$$100 \times \left(observed\ score - scale\ floor\right) / range$$

For a 4-pt CAHPS scale (1=never, 2=sometimes, 3=usually, 4=always) with a mean of 3.5, the transformation would look like this, for example:

$$\frac{3.5-1}{3} \times 100 = 83.3$$

Dichotomous scales where 0=no and 1=yes are simply multiplied by 100 (e.g., if 72% of respondents answer ‘yes’ to the item, the transformed score is 72). Overall ratings, which range from 0 to 10, are simply multiplied by 10 (e.g., a mean of 9.3 becomes 93).

As an example of the proposed approach, consider a sample size estimation assuming a goal of having a half-width probability (power) of 0.80, an alpha of 0.05, and a half-width (margin of error) no greater than 3 points. With these parameters, the power analysis is estimating the number of completes needed to have an 80% chance of obtaining a 95% CI with +/- 3 point margin of error. To put this example in more concrete terms, with an observed score of 83.3 from a sample size calculated using the above inputs, there would be a 95% chance that the true score in the population would be between 80.3 and 86.3, and only a 5% chance that it would be outside of that range.

Exhibit B3 displays the number of completed surveys associated with some different combinations of half-widths (margins of error) and population variances (expressed as standard deviations). This exhibit illustrates the impact of sample size on precision and, thus, indicates the level of precision that might be obtained with the sample sizes proposed for the Marketplace beta test. Observed standard deviations from several of the CAHPS sources consulted ranged from approximately 2 to 28 points for measures on a 100-point scale. Observed standard errors ranged from around 0.30 to 3.2, which represent margins of error of approximately 0.60 to 6.3 points (on a 100-pt scale) for a 95% CI.

Exhibit B3. Precision Associated with Different Sample Sizes and Variances†

Margin of Error	Std=5*	Std=10	Std=15	Std=20	Std=25	Std=30
1	110	410	902	1,585	2,461	3,530
2	32	110	236	410	632	902
3	17	53	110	189	288	410
4	11	32	65	110	167	236
5	8	22	44	73	110	155

Margin of Error	Std=5*	Std=10	Std=15	Std=20	Std=25	Std=30
6	7	17	32	53	73	110

† Assumes an 80% half-width probability and a 95% confidence interval for a CAHPS measure scored on a 100-point scale.

* Std = standard deviation

As an illustration, assuming a standard deviation of 25 for an observed mean of 82, we would expect that, in a series of 100 independent random samples of at least 288 individuals (see blue highlighted cell in Exhibit B3) drawn from the same population, the true population score would fall between 79 and 85 (82 +/- 3) in 95 of those samples. For smaller variances, the precision gets better with smaller samples (e.g., with a sample size of 300 and a standard deviation of around 8 points, the margin of error would be +/- 1 point). For a sample size of at least 1,000, the margin of error would be no more than 2 points, assuming the standard deviation were no greater than 30. Simple t-tests of the difference in scores of handful of survey measures by several subgroup variables show that the standard deviations for the differences across subgroup categories ranges from around 20 to 35 (for measures scored on a 100 point scale), though very few are over 30.

Given the proposed 1,200 completed surveys per State, even if the population standard deviation was as high as 30, the margin of error for State-level estimates would be around +/- 2 (see the red shaded cell in Exhibit B3).

1.1.2.2.2 Assigning Star Ratings and Ranking Marketplaces and States

As described above, one of the objectives of the full national implementation of the survey is to assign star ratings to states based on their performance scores (on items, composites, and global ratings) relative to the average performance across all states. If a global F-test indicates that scores vary across states within the Federal Marketplace, the star rating is then done using a *t*-test of the difference between each state and the overall mean of all states. The discussion below shows that the utility of the scoring system depends on the number of completes. In Section 3.2, we discuss methods to evaluate the possible impact of the potential non-response bias.

Using variances observed from previous CAHPS psychometric tests, CMS conducted a power analysis based on a two-sample *t*-test comparing the mean score on a composite (on a 100-pt scale) from one entity to the pooled mean on that composite from all entities, using a range of variances. The power analysis assumes a balanced design (same number sampled from every entity) and equal variances (single entity variance = pooled variance).[14]

Exhibit B4. Relationship between Sample Size, Variance, and Effect Sizes for Star Rating of Marketplaces†

Number of Completes per State (Assuming 51 States)	Variance	of 15	Variance	of 25
	Mean Diff	ES	Mean Diff	ES
20	9.5	0.63	15.8	0.63

[14] In practice, this test is conducted using a Satterthwaite unpooled *t*-test on the mean difference, which accounts for unequal variances. We reproduced the analyses presented in Exhibit B3 using this test and specifying different variances for the single entity variance and the pooled variance. When the single entity variance is smaller than the pooled variance, the sample size required to detect mean differences of a particular magnitude tends to decrease. When the single entity variance is larger than the pooled variance, the sample size required tends to increase. However, the sample size requirements are still overwhelmingly determined by upper limit of either variance, regardless of how unequal they are. The impact on the estimated number of completes associated with the mean differences and variances presented in the exhibit was negligible.

