Development of a Consumer Experience Survey for the Health Insurance Marketplace

Marketplace Survey Field Test Psychometric Analysis Report – DRAFT

Field Test Phase Deliverable 8.1a

January 9, 2015

Prepared for:
Centers for Medicare & Medicaid Services
Center for Clinical Standards and Quality
Contract No. GS–10F–0112J
Task Order No. HHSM–500–2012–00100G

Submitted to:
Kathleen Jack
Contracting Officer's Representative

Submitted by:

AIR
AMERICAN INSTITUTES FOR RESEARCH®

American Institutes for Research
1000 Thomas Jefferson Street NW
Washington, DC 20007-3835

Development of a Consumer Experience Survey for the Health Insurance Marketplace

Field Test Phase Deliverable 8.1a

Marketplace Survey Field Test Psychometric Analysis Report – DRAFT

January 9, 2015

Christian Evensen, M.S.
HarmoniJoie Noel, Ph.D.
Daniel Harwell, M.S.
Chris Pugliese, M.P.P.
Manshu Yang, Ph.D.
Emily Elstad, Ph.D.
Coretta Mallery, Ph.D.
Tandrea Hilliard, M.P.H.
Stacey Bielick, M.S.
Steve Garfinkel, Ph.D.

For:
Kathleen Jack
Contracting Officer's Representative
Center for Clinical Standards and Quality
Centers for Medicare & Medicaid Services

Contract No. GS-10F-0112J
Task Order No. HHSM-500-2012-00100G
AIR Project No. 01410.083.02

AIR®
1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000 | TTY 877.334.3499
www.air.org

.

Contents

Page

AIR
AMERICAN INSTITUTES FOR RESEARCH®

1.0 Introduction

1.1. Purpose of the Field Test

The 2014 field test of the Health Insurance Marketplace Survey (Marketplace Survey) was designed to enable the evaluation of the reliability and validity of the Marketplace Survey measures, the identification of potential case-mix adjusters, computation of scores, implementation of a key driver analysis, and analysis of differences in Marketplace experience by select subgroups. We employed psychometric analysis techniques to inform revisions to the Marketplace Survey aimed at maximizing the reliability and validity of the instrument. The main goals of the psychometric evaluation are to identify a subset of questions that would efficiently produce precise measurement of Marketplace performance and to identify the most efficient mode of administration for the surveys for the beta test and beyond. In addition, the field test was designed to allow the analysis team to evaluate the performance of the instrument in English, Spanish and Chinese. The analysis of key drivers will provide CMS with guidance for quality improvement (QI). The subgroup analysis (e.g., looking at differences in scores by race, ethnicity, income and disability) will provide CMS with preliminary data regarding disparities during the first open enrollment period and inform the estimation of sample size for the beta test.

The Marketplace Survey was an entirely new questionnaire. The survey development team designed the questionnaire using methods established as part of the Consumer Assessment of Healthcare Providers and Systems (CAHPS) work sponsored by the Agency for Healthcare Research and Quality (AHRQ).

1.2. Overview of the Marketplace Survey field test

Length: The Marketplace Survey included 96 items. AIR estimated that the survey would take about 24 minutes to complete.

Languages: The Marketplace instrument was field tested in three languages—English, Spanish, and Chinese (with traditional Chinese characters). Please refer to Appendix A for a copy of the Marketplace instrument in English.

Sample: The sampling design and methods are summarized in Section 2.2 below. A full account of the sampling design was provided in *Deliverable 7.1b Final Field Test Plan for the Marketplace and Qualified Health Plan Enrollee Experience Surveys*.

Modes: The Marketplace Survey was administered in three modes: mail, telephone, and online. For the English survey, we randomly assigned a predetermined number of subjects from the national sample to five experimental groups shown in Exhibit 1: (1) phone-only, (2) mail with phone follow-up, (3) mail-only with Fed Ex follow-up, (4a,b) Web-only, and (5) mail-only with first-class follow-up. In addition, there was a sub-experiment for the Web-only group where half got the mailed advance pre-notification letter with a URL for the survey (4a) and the other half got an emailed survey link (4b). All English survey respondents in the non-Web modes were also given the option to complete the survey online through a URL with username/password provided in the mailed advance pre-notification letter. Due to budget constraints, the Spanish and Chinese surveys were administered in one mode—mail-only with first-class follow-up.

Exhibit 1 describes the sample allocation among modes and languages for the Marketplace Survey field test.

Exhibit 1. Sample sizes and expected completed survey counts for the Marketplace Survey field test psychometric analyses

Mode†	Target Number of Completed Surveys	Total Number to Sample
English Language		
Exp 1. Phone-only	450	1,500
Exp 2. Mail with phone follow-up	450	1,500
Exp 3. Mail-only with FedEx follow-up	450	1,500
Exp 4a. Web-only with email and pre-notification letter	225	750
Exp 4b. Web-only with email only	225	750
Exp 5. Mail-only with first-class follow-up	450	1,500
Total English	**2,250**	**7,500**
Non-English		
Spanish (Mail-only)	450	1,500
Chinese (Mail-only)	450	1,500
Total Non-English	**900**	**3,000**
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.

Target response rate: The anticipated response rate was 30% on average among all modes. Because the anticipated response rate is below the OMB required 80%, we conducted a nonresponse bias analyses to determine if there were systematic differences between respondents and nonrespondents in terms of demographic or Marketplace related characteristics that could have an impact on the study outcomes.

Data collection timeframe: Sampling occurred in April 2014 and data collection occurred between May 13 and August 19, 2014. Post-data collection processing, including data cleaning, took three weeks and occurred between August 20 and September 12, 2014. Psychometric analyses were conducted after post-data collection processing from September 13 through November 30, 2014. Scoring and analysis was conducted after the psychometric analyses in December 2014 through January 2015. The Marketplace Survey will be revised in December 2014 and early January 2015 in time for inclusion in the materials submitted to OMB in advance of the Marketplace beta test that will begin in late February or Early March of 2015.

Exhibit 2 displays the timeline for the Marketplace Survey.

Exhibit 2. Data collection and analysis timeline for the Marketplace Survey

Marketplace Survey	2014												2015				
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5
Sampling				▓													
Data collection					▓	▓	▓	▓									
Post-data collection processing								▓	▓								
Psychometric analyses									▓	▓	▓						
Scoring and analysis											▓	▓					
Revise surveys											▓	▓					

2.0 Scoring and Analysis Overview

2.1 Goals

The scoring and analysis of the field test data for the Marketplace Survey was designed from the beginning of the project to follow standard CAHPS practice with respect to the initial testing of a new instrument. The methodology for conducting the scoring and analysis was described in detail in *Deliverable 3.1b: Final Scoring Methodology and Analytical Strategies Plan for the Marketplace and Qualified Health Plan Enrollee Surveys*.

The field test data will be used to address three analysis activities:

1. Evaluation of the psychometric properties of the instrument

2. Scoring the resulting final measures (to include single-item patient experience measures, composite measures, and global ratings measuring respondent evaluation of quality of services)

3. Conducting substantive analyses of the field test data

   a. Analysis of disparities

   b. Analysis of key drivers

This report will focus on the first of these analytic goals; the work and results related to the remaining goals will be described in a separate analysis and scoring report.

As part of the psychometric analysis, AIR conducted the following activities:

- Assessed the quality of survey responses (rates of item nonresponse and failed skips) – Data Quality Evaluation (DQE) section.

- Analyzed potential non-response bias – Response rate (RR) and non-response bias (NRB) section.

- Evaluated variation in respondent characteristics by different modes of survey administration – Mode Effects (ME) section.

AIR
AMERICAN INSTITUTES FOR RESEARCH®

- Examined and tested the properties of the proposed measures – Factor Analysis (FA) and Multi-Trait Analysis (MTA) sections.