Number of Completes per State	Variance	of 15	Variance	of 25
(Assuming 51 States)	Mean Diff	ES	Mean Diff	ES
50	6.0	0.40	10.0	0.40
100	4.2	0.28	7.1	0.28
150	3.5	0.23	5.8	0.23
200	3.0	0.20	5.0	0.20
300	2.5	0.16	4.1	0.16
500	1.9	0.13	3.2	0.13
1,200	1.2	0.08	2.1	0.08

† Assumes a balanced design (same number sampled from every entity) and equal variances (single entity variance = pooled variance). ES = effect size; Mean Diff = difference in means between a single State and the mean of all States

Exhibit B4 shows mean differences between a single State and the mean of all States that could be detected with a range of completed survey counts per State, given variances (the Root Mean Square Error) of 15 and 25.[15] Note that when the variance is larger, the mean differences have to be bigger to yield effect sizes of the same magnitude.

As shown, with 300 completes per State-specific subgroup and a variance of 15 points, we would have 80% power (with an alpha of 0.05) to detect a difference of 2.5 points between a single exchange and the overall mean of exchange scores (e.g., 87.5 versus 90). With a wider variance of 25 points, we could detect a difference of just over 4 points (e.g., 68 versus 72). The effect sizes associated with these differences (0.16) are relatively small, and thus a sample size of 300 per State-specific subgroup should be more than sufficient to detect any differences in performance large enough to be relevant. In fact, small effect sizes (0.28) could still be detected with as few as 100 completes per unit.

Moderate effect sizes could be detected with 50 completes per unit (a bit less than the approximate minimum number of completes we could expect in each state for race, ethnicity, or issuer subgroups comprising at least 5% of a Marketplace's population). With 1,200 completes per State, mean differences as small as 1.2 to 2.1 points could be detected, assuming variances of 15 or 25 respectively (effect sizes of 0.08, which are very small).

1.1.3 Marketplace Survey Sampling Methods

1.1.3.1 Psychometric Test

For the English surveys, CMS drew a stratified random sample from the sampling frame described above in Section 1.1.1.1; each of the 36 FFM states comprised a stratum. A total of 208 English-language consumers were drawn from each FFM state, for a total sample of 7,500. From this sample, equal numbers of individuals were randomly assigned to each of the five mode groups (1,500 each). We expected this strategy to produce 450 completed surveys in each of the five modes, yielding a total of 2,250 completed English-language surveys. The web-only group was randomly distributed such that half of the sample of 1,500 (n=750) received *both* an email and a pre-notification letter while the other half (n=750) received only an email; we expected that this strategy would produce 225 completed surveys in each of the web-only groups. See

[15] Results used for input to this power analysis were derived from a series of one-way analyses of variance (ANOVA) of CAHPS data using the entity as a single predictor and composite scores as outcomes. The square root of the mean square error (Root MSE) represents the total unexplained, or residual (within-entity), variance after removing the portion of variance accounted for by the entities (the explained, or between-entity, variance) from the total variance. See pp. 63-65 of the document "Instructions for Analyzing Data from CAHPS Surveys (Document No. 2015)" available here: Instructions for Analyzing Data from CAHPS Surveys (Document No. 2015), for a discussion of star ratings and examples of different effect sizes obtained with different sample sizes.

Exhibit B1 for details of the sample distributions. We expected this sampling approach to yield approximately 62 completed surveys in each of the 36 FFM states.

For the Spanish and Chinese samples, CMS used a systematic random sampling design to yield a sample proportional to the relative size of each group in the 36 States that are part of the FFM. In this design the sampling ratio (k) for each of two sample draws (one for Spanish and one for Chinese) was equal to N/1,500, where N is the number of eligible individuals in the FFM portion of the sampling frame who indicated their respective language preference in their Marketplace applications, summed across all 36 FFM States. We then sorted each sampling frame (one for each language) by State and a random number; then, using a random starting point, we drew a systematic random sample (with implicit stratification by State) by selecting every k*th* unit from the frame, yielding a total sample size of 1,500 for each of the two language groups. As described in Section 1.1.2, we expected the lack of a phone option for non-English speakers might negatively impact the response rates from these two populations. While the ideal would be for these two samples to yield 450 completed surveys each in Chinese and Spanish, CMS expected that the actual number of completes might be lower. To address this, CMS tested whether response rates varied significantly by race, ethnicity, and language among consumers for whom these variables were available in the sampling frame.

1.1.3.2 Beta Test Component

For the English and Spanish surveys, we will draw a stratified random sample from each state in the sampling frame described above in Section 1.1.1.1; each of the 44 participating states will comprise its own stratum. Samples averaging around 3,800 will be drawn from each strata to yield 1,200 completed surveys from each of the 44 States. Because our observed response rates (which are used in estimating sample sizes needed) vary by language, and the proportion of consumers with a Spanish or Chinese language preference varies by state, sample sizes will need to vary by state in the actual beta test. Rather than trying to make those estimates separately for each state now, Exhibit B5 illustrates this effect for states representing a range of language distributions, as well as for two scenarios associated with estimated maximums for both Spanish and Chinese. As can be seen, because of their lower response rate, as the share of consumers with a Spanish language preference increases, the sample size also must increase. Once the beta test sampling frame construction is complete, CMS will calculate the specific sample sizes needed for each state following the logic shown in Exhibit B5.

Exhibit B5. Sample Sizes by State based on Language Variation for Beta Test Component

State	ME	FL	LA	PA	Max Spanish	Max Chinese
English Share (RR = 32%)	100.0%	86.0%	99.5%	98.0%	64.5%	85.0%
Spanish Share (RR = 25%)	0.0%	14.0%	0.4%	1.5%	35.0%	12.0%
Chinese Share (RR = 36%)	0.0%	0.1%	0.1%	0.5%	0.5%	3.0%
Number sampled	3,750	3,900	3,754	3,764	4,115	3,864
Number completes	1,200	1,200	1,200	1,200	1,200	1,200

Consumers who have indicated a Chinese language preference will be oversampled. Oversampling will be conducted using a strategy similar to what was used for the psychometric test:

1. CMS will create a separate sampling frame that includes only consumers who indicate a Chinese language preference; this population will thus comprise its own separate stratum.
2. CMS will sort the Chinese frame by state and a random number and draw a systematic random sample of 2,200 consumers from the frame. With this implicit stratification by state, the size of the sample drawn from each state will be proportional to the population share of consumers with a Chinese language preference in that state.
3. Assuming a 36% response rate, a sample of 2,200 will yield 792 completed surveys in Chinese.
4. The English sample will be reduced by 1,907 (2,200 – 293) to compensate for the larger Chinese sample size, and keep the target number of completed survey constant.

This approach will yield a total sample of around 167,200 individuals, and should result in 52,800 completed surveys. For the beta test, individuals who specified a written or spoken language preference of Spanish or Chinese will be sent the survey in the requested language. All other individuals, including cases that do not include a language preference, will receive the survey in English that includes tag lines in Spanish and Chinese noting that the individual can request the survey in these languages. Individuals who specified a language preference other than English, Spanish, or Chinese will not be included in the frame and thus cannot be sampled. It is estimated that this approach will result in just over 2,640 (5%) surveys completed in Spanish and approximately 792 (1.5%) surveys completed in Chinese.

1.2 QHP Enrollee Survey

1.2.1 Respondent Universe, Study Population, and Sampling Frame

The respondent universe is the theoretical population to which we want our findings to apply. The study population is the population to which we can gain access, and the sampling frame is the means by which we can access this study population.

1.2.1.1 Psychometric Test

The QHP Enrollee Survey psychometric test sampling took place in August 2014, and data collection was carried out from the very end of August through the very beginning of December 2014. We defined the respondent universe for the psychometric test of the QHP survey as any adult (age 18+) enrolled in a QHP through the FFM.[16] We defined the study population as all individuals 18 years or older whose coverage started no later than April 1, 2014, and were enrolled in a QHP for at least 5 consecutive months. This requirement is a departure from the standard CAHPS approach, which includes enrollees who have been enrolled for 6 months or longer with no more than one 30-day break in enrollment during the 6 months. Because we used survey screener questions to identify eligible QHP enrollees, the enrollment time requirement had to be simplified to work within the limitations of survey question wording and be understandable to respondents. Anyone with coverage beginning later than April 1, 2014, would not have been enrolled long enough by the time enrollees began responding to the surveys in September of 2014. The psychometric test sampling frame was list-based and constructed from records contained in CMS databases.

There is some potential for bias in the QHP psychometric test due to website issues and enrollment problems in the first two months of open enrollment. It is partially mitigated by extending the eligibility period to include enrollees whose coverage begins as late as April 1, 2014. This approach will include those who submitted a marketplace application anytime between October 1, 2013 and March 30, 2014. Survey eligibility will be verified using screening questions on the QHP Enrollee Survey.

This limitation could only be mitigated further by relaxing the five-month enrollment requirement for eligibility. However, the consequence of relaxing that requirement is that fewer enrollees will have had any experiences with their plans and providers, which would make them screen out of many of the substantive survey questions. CMS and its contractor will include the month of enrollment in analysis models to test if there are differences in patterns of responses and measurement properties over time.

[16] Note: the definition of a Qualified Health Plan includes any health plan offered outside the Exchange by an issuer that is the same as a plan offered inside the Exchange. To be the "same plan" means that the health plan offered outside the Exchange has identical benefits, premium, cost-sharing structure, provider network, and service area as the QHP offered inside the Exchange. This reflects the fact that some issuers are enrolling persons in the same plan outside the Marketplace insfrastructure as well as through the Marketplace. These will mainly be persons who know that their income exceeds the maximum that would qualify for the Advance Payment Tax Credit wihtout going through the Marketplace and, thus, enroll directly with the issuer. They constitute part of the population enrolled in the QHP, because the plan is identical. In order to represent the entire population of the QHP, they will be eligible to be sampled.

The psychometric test component of this information collection is not designed to provide state-level or QHP-level estimates; as such only the aggregate results of this analysis will be discussed or disseminated.