- Conducted analyses to identify potential case mix adjusters – Case Mix Analysis (CMA) section.

- Made recommendations for revisions to the survey and measure specifications based on those results – Survey Revisions section.

2.2 Sampling and Summary of Returns

AIR created three separate sample frames based on the consumers' language preference. The MIDAS data from which the frame was constructed included both a written language preference (WLP) and a spoken language preference (SLP). Anyone who indicated either a WLP or SLP that was not English, Spanish, or Chinese was assigned to an "other" category.

Consumers were assigned to the English frame if they 1) expressed a WLP for English, 2) indicated "other" or none for their WLP, but indicated an SLP of English, or 3) indicated "other" or none for both SLP and WLP. Written language preference was given precedence over SLP since the majority of surveys were assigned to a written language mode (mail or web).

For the English-language sample, AIR selected a stratified random sample from the English sampling frame. This sample represents all 36 states that used healthcare.gov for their application and enrollment activities, with each of the 36 states comprising a stratum.[1] A total of 209 English-language consumers were selected from each state, for a total sample of 7,524.[2] From this sample, equal numbers of individuals were randomized to each of the five experimental modes; the web sample was further randomized to the two sub-groups. These distributions of sample are shown in Exhibit 3.

Consumers were assigned to the Spanish frame if they 1) expressed a written language preference (WLP) for Spanish, or 2) indicated "other" or none for their WLP, but indicated an SLP of Spanish. Consumers were assigned to the Chinese frame if they 1) expressed a written language preference (WLP) for Chinese, or 2) indicated "other" or none for their WLP, but indicated an SLP of Chinese. Written language preference was given precedence over SLP since all surveys in Spanish and Chinese were to be by mail.

For the Spanish and Chinese samples, AIR used a systematic random sampling approach to produce 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) is equal to N/1,500, where N is the number of eligible individuals in the sampling frame who have indicated their respective language preferences in their Marketplace

---

[1] Although all 36 states are technically not part of the Federally Facilitated Marketplace (FFM), we refer to this group of 36 states collectively as the 'FFM' for convenience.

[2] In practice, the total sample varies from the design due to rounding. A total of 209 were sampled from each state, which was derived from rounding up from the total distributed across 36 states (7,500/36 = 208.33). This produced a total English sample of 7,524, which results in a few extra sampled consumers who were then randomized as equally as possible across the five English modes. In addition, the Web sample was randomized again to the two Web groups.

AIR®
AMERICAN INSTITUTES FOR RESEARCH*

applications, summed across all 36 FFM states. We then sorted each sampling frame (one for each language) by state and a random number using a random starting point, AIR selected a systematic random sample (with implicit stratification by state) by selecting every kth unit from the frame, yielding a total sample size of 1,500 for each of the two language groups. These distributions are shown in Exhibit 3.

Exhibit 3. Sample size, respondents, and completions by assigned mode*

Assigned Mode	Sample Size	Number of Respondents	Raw RR	Number of Completes	Completion Rate
Mail-FedEx	1,504	403	27%	393	98%
Mail-Mail	1,505	301	20%	296	98%
Mail-Mail Chinese	1,500	540	36%	532	99%
Mail-Mail Spanish	1,500	381	25%	372	98%
Mail-Phone	1,505	473	31%	441	93%
Phone-Only	1,505	323	21%	259	80%
Web: Email Invite	753	54	7%	50	93%
Web: Postal Invite	752	79	11%	70	89%
Totals	10,524	2,554	27%	2,413	95%

*RR = response rate. The raw RR was equal to the number of respondents divided by the sample size; a more refined RR that was calculated according to AAPOR guidelines will be presented in Section 4. The completion rate was the number of completed surveys divided by the number of respondents. The total for both of these percentages was weighted (i.e., the sum of the weighted rates divided by the total sum of consumers for that rate).

A respondent was defined as any sampled consumer who answered at least one question on the survey (n = 2,554). A completed survey was defined as one where a respondent answered at least half of the survey items all respondents were eligible to answer (not including the "About You" items) – referred to as the set of "key" items (n = 2,413). The key items included: q01, q06, q16, q18, q26, q36, of q46-q48, q50, q52, q54, q56, q58, q60, q62, q66, and q67.

A more detailed analysis of response rates, potential non-response bias, and mode bias appear in Section 4.

3.0 Data Quality Evaluation

The goal of the data quality evaluation (DQ) is to identify any survey questions that may have been confusing or burdensome to respondents by flagging items with high rates of nonresponse and identifying areas of the survey where skip instructions were not correctly followed by the respondent. Inordinate rates of nonresponse to a question suggest that the question was poorly understood by respondents (who skipped the item because they were not sure how to reply), that the item did not apply to the respondent, or that the item asked for sensitive information that respondents may have been unwilling to give. Failure to follow skip instructions can suggest that those instructions were unclear or inappropriate (and thus the respondent ignored or chose not to follow them), or that the format of the survey made it difficult for respondents to understand the skip instructions (and thus they were unable to follow them).

The DQ evaluation was conducted at both the respondent level and the item level. The latter is used to identify problem *items*, whereas the former is designed to identify *respondents* who might have had problems completing the survey.

3.1 Methods

In the raw survey data file provided by Ipsos (the survey data collection vendor), every field associated with a question on the survey includes either a numeric value associated with a response or is blank. Some survey items were used as screeners designed to determine whether a consumer was eligible to respond to follow-up questions about specific experiences. Responses were either answers to questions that the respondent should have answered (referred to as 'legitimate responses,' which are good) or, following a screener, failed skips (bad). Blanks were either correct skips (good) when following a screener or they can be missing values (bad). Blank screeners created a particular type of uncertainty that affects items that follow the screeners. Each is described below in more detail.

The first step in this analysis was to initialize new variables corresponding to each survey question that contains *item disposition* codes in place of the actual responses. Some item disposition codes applied only to screener-item pairs that were part of skip patterns, and others applied to all survey items. One of five possible item disposition codes was assigned to every item within each respondent record. This coding was done for all respondents, regardless of whether or not the respondent met the criteria to for a complete survey described in Section 2.2.

The five item disposition codes include:

1. Correct Skips (CS)—This code was applied to a follow-up item where the respondent answered the screener question with a response that should have triggered a skip (i.e., a response that *should* result in skipping the next item or several items), and then did in fact follow the instructions and skipped the follow-up item(s). For each respondent, two CS rates were calculated. The first was calculated relative to the total number of responses in the survey. This rate was not a direct indicator of the quality of response; rather, it was added to the legitimate response (LR) rate to obtain the total percentage of appropriate responses given by each respondent to the survey. We also calculated a true CS rate (TCS), which was the total number of correct skips divided by the sum of the number of correct skips and failed skips. This denominator was equal to the total number of items respondents *should* have skipped based on their response to the screener question. A higher TCS rate indicates a higher quality in responses—that is, fewer items where the respondent failed to follow skip instructions. These codes were only assigned to items controlled by screeners; they weare not assigned to screeners unless they were part of a nested skip.

2. Failed Skips (FS)—This code was applied to a follow-up item where the respondent answered the screener question such that the next item or several items should have been skipped and then *failed* to follow the instructions and gave a response to the follow-up item(s) anyway. Similar to the CS rate, two FS rates were calculated. The first was calculated relative to the total number of items in the survey. This rate is not a direct indicator of the quality of response, and thus we also calculated a true FS rate (TFS), which was the total number of failed skips divided by the same denominator used for the TCS rate. A higher TFS rate indicates a lower quality of responses. Like the CS and TCS, these codes were only assigned to items controlled by screeners; they were not assigned to screeners unless the screeners were part of a nested skip. The TFS and the TCS sum to 100 percent.

3. Indeterminate Eligibility (IE)—This code was applied to a follow-up item for respondents who left the associated screener blank (or did not answer a screener, in the case of a phone survey), even if the follow-up items contains a valid response. This code thus indicated that the respondents' eligibility to answer the follow-up questions associated with the screener could not be determined. As with CS and FS, this code was only assigned to items controlled by screeners; it was not assigned to screeners unless they were part of a nested skip.

4. Truly Missing (TM)—This code was applied to all blank or unanswered survey items that did not qualify as either a CS or IE. Note that the rate of TM for a screener item matched exactly the rate of IE for all follow-up items linked to that screener.

5. Legitimate Response (LR)—This code was applied to any non-missing response that was not coded as FS or IE.

There were 168 total items in the Marketplace Survey. This total is based on counting every response option for a code-all-that-apply item (e.g., q02 or q07) in the survey as a separate item. Any respondent who provided a legitimate response to *any* of the response options for such items was considered to have provided a legitimate response to *all* of the response options.[3] These item disposition codes were summed across all of the items in the survey for each respondent; thus, each respondent record contained seven additional variables indicating the total number of each of the seven item disposition codes assigned (CS, TCS, FS, TFS, IE, TM, LR). These rates were calculated for each respondent by dividing each of these count variables by the total number of items on the survey.

In addition we calculated an item response rate (IRR) for each respondent, which was equal to the total number of LRs provided by the respondent, divided by the total number of survey items minus the total number of CS and FS combined. For each respondent, the following formula was used to calculate the IRR:

Item RR = (LR) /(168−(CS+FS))

The item nonresponse rate (INRR) was equal to 1−IRR. The INRR was the percentage of items the respondent was eligible to answer but did not. Contrast this with the TM *rate*, which simply included the total number of survey items in the denominator (n=168) regardless of the respondent's eligibility to answer those items. While each of these codes provides information that is useful to assess across items and respondents, some are more easily interpreted for the evaluation of data quality when combined. Measures were thus grouped into positive and negative indicators:

- Positive indicators included dispositions that indicated desirable responses: the IRR and the sum of LR and CS.

- Negative indicators included dispositions that indicated problematic responses: the INRR and the sum of FS, CS, and IE.

---

[3] For the code-all-that-apply items, there was no opportunity give for the respondent to indicate 'no' or 'not applicable' to the various response options. Thus, we relied on the screener to determine who was eligible to respond to any of the options presented, and viewed a response to any item as a response to all of the items.

3.1.1 Respondent Level Results

One goal of the DQ analysis at the respondent level was to examine differences in DQ by mode and language. With the telephone and Internet modes, the technology used to implement the surveys enforced skip pattern logic and thus respondents did not have the opportunity to violate the skip patterns. They could, however, refuse to respond to items and thus rates of IE, TM, and item response could vary by mode.

Another goal of the DQ analysis at the respondent level was to identify individuals whose response patterns suggested that they might have had problems completing the survey. Such problems would be indicated by respondents with unusually high rates of the negative item dispositions (FS, IE, and TM), high rates of TFS, or low item response rates. We calculated some univariate statistics for the various item dispositions, both individually and collapsed into positive (CS and LR) and negative groupings. The means for these indicators are displayed in Exhibit 4 for all respondents and separately by completion status.[4]

Exhibit 4. Respondent level data quality rates by completion status

Item Dispositions	Overall Mean n=2,554	Not Complete Mean n=141	Complete Mean n=2,413
Rate of Correct Skips (CS)	43%	12%	45%
Rate of Failed Skips (FS)	2%	0%	2%
Rate of Indeterminate Eligibility (IE)	6%	52%	3%
Rate of Truly Missing (TM)	4%	23%	3%
Rate of Legitimate Responses (LR)	46%	12%	48%
Total	**100%**	**100%**	**100%**
LR + CS	89%	24%	93%*
FS + IE + TM	11%	76%	7%*

*Difference between completes and incompletes is statistically significant at (model F-test from a one-way analysis of variance; p < 0.001).

As can be seen in Exhibit 4, item dispositions obviously varied by completion status. This makes sense since to be considered to have completed a survey, the respondent must have answered a minimum number of questions. Positive item dispositions make up only 24% of the total for incompletes, compared to 93% of the total for completes.

Exhibit 5 displays item dispositions separately by mode of survey completion while controlling for survey language. As expected, there were no failed skips in either the Internet or telephone mode. The completion rate differed by mode, with mail producing the highest rate (98%), followed by Internet (93%) and telephone (80%).

---

[4] Those respondents with the highest rates of TM and IE and the lowest IRRs overlap considerably with those who are classified as incompletes. Given that the propensity to *complete* a survey almost completely overlaps with the propensity to *respond* (the completion rate among respondents is 94% overall and close to 100% for several modes), this section presents only a limited analysis of completion rates. The propensity to respond and response bias is analyzed in Section 4, and differences in respondent characteristics by mode or language are discussed in Section 5.

AIR ®
AMERICAN INSTITUTES FOR RESEARCH®

Exhibit 5. Respondent level data quality rates by survey mode

Item Dispositions Completed Surveys Only	Overall Mean n=2,554	Internet Mean n=225	Mail Mean n=1,889	Phone Mean n=440
Rate of Correct Skips (CS)	43%	41%	44%	39%
Rate of Failed Skips (FS)	2%	0%	2%	0%
Rate of Indeterminate Eligibility (IE)	6%	4%	3%	15%
Rate of Truly Missing (TM)	4%	3%	3%	8%
Rate of Legitimate Responses (LR)	46%	52%	47%	38%
Completion Rate	94%	93%	98%	80%*
Total	**100%**	**100%**	**100%**	**100%**

*Difference in completion rates among modes was statistically significant (model F-test from an analysis of covariance controlling for survey language; $p < 0.001$). Post-hoc comparisons showed that the average completion rate in each mode differed from the other two modes (multiple comparison procedures from a one-way analysis of variance, using the Tukey method to adjust for multiplicity, $p < 0.05$)

It is not clear why the completion rate for telephone respondents was lower than mail respondents. We further examined this survey mode effect among the English survey respondents by looking at the differences in completion rates by the five experimental groups. We found that the phone-only mode seemed to be driving the low completion rate for all phone respondents; the mail mode that included phone follow-up had a completion rate of 93% compared to 80% for the phone-only mode (see Exhibit 6).

Exhibit 6. Survey completion rates by experimental mode – English only

Experimental Mode	Completion Rate*
Mail - Mail	98%
Mail - FedEx	98%
Mail - Phone	93%
Web	90%
Phone - Only	80%

*Difference in completion rates among experimental modes was statistically significant (model F-test from an analysis of covariance; $p < 0.001$). Post-hoc comparisons showed that the average rate for Web was significantly lower than the two all-mail modes and that the completion rate for the phone-only mode was significantly lower than all other modes (multiple comparison procedures from a one-way analysis of variance, used the Tukey method to adjust for multiplicity, $p < 0.05$).

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Exhibit 7 displays the un-collapsed item disposition rates by language while controlling for survey mode. As shown, completion rates did not differ by language.

Exhibit 7. Respondent level data quality rates by survey language

Item Dispositions Completed Surveys Only	Overall Mean n=2,554	English Mean n=1,633	Spanish Mean n=381	Chinese Mean n=540
Rate of Correct Skips (CS)	43%	43%	42%	44%
Rate of Failed Skips (FS)	2%	1%	3%	2%
Rate of Indeterminate Eligibility (IE)	6%	6%	5%	3%
Rate of Truly Missing (TM)	4%	4%	3%	4%
Rate of Legitimate Responses (LR)	46%	45%	47%	47%
Total	**100%**	**100%**	**100%**	**100%**
Completion Rate	94%	92%	98%	99%*

*Difference in completion rates among languages wais not statistically significant (model F-test from an analysis of covariance controlling for survey mode).