1.2.1.2 Beta Test Component

The respondent universe for the beta test of the QHP survey is defined as any adult (age 18+) enrolled in a QHP through both the Federal and State-based maketplaces. The study population includes all individuals 18 years or older who have been enrolled in a QHP for 6 months or longer, with no more than one 30-day break in enrollment during the 6 months. The beta test sampling frames will be constructed by insurance issuers following instructions provided by CMS; the issuers will draw the samples. Sampling will be validated by a CMS contractor (Booz Allen Hamilton). This second component of the information collection (in 2015) includes a provision for initial feedback to QHPs on the operations of the data submission, collection and scoring processes, subject to the limitations associated with response rates and data quality findings, including non-response bias anlayses.

1.2.2 Sample Size Calculations

Sample size is calculated first by determining the minimum number of completed responses needed to meet the goals of the data collection, and second by inflating that number by a large enough factor to account for the estimated rate of survey non-response. Our assumptions for and approach to calculating response rates is described above in Section 1.1.1, and apply here. Response rate targets and the response rate calculation for the psychometric test of the QHP Enrollee Survey are the same as those for the psychometric test of the Marketplace survey. CMS assumes a 30% response rate.

1.2.2.1 QHP Psychometric Test Sample Size Estimates

Our sample size estimates for the QHP Enrollee Survey psychometric test reflect the sample sizes necessary for fully evaluating reliability and validity of the instrument. The reliability and validity testing for the QHP psychometric test will include the same analyses being conducted for the Marketplace Survey psychometric test (see Section 1.1.2.1 above).

As with the Marketplace survey, our sample size recommendations are based on our estimate of the minimum number of responses per equivalence group (i.e., mode and language groups) needed at the national level to conduct the psychometric analyses described in Section 1.1.2.1. This estimate is described in more detail immediately below. In order to evaluate unit-level reliability, which requires that we have a consistent number of completed surveys from each reporting unit (RU), we propose distributing the total national sample evenly across a purposively selected group of 30 RUs. This strategy, including the precise definition of the RU, is described in more detail in Section 1.2.3.1, which describes our proposed sampling methods for the QHP Enrollee survey psychometric test.

Sample Size

Factor analyses, multi-trait analyses, and the estimates of equivalence and internal consistency reliability will all be conducted separately for each survey administration mode using all complete responses from eligible sample members across the whole nation. The generalizability of the results from this psychometric analysis is obtained by attempting to capture the full range of experiences, and thus potential response patterns, in the population. As discussed in Section 1.1.2.1, standard psychometric practice is to obtain a *minimum* of 10 *complete responses* for each item that will be used in the psychometric analysis (this includes substantive questions that will be combined into composites, but not screeners, 'About You' items, or questions designed to determine survey eligibility).

At this time, the QHP Enrollee survey includes 40 assessment items, which translates into a *minimum* of 400 completed surveys nationwide, 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 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, we will propose to obtain a minimum of 15 complete responses for each assessment item. This translates into a minimum number of completes of 600 (15*40) for any grouping on which psychometric analyses will be conducted. For surveys conducted in English, there are five mode experiment groups (telephone-only, mail with telephone follow-up, mail-only with 1st class mail follow-up, mail-only with FedEx follow-up mailing, and web-only), and thus we need a minimum of 3,000 completed surveys to conduct psychometric testing for each mode (5*600 = 3,000). In addition, we would want 600 completed surveys each for both the Spanish and Chinese surveys to conduct psychometric analyses separately for each language.

To be useful for making sample size recommendations for future rounds of data collection, past experience demonstrates that, where the universe of accountable units is theoretically not finite (as with QHPs), it is best to have data from at least 30 accountable units selected across the full range of unit performance (i.e., from the poorest performing units to the best performing units). The CAHPS consortium recommends a minimum of 100 completed surveys per plan for the various Health Plan surveys, which should be sufficient for producing stable IUR estimates. With 30 QHPs, this translates into the requirement for a total of 3,000 completed surveys.

Taking into consideration the analysis requirements, a sample size sufficient to adequately conduct the psychometric analyses (3,000 completed surveys) will also be sufficient to evaluate unit-level reliability. Thus, CMS will sample equally across all 30 RUs with the goal of obtaining 100 completed surveys from each RU, for a total of 3,000 completed surveys.

Sampled consumers from each RU will be randomly assigned to each of the five mode groups, and we would control for mode in the IUR analysis to avoid confounding mode differences with differences across RUs. CMS will distribute the survey in Spanish and Chinese following the methods described for the Marketplace Survey psychometric test. Surveys in those languages will only be administered in the mail-only mode .

In our sampling frame data, we know who "enrolled" in a QHP (i.e., put a QHP in their shopping cart), but we do not know who among these enrollees: a) paid their first premium, and b) had coverage that remained in effect for at least 5 consecutive months from April through August of 2014. In other words, we do not know who among the enrollees in our QHP psychometric test sampling frame is eligible to complete the survey. The proposed solution is to inflate the initial sample to account for the fact that some percentage of enrollees will not meet our eligibility requirements. Assuming that 80% of enrollees paid their first premium in time for coverage to be *in effect* no later than April 1, 2014 (such an enrollee would at least have had the opportunity to be covered for 5 months prior to the beginning of survey data collection), then assuming that 75% of those enrollees continued to have their coverage remain in effect for the requisite number of months, our eligibility rate is 60% (0.80 x 0.75 = 0.60). Thus, we will need to inflate our initial sample size by 67% (inflation factor = 1/0.6 = 1.67). When the eligibility rate (60%) is combined with the assumed response rate (30%), we get an overall yield rate of 18% (0.60 x 0.30 = 0.18). Exhibit B6 displays the sample size requirements for the QHP Enrollee survey psychometric test required with an assumed yield rate of 18% (e.g., 3,340 x 0.18 = 601).

Exhibit B6. Sample Sizes and Completed Survey Counts for the QHP Psychometric Test

	Target Number of Completed Surveys	Total Number to Sample
English Language		
Exp 1. Phone only	600	3,340
Exp 2. Mail with phone	600	3,340
Exp 3. Mail only with third survey mailed Fed Ex	600	3,340
Exp 4. Web only	600	3,340
Exp 5. Mail only	600	3,340
Total English	**3,000**	**16,700**
Non-English		
Spanish (mail only)	600	3,340
Chinese (mail only)	600	3,340
Total non-English	**1,200**	**6,680**
Overall Total	**4,200**	**23,380**

1.2.2.2 QHP Beta Test Sample Size Estimates

The estimation of sample size for the beta test of the QHP survey will be driven by sample size estimates that result from the IUR analysis described above, as well as the analysis and reporting goals associated with this round of data collection (see Exhibit A1 in Part A). Once the analysis of the psychometric test data are complete, CMS will make final recommendations for sample size requirements to issuers and survey vendors.

As described in Section 1.1.2.2, to have a sufficient number of responses for *analysis* and *reporting* based on surveys of enrollees in health plans, CAHPS generally recommends obtaining completed questionnaires from 300 respondents per reporting unit.[17] With an assumed response rate of 30%, issuers would have to draw samples of 1,000 enrollees from each RU; however this number may be updated based on the observed response rates from the psychometric test of the Marketplace survey.

1.2.3 QHP Enrollee Survey Sampling Methods

1.2.3.1 Psychometric Test

While we use the term QHP as a semantic convenience, the operational definition of QHP for use in sampling (and ultimately reporting as well) is not as straightforward as common usage would suggest.

[17] See p. 65 in the document "Instructions for Analyzing Data from CAHPS® Surveys: Using the CAHPS Analysis Program Version 4.1," Document No. 2015, updated on 04/02/2012; available here: Instructions for Analyzing Data from CAHPS® Surveys: Using the CAHPS Analysis Program Version 4.1 .

If QHP is defined in terms of the unique Standard Component ID (SCID) provided by the HIOS system at the request of insurance issuers, then early data indicate that just over 200 issuers offer over 4,400 separate QHPs in just the FFM. Comments received from Blue Cross Blue Shield Association (BCBSA) in response to the 60-day FRN posting explained that using the SCID to define the sampling, data collection, and reporting unit would expose issuers to excessive burden by possibly requiring them to conduct dozens of separate surveys in a given state from individuals enrolled in products that are virtually identical (at least in terms of actuarial value). For example, one issuer in Arizona has 84 separate HMO plans across all five metal levels (including 30 silver plans, 22 gold plans, and 22 platinum plans), each with its own SCID; another issuer in Indiana has 137 HMO plans across the five metal levels, including 58 silver plans. BCBSA also described possible scenarios where a given issuer with a large number of plans (as defined by SCID) might have enrollments in each product offering that are small enough ($n < 500$) to result in a situation where that issuer would not be required to conduct any surveys at all beginning with the beta test and beyond.

Given these issues, it is apparent to CMS that the sampling and data collection unit for the QHP survey will have to be defined in terms of an aggregation of individual product offerings as defined by the SCID. Aggregating SCIDs up to the product *type* (EPO, PPO, POS, HMO) within issuer within state is a strategy that produces 268 unique units (this excludes child-only and dental-only plans). The current plan for the psychometric and beta tests is to aggregate to the level of product type (HMO, PPO, EPO, POS) offered by a given issuer in a given state (e.g., an Aetna HMO in Texas, or Coventry PPO in Florida). We refer to this unit as a State-Issuer-Product type (SIP). This approach yields 226 units from which to sample our 30 “QHPs” for the psychometric test.

In order to be eligible to be in this sampling frame, a SIP will have to have a enough enrollees to produce the required number of *completed surveys* (n=100). CMS has determined that, in order to obtain 100 completed surveys per unit, the SIP frame would need to exclude any SIPs with less than 660 de-duplicated, survey-eligible enrollees. This estimate was based on assuming a 30% response rate, an effectuation rate of 67%, and assuming that 75% of effectuated enrollees would be enrolled for the minimum of 5 months.[18] This results in a yield rate of 15% (0.30 x 0.67 x 0.75 = 0.15), which corresponds to an inflation rate of 6.67 (1.0/0.15). Multiplying 100 by the inflation rate of 6.67 indicates that we will need 667 survey-eligible enrollees per RU. We rounded this estimate down to 660 to account for imprecision in the various rates described above. A total of 226 SIPs meet this enrollment size requirement.