Exhibits 8 and 9 summarize the differences in the rates of positive and negative item disposition indicators by survey mode and language respectively. The phone mode had the highest rate of negative item dispositions (9%), followed by mail (7%), with Internet at 1%. Since failed skips could not occur in the phone and Internet modes, the differences in this rate by mode is driven by the IE and TM rates.

Exhibit 8. Respondent level data quality rates by survey mode

Item Dispositions Completed Surveys Only	Overall Mean n=2,413	Internet Mean n=210	Mail Mean n=1,853	Phone Mean n=350
FS + IE + TM	7%	1%	7%	9%*
Rate of True Failed Skips (TFS)	4%	NA	5%	NA
Item Response Rate (IRR)	94%	98%	91%	83%†

*Difference in negative indicator rates among modes was statistically significant (model F-test from a one-way analysis of variance; p < 0.001). Post-hoc comparisons showed that the average negative indicator rate within each mode differed from each of the other two modes (multiple comparison procedures from a one-way analysis of variance used the Tukey method to adjust for multiplicity, p < 0.05).
†Difference in IRR among modes wais statistically significant (model F-test from a one-way analysis of variance; p < 0.001). Post-hoc comparisons showed that the average IRR for phone mode was lower than both of the other two modes (multiple comparison procedures from a one-way analysis of variance used the Tukey method to adjust for multiplicity, p < 0.05).

The Spanish language surveys had the highest rate of negative item dispositions (10%), true failed skips (10%), and the lowest item response rate, compared to English and Chinese. The English surveys, on average, had the lowest rates (Exhibit 9).

Exhibit 9. Respondent level data quality rates by survey language

Item Dispositions Completed Surveys Only	Overall Mean n=2,413	English Mean n=1,509	Spanish Mean n=372	Chinese Mean n=532
FS + IE + TM	7%	6%	10%	8%*
Rate of True Failed Skips (TFS)	4%	2%	7%	6%*
Item Response Rate (IRR)	94%	92%	88%	90%*

*Difference in rates among languages was statistically significant (model F-test from a one-way analysis of variance; p < 0.001). Post-hoc comparisons showed that the average rate within each language differed from each of the other two languages (multiple comparison procedures from a one-way analysis of variance used the Tukey method to adjust for multiplicity, p < 0.05)

3.1.2 Item Level

We also examined each of the five item dispositions at the *item level* across all respondents to assess the degree to which *each item* elicited quality responses (legitimate answers and correct skips) relative to problematic responses (missing responses and failed skips). Exhibit 10 provides an example for one of the survey items. As shown in the exhibit, around 8% of respondents correctly skipped this question based on their response to the screener question (q07), while around 4% failed to skip this item. Note also that the *true* failed skip rate here is 32%, which is the number of failed skips over the total number of items that the respondent should have skipped (96/301). The vast majority of respondents gave a legitimate response to this question.

Exhibit 10. Example of Item-level item dispositions

Q08: When you gave your household income information, was it easy to find out if you could get help paying for your health insurance?	Frequency	Percentage	Cumulative frequency	Cumulative percentage
Correct Skip	205	8.03	205	8.03
Failed Skip	96	3.76	301	11.79
Indeterminate Eligibility	89	3.48	390	15.27
Truly Missing	71	2.78	461	18.05
Legitimate Response	2093	81.95	2554	100.00

As illustrated above, the item-level results indicated, for each item, the proportion of respondents who provided each type of response. Exhibit 11 displays the distribution of each type of response across all respondents for a select set of survey item where the prevalence of a problematic item disposition (either FS, TM, or IE) was more than 5 percent. The rates sumed to 100% across the rows for each item, and the denominator for each percentage in a row was the 2,254 respondents (those who answered at least one question on the survey). The FS rate across all survey items ranged from a low of 0% to a high of 10%, while both the IE and TM rates ranged from 0% to 19%. As can be seen in the table, the LR rate varied quite a bit by item, ranging from a low of 0.5% (q88) to a high of 96% (q06, which is not shown in Exhibit 11).

Exhibit 11. Item dispositions by survey question (n = 2,254)

Question	CS Rate	FS Rate	IE Rate	TM Rate	LR Rate
q04	7.7%	3.9%	2.3%	6.7%	79.4%
q05	8.0%	3.6%	2.3%	13.0%	73.1%
q09	8.2%	3.6%	3.5%	5.4%	79.4%
q10	8.5%	3.3%	3.5%	9.7%	75.0%
q11	7.9%	3.9%	3.5%	9.0%	75.7%
q12	22.7%	7.4%	3.5%	7.4%	59.0%
q13	8.3%	3.4%	3.5%	9.7%	75.0%
q14	52.6%	7.4%	13.2%	1.6%	25.1%
q15	61.8%	6.1%	14.8%	0.4%	16.8%
q16	0.0%	0.0%	0.0%	6.2%	93.8%
q17	42.1%	6.0%	6.2%	1.8%	43.9%
q21_1	44.4%	0.6%	5.5%	3.7%	45.8%
q21_2	44.2%	0.8%	5.5%	3.7%	45.8%
q21_3	44.3%	0.7%	5.5%	3.7%	45.8%
q21_4	44.2%	0.8%	5.5%	3.7%	45.8%
q21_5	44.4%	0.6%	5.5%	3.7%	45.8%
q21_6	44.8%	0.2%	5.5%	3.7%	45.8%
q21_7	44.4%	0.5%	5.5%	3.7%	45.8%
q21_8	44.9%	0.1%	5.5%	3.7%	45.8%
q21_9	43.8%	1.2%	5.5%	3.7%	45.8%
q23_1	45.7%	1.1%	5.8%	4.2%	43.3%
q23_2	46.3%	0.5%	5.8%	4.2%	43.3%
q23_3	45.8%	0.9%	5.8%	4.2%	43.3%
q23_4	46.0%	0.7%	5.8%	4.2%	43.3%
q23_5	46.6%	0.2%	5.8%	4.2%	43.3%
q23_6	45.8%	0.9%	5.8%	4.2%	43.3%
q23_7	45.8%	0.9%	5.8%	4.2%	43.3%
q23_8	46.0%	0.8%	5.8%	4.2%	43.3%
q23_9	46.2%	0.5%	5.8%	4.2%	43.3%
q23_10	46.1%	0.6%	5.8%	4.2%	43.3%
q23_11	46.5%	0.2%	5.8%	4.2%	43.3%
q23_12	45.8%	1.0%	5.8%	4.2%	43.3%
q23_13	45.8%	0.9%	5.8%	4.2%	43.3%
q26	0.0%	0.0%	0.0%	5.8%	94.2%
q27	41.8%	1.6%	5.8%	1.0%	49.8%
q28_1	58.7%	0.1%	6.8%	2.5%	31.9%
q28_2	58.3%	0.5%	6.8%	2.5%	31.9%
q28_3	58.5%	0.4%	6.8%	2.5%	31.9%
q28_4	58.5%	0.3%	6.8%	2.5%	31.9%
q28_5	58.7%	0.1%	6.8%	2.5%	31.9%
q28_6	58.7%	0.1%	6.8%	2.5%	31.9%
q28_7	58.6%	0.2%	6.8%	2.5%	31.9%
q28_8	58.4%	0.4%	6.8%	2.5%	31.9%
q28_9	58.4%	0.4%	6.8%	2.5%	31.9%
q28_10	58.8%	0.0%	6.8%	2.5%	31.9%
q28_11	58.1%	0.7%	6.8%	2.5%	31.9%
q29	41.6%	1.7%	5.8%	1.6%	49.2%