Sampling for the QHP psychometric test took place in two stages. First, we drew a sample of 30 SIPs from the sample frame of all 226 SIPs.CMS selected 30 SIPs using a mix of random and purposive sampling techniques. CMS considered purposively selecting the 30 SIPs from that frame on the basis of several criteria, such as maximizing geographic variation, including plans for specific states that are likely to span the full range of likely enrollee experience, including plans that vary in the racial and ethnic composition of their enrollee populations, or ensuring that specific states are represented. However, after reviewing the results of selecting the 30 SIPs at random, it became obvious that random selection produced a sample of SIPs that very closely

[18] Note that we used a more conservative effectuation rate assumption at this stage of sampling. CMS assumed a conservative 67% based on the best information available at the time. Once the 30 SIPs had been finalized, we did not want to go back and resample based on including an additional 5 RUs that would have met the lower requirement of 600 survey-eligible enrollees that results from applying an 80% effectuation rate assumption.

reflected the distribution of enrollee characteristics in the frame. The only drawback to that approach was that it would be virtually impossible for any set of 30 randomly selected SIPs to provide the 3,340 individuals needed for the Chinese sample.

To address this problem, we tested a blended approach that included purposively selecting the two SIPs with the most Chinese-language individuals and randomly selecting the remaining 28 SIPs. This approach successfully produced a 30-SIP sample with a sufficient number of both Spanish-language and Chinese-language individuals from which to draw our person-level sample, and still produces a sample of SIPs that very closely reflects the distribution of enrollee characteristics in the frame.

Next, CMS drew a simple random sample of 557 enrollees from each of the 30 SIPs sampled at the first stage, producing a total sample of 23,380 enrollees across the three languages (16,700 English; 3,340 each for Spanish and Chinese). The English enrollees were randomly assigned to each of the five experimental mode groups: 3,340 to each group (see Exhibit B5). For the Spanish and Chinese samples, CMS used a systematic random sampling design to yield a sample proportional to the relative size of each group in the 30 SIPs sampled at the first stage. In this design the sampling ratio (k) for each of two sample draws (one for Spanish and one for Chinese) will be equal to N/3,340, where N is the number of eligible individuals in the sampling frame who have indicated their respective language preference in their Marketplace applications, summed across all 30 SIPs. We then sorted each sampling frame (one for each language) by State and, using a random starting point, draw a systematic random sample (with implicit stratification by SIP) by selecting every kth unit from the frame.

As described in Section 1.1.2, the lack of a phone option for non-English speakers may negatively impact the response rates from these two populations. While the ideal is for these two samples to yield 600 completed surveys each for those consumers whose preferred language is either Chinese or Spanish, CMS is aware that the actual number of completes may be lower. To address this, CMS will test if response rates vary significantly by race, ethnicity, and language among consumers for whom these variables are available in the sampling frame.

1.2.3.2 Beta Test Phase

For the beta test, HHS-approved QHP Enrollee Survey vendors will draw samples from each SIP using instructions and guidelines provided by CMS. For the 2015 Beta test, the sampling and reporting unit has been defined at the level of product type (i.e., EPO, HMO, PPO, POS) offered by an insurance carrier in a particular state. For example, XYZ Insurance Company’s HMOs in Florida would be considered a single sampling unit. Note that, depending on the way a QHP issuer packages its plan offerings, the sampling unit might include anywhere from a single QHP to dozens of QHPs spanning all categories of coverage (i.e., bronze, silver, gold, platinum, and catastrophic). QHP issuers will create a sample frame for each product type they offer through the Marketplace within a particular state. For convenience, we refer to this unit as a State-Issuer-Product-type, or SIP sampling unit.

CMS will explore data collection at a more granular level of QHP issuer coverage (e.g., HMO Bronze level) in the future, keeping in mind the need to balance the value of this information for consumers with QHP issuer data collection, validation, and reporting efforts.

The list of all eligible individuals who are enrolled in eligible QHPs subsumed by a SIP sampling unit will constitute the sample frame. The guidelines for determining who to include in the sample frame for a given SIP sampling unit follow:

- Include all individuals who are enrolled in an eligible QHP offered through any of the Marketplace types (i.e., Federally-Facilitated Marketplace [FFM], State Partnership Marketplace [SPM], or State-based Marketplace [SBM]). An eligible QHP includes:
    - QHP offered through the individual Marketplace or SHOP Marketplace for at least one coverage year;
    - QHP that offers family or adult coverage;
    - QHP associated with one of the following categories of coverage: bronze, silver, gold, platinum, and catastrophic;
    - QHP associated with one of the following product types: EPO, HMO, POS, PPO;
    - QHP associated with a SIP that has more than 500 enrollees as of July 1, 2014. (Note: Enrollees in otherwise eligible QHPs that are suppressed by CMS are included in determining the number of enrollees in the SIP and are included in the sampling frame if they meet the other eligibility criteria below).

  The following QHPs or plan types are <u>not</u> included:

    - Stand-alone dental plans.
    - Child-only plans.
- Include individuals 18 years or older who have been enrolled in an eligible QHP for at least the last 6 months, with no more than one 30-day break in enrollment during the 6 months.
    - Use December 31, 2014, as the anchor date to determine whether the individual meets the age and 6-month enrollment requirements. For example, include all individuals enrolled in an eligible QHP who are 18 years or older as of December 31, 2014, and who have been enrolled since July 1, 2014 (still allowing for the single break of no more than 30 days).
- Include only individuals with primary health coverage through the eligible QHP in which they are enrolled.
- Include individuals enrolled in an eligible QHP and who meet other eligibility criteria, regardless of the route through which they enrolled in the eligible QHP (e.g., whether they enrolled through the individual Marketplace, the SHOP Marketplace, or directly with the QHP issuer).

- Allow the sample frame to include multiple adults (age 18 and older) from the same policy. The survey vendor will implement sampling procedures to ensure that the sample will not have more than one adult per policy.

- In the event that an identical plan is offered on and off of the Marketplace, include enrollees in the off-Marketplace version of the plan only if the off-Marketplace version of the plan is certified as a QHP by the state insurance commissioner and has the same plan ID (standard component ID) as the version offered on the Marketplace.

An auditor will audit the sample frame created by the issuer to verify that it follows the specifications and includes all eligible individuals. The issuer will then provide this sample frame to the survey vendor with whom they have contracted, who will draw a sample of individuals from the sample frame and conduct the QHP Enrollee survey. By following these guidelines, issuers can be confident that their results will be comparable to those produced by other issuers and vendors.

2. Information Collection Procedures

Both surveys will follow standard CAHPS procedures with respect to defining the sampling frame and determining respondent eligibility, and survey operations.[19]

For the Marketplace surveys and the QHP psychometric test in 2014, data will be collected by a single survey vendor; for the QHP beta test (and full implementation surveys), data will be collected by multiple approved commercial vendors on behalf of QHP issuers. The mode of administration will be mail with phone follow-up. Survey operations for both surveys will follow standard CAHPS practice:

- Mail an advance letter
- Mail the questionnaire package one week after the advance letter. Include a postage-paid envelope to encourage participation.
- Send a postcard reminder to nonrespondents 10 days after sending the questionnaire.
- Send a second questionnaire with a reminder letter to those still not responding thirty days after the first mailing.
- Begin follow-up by telephone or send final mail survey with nonrespondents three weeks after sending the second questionnaire. Interviewers will attempt to locate respondents who have not responded to the mailed survey
- Telephone numbers for sample respondents will be verified prior to calling
- A maximum of 9 attempts will be made by phone

[19] As described in Document No. 13b in the CAHPS Health Plan Reporting Kit, which is titled “Fielding the CAHPS Health Plan Survey: Commercial Version.”

3. Methods to Maximize Response Rates and Address Non-Response Bias

3.1 Maximizing Response Rates

Every effort will be made to maximize the response rate, while retaining the voluntary nature of the effort. Below are several options recommended by CAHPS for maximizing response rates that may be employed:

- We will set up a toll-free number and publish it in all correspondence with respondents. Assign a trained project staff member to respond to questions on that line. Maintain a log of these calls and review them periodically.
- For the *psychometric tests* of both the Marketplace and QHP Enrollee surveys, a persuasive advance letter will be sent to the respondent. Cover letters describing the survey and encouraging participation will also be included in the survey packets. Reminder postcards will also be sent to encourage participation. The letters will be printed on CMS letterhead with an official logo and include an official signature of a representative from CMS; it will be personalized with the name and address of the intended recipient. Postcards will include an official signature of a representative from CMS.
- In subsequent data collections using the *Marketplace* survey (beta test and national implementations in 2016 and 2017), where samples will be pulled from CMS administrative files, advance letters and cover letters will be sent on CMS letterhead and signed by the CMS privacy officer the same as in the psychometric test.
- For subsequent data collections for the *QHP* survey, both advance letters and cover letters will use the letterhead and logo of the survey vendor or, alternatively, the letterhead and logo from the QHP issuer.
- The envelope will also include the appropriate official logo and include a return address; envelopes will be marked “forwarding and address correction” in order to update records for respondents who have moved and to increase the likelihood that the survey packet will reach the intended respondent.
- For the telephone interviews:
  - Interviewers will be trained and monitored
  - Interviewers will read questions exactly as worded so that all respondents are answering the same question.
  - When a respondent fails to give a complete or adequate answer, interviewer probes will be nondirective.
  - Interviewers will maintain a neutral and professional relationship with respondents. The primary goal of the interaction from the respondent’s point of view should be to provide accurate information. The less interviewers communicate about their personal characteristics and, in particular, their personal preferences, the more standardized the interview experience becomes across all interviewers.
  - Interviewers will record only answers that the respondents themselves choose. The instrument is designed to minimize decisions that interviewers might need to make about how to categorize answers.

– The single vendor for the Marketplace surveys and the multiple vendors for the QHP Enrollee Surveys will be required to use CATI.

The mode-of-administration experiment is being conducted in the psychometric test to determine the most efficient and least burdensome modes that should be used in the subsequent surveys.