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Question	CS Rate	FS Rate	IE Rate	TM Rate	LR Rate
q30_1	59.5%	0.8%	7.4%	4.3%	28.0%
q30_2	60.1%	0.2%	7.4%	4.3%	28.0%
q30_3	59.5%	0.7%	7.4%	4.3%	28.0%
q30_4	59.8%	0.5%	7.4%	4.3%	28.0%
q30_5	60.2%	0.1%	7.4%	4.3%	28.0%
q30_6	59.7%	0.5%	7.4%	4.3%	28.0%
q30_7	59.7%	0.5%	7.4%	4.3%	28.0%
q30_8	59.9%	0.4%	7.4%	4.3%	28.0%
q30_9	59.9%	0.3%	7.4%	4.3%	28.0%
q30_10	60.1%	0.2%	7.4%	4.3%	28.0%
q30_11	60.3%	0.0%	7.4%	4.3%	28.0%
q30_12	59.7%	0.5%	7.4%	4.3%	28.0%
q30_13	59.4%	0.9%	7.4%	4.3%	28.0%
q31	41.7%	1.6%	5.8%	1.8%	49.1%
q32	41.9%	1.5%	5.8%	2.5%	48.3%
q33	41.7%	1.6%	5.8%	1.4%	49.4%
q34	45.9%	1.6%	7.2%	0.9%	44.3%
q35	41.2%	2.2%	5.8%	1.8%	49.0%
q36	0.0%	0.0%	0.0%	7.3%	92.7%
q37	23.9%	3.6%	7.3%	3.5%	61.7%
q38	58.8%	6.4%	7.3%	1.6%	26.0%
q39_1	78.9%	0.9%	8.8%	2.9%	8.4%
q39_2	79.1%	0.7%	8.8%	2.9%	8.4%
q39_3	79.4%	0.4%	8.8%	2.9%	8.4%
q39_4	79.4%	0.4%	8.8%	2.9%	8.4%
q39_5	79.1%	0.7%	8.8%	2.9%	8.4%
q39_6	76.7%	3.1%	8.8%	2.9%	8.4%
q40	59.9%	5.3%	7.3%	1.8%	25.8%
q41_1	78.3%	0.8%	9.0%	2.0%	9.9%
q41_2	78.7%	0.4%	9.0%	2.0%	9.9%
q41_3	78.1%	1.0%	9.0%	2.0%	9.9%
q41_4	78.3%	0.7%	9.0%	2.0%	9.9%
q41_5	78.9%	0.2%	9.0%	2.0%	9.9%
q41_6	78.3%	0.8%	9.0%	2.0%	9.9%
q41_7	78.1%	1.0%	9.0%	2.0%	9.9%
q41_8	78.2%	0.9%	9.0%	2.0%	9.9%
q41_9	78.3%	0.7%	9.0%	2.0%	9.9%
q41_10	78.9%	0.2%	9.0%	2.0%	9.9%
q41_11	78.9%	0.2%	9.0%	2.0%	9.9%
q41_12	78.1%	1.0%	9.0%	2.0%	9.9%
q41_13	76.9%	2.2%	9.0%	2.0%	9.9%
q42	60.2%	5.0%	7.3%	2.1%	25.4%
q43	59.9%	5.3%	7.3%	2.3%	25.3%
q44	60.2%	5.0%	7.3%	2.0%	25.5%
q45	61.0%	4.2%	7.3%	1.9%	25.6%
q46	0.0%	0.0%	0.0%	6.1%	93.9%
q47	0.0%	0.0%	0.0%	6.7%	93.3%
q48	0.0%	0.0%	0.0%	7.0%	93.0%
q49	20.6%	2.4%	7.0%	1.4%	68.6%

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Question	CS Rate	FS Rate	IE Rate	TM Rate	LR Rate
q50	0.0%	0.0%	0.0%	7.8%	92.2%
q51	41.0%	4.6%	7.8%	0.7%	45.9%
q52	0.0%	0.0%	0.0%	8.0%	92.0%
q53	55.0%	6.2%	8.0%	0.5%	30.3%
q54	0.0%	0.0%	0.0%	7.6%	92.4%
q55	79.5%	6.5%	7.6%	0.4%	6.0%
q56	0.0%	0.0%	0.0%	7.6%	92.4%
q57	83.3%	7.8%	7.6%	0.1%	1.2%
q58	0.0%	0.0%	0.0%	7.6%	92.4%
q59	31.6%	3.2%	7.6%	2.0%	55.6%
q60	0.0%	0.0%	0.0%	7.9%	92.1%
q61	74.1%	3.5%	7.9%	0.6%	13.9%
q62	0.0%	0.0%	0.0%	9.3%	90.7%
q63	46.7%	3.1%	9.3%	1.5%	39.4%
q64	44.4%	5.4%	9.3%	1.3%	39.5%
q65	82.6%	6.0%	10.6%	0.0%	0.8%
q66	0.0%	0.0%	0.0%	9.7%	90.3%
q67	0.0%	0.0%	0.0%	8.1%	91.9%
q68	0.0%	0.0%	0.0%	6.2%	93.8%
q69	0.0%	0.0%	0.0%	6.3%	93.7%
q70	0.0%	0.0%	0.0%	8.2%	91.8%
q71	65.2%	7.0%	8.2%	0.4%	19.2%
q72	0.0%	0.0%	0.0%	8.3%	91.7%
q73	43.2%	4.5%	8.3%	0.7%	43.2%
q74	0.0%	0.0%	0.0%	7.7%	92.3%
q75	0.0%	0.0%	0.0%	7.9%	92.1%
q76	0.0%	0.0%	0.0%	7.9%	92.1%
q77	0.0%	0.0%	0.0%	7.8%	92.2%
q78	0.0%	0.0%	0.0%	7.9%	92.1%
q79	0.0%	0.0%	0.0%	8.0%	92.0%
q80	0.0%	0.0%	0.0%	7.6%	92.4%
q81	0.0%	0.0%	0.0%	8.1%	91.9%
q82	0.0%	0.0%	0.0%	8.0%	92.0%
q83	0.0%	0.0%	0.0%	7.8%	92.2%
q84	0.0%	0.0%	0.0%	9.2%	90.8%
q85_1	72.7%	0.2%	9.2%	0.0%	17.9%
q85_2	72.9%	0.0%	9.2%	0.0%	17.9%
q85_3	72.9%	0.0%	9.2%	0.0%	17.9%
q85_4	72.4%	0.5%	9.2%	0.0%	17.9%
q87	0.0%	0.0%	0.0%	9.8%	90.2%
q88	79.4%	10.1%	9.8%	0.2%	0.5%
q89	0.0%	0.0%	0.0%	6.0%	94.0%
q90	48.7%	8.4%	6.0%	0.4%	36.6%
q91	0.0%	0.0%	0.0%	6.3%	93.7%
q92	0.0%	0.0%	0.0%	6.5%	93.5%
q93	0.0%	0.0%	0.0%	6.1%	93.9%
q94	0.0%	0.0%	0.0%	19.2%	80.8%
q95_1	70.6%	0.9%	19.2%	0.1%	9.2%
q95_2	71.1%	0.4%	19.2%	0.1%	9.2%

Question	CS Rate	FS Rate	IE Rate	TM Rate	LR Rate
q95_3	71.0%	0.5%	19.2%	0.1%	9.2%
q95_4	70.9%	0.6%	19.2%	0.1%	9.2%
q95_5	67.7%	3.8%	19.2%	0.1%	9.2%

Because the CS and FS rates in Exhibit 11 used the total number of respondents as a denominator, the FS rate is somewhat misleading. The 'true' CS and FS rates that used as the denominator the total number of items the respondent should have skipped is a better indicator of the quality of response. This denominator was calculated only for those respondents who provided a valid response for the screening items that controlled the follow-up item(s). For example, the denominator for the five selections that were part of q02 was the number of respondents who answered 'no' to q01 (n = 2,199).

Exhibit 12 displays the true FS rates for all follow-up items on the survey, both overall and by language. Since the denominators vary by language and across sets of follow-up items, they are not displayed in Exhibit 12, though follow-up items controlled by the same screener (e.g., q07, q08 to q11) will have the same denominator. Note that failed skips could not occur in either the telephone or Internet modes because skip patterns were enforced by the survey technology such that respondents simply cannot make skip errors. We found exceptionally high FS rates for two sets of items: 1) q03 to q05, and 2) q08 to q13. When broken out by language, we found that the FS rates were even worse for Spanish and Chinese respondents, with close to half or more failing to follow the skip patterns. Both sets of items were part of skip patterns initiated by q02 and q07, respectively. These items were part of a complex set of nested skips that were clearly difficult to follow for at least one-third of English respondents and half or more of Chinese and Spanish respondents who should have skipped these items based on their screener responses. In addition, q11 to q16 included some additional nested skips branching off the legitimate path initiated by a 'yes' response to q06. The nested skips beginning with q13 appear to have reset some respondents, as the FS rate falls by about 50% for q14, q15, and q17.

Exhibit 12. True failed skip rates by language – mail mode only

Question	Overall	English	Spanish	Chinese
q02_1	1.1%	0.9%	2.6%	2.1%
q02_2	0.3%	0.2%	0.3%	0.6%
q02_3	0.2%	0.4%	0.3%	0.2%
q02_4	1.5%	1.6%	2.6%	2.3%
q02_5	1.7%	1.1%	6.1%	2.1%
q03	35.4%	34.1%	63.2%	53.3%
q04	33.7%	31.8%	63.2%	48.9%
q05	31.0%	31.1%	56.1%	42.2%
q07_1	0.9%	0.7%	2.0%	1.5%
q07_2	0.1%	0.1%	0.7%	0.0%
q07_3	0.1%	0.1%	0.0%	0.2%
q07_4	0.8%	0.5%	2.0%	1.5%
q07_5	1.2%	1.0%	4.0%	1.5%
q08	31.9%	31.9%	55.9%	40.4%
q09	30.2%	30.2%	55.9%	34.6%

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Question	Overall	English	Spanish	Chinese
q10	28.2%	25.9%	54.4%	34.6%
q11	33.2%	35.3%	55.9%	40.4%
q12	24.7%	22.2%	44.6%	31.4%
q13	29.2%	32.8%	48.5%	32.7%
q14	12.4%	11.7%	26.5%	15.5%
q15	9.0%	8.3%	20.4%	10.8%
q17	12.4%	11.7%	23.6%	19.0%
q19	10.4%	10.6%	13.0%	14.8%
q20	9.9%	10.3%	11.0%	14.8%
q21_1	1.3%	1.0%	2.1%	2.3%
q21_2	1.8%	2.4%	2.9%	1.5%
q21_3	1.5%	1.2%	4.2%	0.8%
q21_4	1.7%	1.4%	2.9%	2.7%
q21_5	1.4%	1.0%	2.1%	2.7%
q21_6	0.3%	0.5%	0.8%	0.0%
q21_7	1.2%	0.2%	0.8%	4.2%
q21_8	0.3%	0.0%	1.3%	0.0%
q21_9	2.6%	2.6%	3.8%	3.8%
q22	9.8%	9.9%	12.0%	13.9%
q23_1	2.3%	3.1%	2.5%	3.3%
q23_10	1.3%	2.4%	0.4%	1.9%
q23_11	0.5%	0.7%	0.4%	0.7%
q23_12	2.1%	2.4%	2.9%	3.0%
q23_13	1.9%	2.1%	2.9%	2.6%
q23_2	1.0%	1.7%	0.4%	1.5%
q23_3	1.9%	2.6%	1.6%	3.0%
q23_4	1.5%	2.1%	1.6%	1.9%
q23_5	0.4%	0.2%	0.8%	0.7%
q23_6	2.0%	2.6%	1.6%	3.3%
q23_7	2.0%	2.4%	2.5%	3.0%
q23_8	1.7%	2.1%	2.1%	2.2%
q23_9	1.2%	1.2%	0.8%	2.6%
q24	9.5%	9.6%	12.0%	13.5%
q25	9.4%	10.9%	12.0%	11.4%
q27	3.6%	2.4%	6.9%	7.3%
q28_1	0.2%	0.3%	0.4%	0.0%
q28_10	0.0%	0.0%	0.0%	0.0%
q28_11	1.3%	0.8%	4.0%	1.6%
q28_2	0.9%	0.8%	0.4%	2.5%
q28_3	0.6%	1.0%	0.4%	0.6%
q28_4	0.5%	0.7%	0.9%	0.6%
q28_5	0.2%	0.3%	0.4%	0.0%
q28_6	0.1%	0.2%	0.4%	0.0%
q28_7	0.3%	0.7%	0.4%	0.0%
q28_8	0.7%	0.7%	0.9%	1.6%
q28_9	0.7%	0.2%	0.0%	2.8%
q29	4.0%	2.8%	6.9%	8.1%
q30_1	1.3%	1.1%	2.2%	2.4%
q30_10	0.3%	0.2%	0.9%	0.3%

Question	Overall	English	Spanish	Chinese
q30_11	0.0%	0.0%	0.0%	0.0%
q30_12	0.8%	0.6%	2.2%	1.2%
q30_13	1.4%	1.1%	3.1%	2.4%
q30_2	0.3%	0.3%	0.9%	0.3%
q30_3	1.2%	1.6%	1.8%	1.5%
q30_4	0.8%	1.3%	0.9%	0.6%
q30_5	0.1%	0.2%	0.0%	0.3%
q30_6	0.8%	0.8%	1.8%	1.2%
q30_7	0.9%	0.6%	0.9%	2.4%
q30_8	0.6%	1.0%	0.9%	0.3%
q30_9	0.5%	0.5%	0.4%	1.2%
q31	3.8%	2.6%	6.9%	7.7%
q32	3.4%	2.4%	6.3%	6.9%
q33	3.8%	2.8%	6.9%	7.3%
q34	3.5%	2.3%	6.5%	7.0%
q35	5.0%	5.0%	8.3%	7.7%
q37	13.2%	16.1%	17.4%	13.3%
q38	9.8%	10.7%	20.8%	16.7%
q39_1	1.2%	1.4%	1.0%	2.2%
q39_2	0.9%	1.0%	2.1%	1.0%
q39_3	0.5%	0.8%	0.3%	0.7%
q39_4	0.5%	0.6%	1.0%	0.5%
q39_5	0.9%	0.2%	0.7%	3.4%
q39_6	3.9%	5.3%	4.5%	5.4%
q40	8.2%	8.1%	19.5%	14.6%
q41_1	1.0%	1.4%	0.7%	1.5%
q41_10	0.2%	0.2%	0.0%	0.8%
q41_11	0.2%	0.4%	0.0%	0.3%
q41_12	1.2%	1.4%	0.7%	2.8%
q41_13	2.7%	3.6%	5.8%	2.3%
q41_2	0.4%	0.6%	0.7%	0.5%
q41_3	1.2%	1.4%	0.7%	2.8%
q41_4	0.9%	1.2%	0.0%	2.3%
q41_5	0.2%	0.4%	0.0%	0.3%
q41_6	1.0%	1.4%	0.7%	1.5%
q41_7	1.2%	1.6%	1.1%	2.3%
q41_8	1.1%	1.3%	0.7%	2.3%
q41_9	0.9%	1.1%	1.1%	1.8%
q42	7.7%	7.5%	18.8%	13.7%
q43	8.2%	8.5%	20.1%	13.4%
q44	7.6%	7.6%	20.1%	12.5%
q45	6.5%	6.5%	15.6%	11.2%
q49	10.5%	11.3%	12.6%	14.8%
q51	10.0%	10.6%	15.2%	12.2%
q53	10.2%	8.7%	17.2%	15.8%
q55	7.6%	5.9%	15.7%	12.3%
q57	8.6%	7.3%	15.3%	15.4%
q59	9.1%	11.2%	13.5%	13.3%
q61	4.5%	3.4%	8.6%	13.2%