Unduplicating the samples for the Marketplace and QHP surveys is another way to improve response by minimizing burden on specific sample members who might be selected for both surveys. The psychometric test samples for both the Marketplace and QHP surveys are being drawn by CMS and its contractor, so the two samples will be unduplicated. For the beta test, the sample for the Marketplace Survey will be drawn by CMS and its contractor, but the samples for the QHP Enrollee Survey will be drawn by commercial survey vendors hired by the QHP issuers. The data will be supplied to CMS and its contractor for analysis without identifiers, so it will be impossible to unduplicate the Marketplace Survey and QHP Enrollee Survey samples beginning with the beta test or to know the extent to which duplication occurred. CMS believes that the population for the QHP sample will eventually be so large that the chances of the same individual being selected for both the QHP and Marketplace Surveys will be small, but we will not be able to estimate the likelihood of duplicate selection until the sampling frames for the beta test and subsequent annual rounds of the surveys are constructed.

As part of testing the performance of the surveys in the psychometric test, CMS will determine if the goal of 30 percent response can be achieved. The actual response rates obtained in the psychometric test will be used to adjust the response rate goals for the beta test and subsequent rounds. If 30 percent is not achieved in the psychometric test, the reliability of the surveys as determined at the national level and the ability to conduct subgroup analyses will depend on the presence of non-response bias.

3.2 Evaluating Non-Response BiasIf response rates are less than 80 percent, which we expect to be the case based on the results from other CAHPS surveys (we are targeting 30 percent), CMS will conduct nonresponse bias analyses to determine if there are systematic differences between respondents and nonrespondents in terms of demographic, Marketplace, or QHP related characteristics that could have an impact on the study outcomes. Some of the potentially related characteristics that will be available on the sampling frame for respondents and nonrespondents of both the Marketplace and QHP Enrollee surveys include: the mode of application (phone, web, in-person, or a combination), applicant status (PA, PE, EE, E), Medicaid eligibility, language preference, race, ethnicity, gender, income, disability status, and state. Additionally, CMS will know the QHP issuer, product type, and metal level for respondents and nonrespondents of the QHP Enrollee survey. Of particular interest is the extent to which response rates vary by language, state, or mode and the extent to which response rates within these groups differ by sociodemographic characteristics. For example, a nonresponse bias analysis could investigate whether the sociodemographic characteristics of the mail mode respondents and nonrespondents are systematically different. If bias is found CMS will employ post-stratification to lessen the effects of non-response bias. CMS will also consider the possibility of conducting non-English surveys by telephone if the results of these analyses suggest that there is a significant bias associated with limiting non-English surveys to mail only.

If response rates vary by mode in the psychometric test, CMS will compute a cost per complete for each mode and relate the response rate for that mode to its unit cost to determine if the benefit in terms of better response is worth any additional cost that might be required. This assessment will be made qualitatively once we see the variation in costs and response rates among the modes. There is no *a priori* assumption about an acceptable benefit-cost tradeoff; however, CMS also wants to remain consistent with standard CAHPS survey administration procedures to the extent possible.

Thus far, the response rates discussed have been at the unit level, where respondents either completed or did not complete the entire survey. There is also item-level nonresponse where a respondent answers some, but not all of the questions they are eligible for in the survey. Although highly unlikely, if the item response rate is less than 70% for any survey questions, CMS will conduct an item nonresponse analysis similar to that discussed above for unit nonresponse as required by Guideline 3.2.10 of the Office of Management and Budget's Standards and Guideline for Statistical Surveys.

4. Tests of Procedures

The survey development team conducted nine interviews with key stakeholders to help inform aspects of the Marketplaces that would be important to capture in the surveys; four focus groups with 33 individuals about their perspectives on health insurance, health care, and the new Health Insurance Marketplaces; and two rounds of cognitive testing in all three languages (English, Spanish, and Chinese) for both surveys. To avoid duplicating efforts we relied heavily on cognitive testing that had already been done on the CAHPS questions used in the QHP Enrollee Survey and only tested new or modified questions. Thus, cognitive testing focused mainly on the Marketplace Survey. The first round of testing was conducted with proxy Marketplace users from the Massachusetts Health Connector because it had to be done before Marketplace open enrollment began. The nine interviews in each language were sufficient to understand respondents' experiences with the Massachusetts Health Connector. The second round of testing was conducted in the first weeks of Marketplace open enrollment when people had varying experiences with the Marketplaces. The nine respondents in each language provided a balanced perspective of positive and negative experiences interacting with the Marketplace in a variety of ways such as on the website, over the phone, and in person. The final cognitive testing report was provided as part of this submission earlier. The CCSQ survey team worked closely with CCIIO's state-based marketplace team, who collected state level information about enrollees. CMS intends that the psychometric tests will verify and validate the cognitive testing and identify any additional testing needs.

The Marketplace and QHP psychometric and beta test surveys are intended to test and refine the questionnaires and survey procedures prior to the full national implementation of both surveys, with public reporting, which will take place annually beginning in 2016.

5. Statistical Consultants

This sampling and statistical plan was prepared and reviewed by staff of CMS and by the American Institutes for Research. The primary statistical design was provided by Chris Evensen, MS, of the American Institutes for Research at (919) 918-2310; Michael P. Cohen, PhD, of the American Institutes for Research at (202) 403-6453; Steven Garfinkel, PhD, of the American Institutes for Research at (919) 918-2306, and HarmoniJoie Noel, PhD, of the American Institutes for Research at (202) 403-5779.