Question	Overall	English	Spanish	Chinese
q63	6.3%	5.7%	7.3%	12.8%
q64	10.9%	11.5%	16.1%	16.7%
q65	6.7%	5.2%	11.7%	13.2%
q71	9.7%	8.8%	18.8%	13.4%
q73	9.5%	6.8%	15.7%	14.7%
q85_1	0.3%	0.4%	50.0%	0.4%
q85_2	0.0%	0.0%	0.0%	0.0%
q85_3	0.0%	0.0%	0.0%	0.0%
q85_4	0.7%	0.5%	0.0%	1.8%
q88	11.2%	11.2%	23.9%	14.6%
q90	14.7%	22.9%	33.3%	30.0%
q95_1	1.3%	0.9%	2.5%	1.8%
q95_2	0.5%	0.2%	1.3%	0.9%
q95_3	0.7%	0.3%	1.9%	0.9%
q95_4	0.9%	0.1%	1.6%	2.3%
q95_5	5.3%	6.4%	7.6%	3.9%

True Failed Skip: the total number of failed skips divided by the sum of the number of correct skips and failed skips.

Denominators for these items (q03 to q17) were somewhat small (i.e., few respondents screened into them), and thus the total number of problem response was small in some cases even though the rates were high. These denominators are shown in Exhibit 13.

Exhibit 13. Denominators for fs rate – mail mode only

Question	Overall	English	Spanish	Chinese
q03	297	132	57	45
q04	297	132	57	45
q05	297	132	57	45
q06	0	-	-	-
q08	301	116	68	52
q09	301	116	68	52
q10	301	116	68	52
q11	301	116	68	52
q12	769	261	186	156
q13	301	116	68	52
q14	1534	643	234	342
q15	1735	709	265	398
q16	0	-	-	-
q17	1229	506	127	336

3.3   *Cleaning Response Inconsistencies*

The skip-logic cleaning is based on the results of the preceding DQ analysis. Two general problems were addressed by the data cleaning logic, both of which were problems related to skip logic:

1. IEs or follow-up items where a respondent's eligibility to respond cannot be determined (screener item was left blank)

2. FSs, that is, where the respondent answered the screener negatively but then disregarded the skip instructions and answered the follow-up items.

In dealing with the first issue (IEs), we applied the following logic:

a. If the screener was a yes/no type question and is blank or missing, AND if the follow-up question is a "how often" question that was answered "never," THEN we recoded the follow-up as missing (this assumes that the respondent was *ineligible* to answer the follow-up questions, which is supported by their response of "never" to the follow-up, but left the screener blank).

b. If the screener was a yes/no type question and was blank or missing, AND if the follow-up question was a "how often" question and the response given is "sometimes," "usually," or "always," THEN we kept the response to the follow-up and back-code the screener question to whatever response is associated with the respondent not skipping the follow-up item (this assumes that the respondent was *eligible* to respond to the follow-up, but left the screener blank).

c. For situations that do not correspond to the *a* or *b* logic above, if the screener was blank or missing AND the follow-up question was NOT a "how often" question BUT contained a valid response, THEN we kept the response to the follow-up (this requires no recoding). If the screener was a yes/no item, we back-coded the screener to whatever response was associated with the respondent not skipping the follow-up item (this assumes that the R was

eligible to respond but left the screener blank). This logic is essentially the same as *b*, but for follow-up items that were not "how often" items.

In dealing with the second issue (FSs), we followed this procedure: if the response to a screener question was valid but the respondent violated the skip instruction by answering survey items that should have been skipped, we kept the response to the screener and set the response to the follow-up as missing.

4.0 Response Rate and Response Bias

This section discusses survey response variations over the 12-week field test period, describes the calculation of the final response rate, and presents the methods and results of the nonresponse bias analysis.

4.1 Response across Field Period

As detailed in the Field Test Survey Design Report (Field Test Period, Deliverable 5.1b), the Marketplace Survey was administered in three languages (English, Spanish, and Chinese) with the English sample randomized into different groups as part of a survey mode experiment.

For the English survey, AIR randomly assigned sampled individuals into one of five experimental survey administration modes: (1) phone-only, (2) mail with phone follow-up, (3) mail-only with Fed Ex follow-up, (4) Web-only, and (5) mail-only with first-class follow-up. In addition, there is a sub-experiment for the Web-only group where a random half were sent an advance pre-notification letter by mail with a URL for the survey (4a) and remaining random half (4b) were surveyed using an entirely electronic mode (i.e., an emailed survey link instead of a mailed advance letter, with all reminders sent via email).

All English survey respondents in the non-Web modes were given the option to complete the survey online through a URL with username and password provided in the mailed advance pre-notification letter. Due to budget constraints, all respondents who received the survey in Spanish and Chinese received the survey only by mail. This experiment allowed AIR to evaluate which survey mode was the most cost-effective mode of administration on a per case basis.

Figure 1 shows the trends in survey response by mode across the twelve week field period (note that the final response rates shown in Figure 1 do not reflect the exclusion of ineligibles from the response rate denominator). Overall, the Chinese mail survey achieved the highest overall response rate (36%). Additionally, the Chinese mail survey yielded a higher rate of response earlier than the other modes, obtaining a 20% response rate by the second week of the field test. Of the experimental modes, the mail with phone follow-up mode achieved the highest response rate (31%). The experimental mode with the second highest response rate (27%) was the mail with FedEx follow-up. As shown in Figure 1, the mail with FedEx and English mail-only modes have similar response rates until the final mailing in week nine when respondents received the mailing via FedEx. Additionally, while data collection for the phone-only mode was completed more quickly than all other modes (telephone interviews occurred between weeks five and eight), the response rate for that mode was only 22 percent. Responses prior to week five for this mode

were due to consumers visiting the URL provided in their advance letter and completing the Internet version of the survey prior to the start of the telephone-based data collection.

Figure 1: Survey responses by week

4.2 Response Rate

In their review of Supporting Statement B of the OMB submission for this information collection (OMB Control Number 0938-1221), OMB asked us to use a more liberal response rate calculation than that which is the CAHPS standard.[5] Therefore, as noted in Statement B of the OMB submission, the response rate we proposed to use for the Marketplace Survey utilizes one of the industry standard response rates developed by the American Association for Public Opinion Research (AAPOR, 2011),[6] RR3, which is calculated as:

$$Response\,Rate\,(RR3) = \frac{C+P}{C+P+E+(X*U)}$$

where,

        C = Eligible, complete interview,

---

[5] The CAHPS standard calculation does *not* reduce the size of the denominator in the RR equation based on an estimate of the proportion of potentially eligible persons among all sampled persons whose eligibility cannot be confirmed. See p. 15 of this document for an example of the CAHPS RR calculation: https://cahps.ahrq.gov/surveys-guidance/survey4.0-docs/1033_CG_Fielding_the_Survey.pdf

[6] The American Association for Public Opinion Research. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* AAPOR, 2011. <http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions2&Template=/CM/ContentDisplay.cfm&ContentID=3156>

P = Eligible, partial interview

E = Eligible and not interviewed,

I = Ineligible (e.g., out of scope; only potential respondents who have explicitly indicated ineligibility were included here),

U = Unable to determine eligibility,

X = An estimate of the proportion of potential respondents in U who might be eligible, which is calculated as:

$$X ¿ \frac{C+P+E}{C+P+E+I} \, .$$

All sampled consumers were assumed to be eligible for the survey based on the design of the sampling methodology; that is, only those consumers who met the eligibility criteria for the survey were included in the sampling frame. Therefore, the term 'U' in the above equation is null, and all sampled consumers who did not respond (n=7,970) were classified under the 'E' term in the RR equation.

Furthermore, the survey itself did not include any eligibility screening questions, and thus ineligibility could only be determined based on some kind of explicit response provided by the sampled person. In all, 25 respondents provided such an explicit response and thus, even though 'X' could be calculated, the fact the 'U' is null causes the (X*U) portion of the RR equation to drop out. As a result, the actual RR calculation for this survey defaults to a standard CAHPS calculation:

$$CAHPS\ Response\ Rate\,(RR)=\frac{C+P}{C+P+E-I}$$

Because 'I' is so small, it had very little impact on the RRs.

For purposes of the Field Test, a completed survey was defined as a survey that had at least one question completed. Of the 2,554 surveys that were returned with at least one question completed, 2,411 of these surveys (94.4%) would meet the more traditional CAHPS definition of a completed survey which requires that 50 percent of the questions that were applicable to all respondents were answered, excluding the *About You* section, in order to be considered a complete. Response rates and completion rates based on the CAHPS standard calculation shown above are displayed in Exhibit 14 for each experimental mode group (all English), as well as for the Chinese and Spanish samples.

Exhibit 14. Response rate and completion rate by experimental mode group and language

Mode (Sample Size)	Number of Respondents	Response Rate (RR)*	Number of Completes	Completion Rate (CR)	Yield Rate (RR x CR)
Mail Total (n=7,514)	2,098	28.0%	2,034	96.9%	27.1%

Mode (Sample Size)	Number of Respondents	Response Rate (RR)*	Number of Completes	Completion Rate (CR)	Yield Rate (RR x CR)
Mail w/ Phone Follow-up (n=1,505)	**473**	**31.7%**	**441**	**93.2%**	**29.5%**
Mail Completes	306	20.5%	302	98.7%	20.2%
Online Completes	22	1.5%	20	90.9%	1.3%
Phone Completes	145	9.7%	119	82.1%	8.0%
Mail w/ Fed-Ex Follow-Up (n=1,504)	**403**	**26.8%**	**393**	**97.5%**	**26.2%**
Mail Completes	385	25.6%	375	97.4%	25.0%
Online Completes	18	1.2%	18	100.0%	1.2%
Mail (first class): English (1,505)	**301**	**20.0%**	**296**	**98.3%**	**19.7%**
Mail Completes	277	18.4%	272	98.2%	18.1%
Online Completes	24	1.6%	24	100.0%	1.6%
Mail (first class): Spanish (n=1,500)	**381**	**25.4%**	**372**	**97.6%**	**24.8%**
Mail (first class): Chinese (n=1,500)	**540**	**36.0%**	**532**	**98.5%**	**35.5%**
Web Only Total (n=1,505)	**133**	**8.8%**	**120**	**90.2%**	**8.0%**
Mail Pre-note (n=752)	79	10.5%	70	88.6%	9.3%
E-mail Only (n=753)	54	7.2%	50	92.6%	6.6%
Phone Only Total (n=1,505)	**323**	**21.6%**	**259**	**80.2%**	**17.3%**
CATI	295	19.8%	231	78.3%	15.5%
Web Option	28	1.9%	28	100.0%	1.9%
All Modes Total (n=10,524)	**2,554**	**24.3%**	**2,413**	**94.5%**	**23.0%**

*Ineligibles have been excluded from the denominators when calculating the response rate. The number ineligible include 12 for mail with phone follow-up, two for mail with Fed-Ex follow-up, and 11 for phone only (all among the CATI responses).

As shown in Exhibit 14, the overall response rate across all modes and languages was 24.3 percent, with the best rate obtained using mail with phone follow-up (for the English language sample). The yield rate shows a slight reduction in the number of useable surveys for each mode and language based on combining the response rate and completion rate. For those modes that only include mail surveys, the completion rate is close to 100 percent and thus does not impact the overall yield very much.

However, the completion rates were notably lower for respondents utilizing the online and phone modes. As shown, the online completion rate for those in the mail with phone follow-up mode matched the overall completion rate for the web-only mode (~90 percent). It is also notable that respondents using the telephone had a completion rate of only around 80 percent: CATI respondents come in at just over 78 percent and phone respondents in the mail with phone follow-up mode have an 82 percent completion rate. Still, even when taking the completion rate into consideration, the mail with phone follow-up mode produced usable surveys at a higher rate than any other mode (for English). Given these results, AIR plans to conduct the beta test survey data collection using this mode for all three languages.

As shown in Exhibit 14, the response rates for the Chinese and Spanish samples were 36 percent and 25 percent respectively; the response rate for the English sample across all modes was 22 percent (not shown), but was 32 percent for the mail with phone follow-up mode. Evidence from the field test presented here suggests that 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 percent, and thus the phone follow-up appears to produce a 12 percentage point increase in the response rate (from

20 percent to 32 percent, which is a 60 percent increase). Even when taking the completion rate into consideration, the yield rate increases by 10 percentage points (from 20 to 30 percent, which is a 50 percent increase).

Comparable increases among the Chinese and Spanish samples obtained by using this mode during the beta test could result in response rates of 54 to 58 percent and 38 to 40 percent respectively; however, without knowing how Chinese and Spanish respondents will respond to a telephone survey in their preferred languages, it is probably unwise to expect that actual response rates will be this high. For the beta test we used the response rates observed from the field test data collection to calculate the beta test sample sizes. 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.

4.3 Nonresponse Bias Analysis

Unit nonresponse, which is typically referred to simply as nonresponse, occurs when a sampled individual fails to complete the survey. While nonresponse alone is not problematic, if the topic being measured in the survey is directly related to the reason for nonresponse it can result in nonresponse bias and inaccurate survey estimates. For example, if individuals who had poor experiences with the Health Insurance Marketplace were significantly less likely to complete the Marketplace Survey, then the survey results may overestimate the percentage of Marketplace users who had a positive experience.

Given the potential detrimental impact that nonresponse bias can have on survey estimates, the Office of Management and Budget (OMB) requires that all federal surveys that achieve a response rate below 80 percent perform a nonresponse bias analysis (Office of Management and Budget, 2006). The analysis presented in this section fulfills this requirement for the Marketplace Survey.

Because the Marketplace Survey was a sample of individuals who interacted with the Health Insurance Marketplace, the information that sampled individuals provided as part of their Marketplace application, and included in the sampling frame, provided several variables to use for nonresponse bias analysis. Exhibit 15 provides a full listing of all the variables that were reviewed in this analysis.

Exhibit 15: Variables analyzed in nonresponse bias cross-tabs

Demographic	Geographic	Administrative
Sex Age Race Disability Status Citizenship Status Language Preference	State Census Region	Applicant Status Mode of Application Applicant provided Telephone Number Applicant provided Email Address Eligibility for Medicaid Eligibility for Advanced Premium Tax Credit (APTC)

First, AIR examined cross-tabs of respondents and nonrespondents across the demographic, geographic, and administrative variables that were included on the sampling frame. In reviewing these cross-tabs, AIR found statistically significant differences (Chi-sq test, $p < 0.05$) between respondents and nonrespondents for sex, age, race, disability status, Census region, applicant status, mode of application, and eligibility for the APTC.

Additionally, AIR compared distributions of variables that were included on the sampling frame among the full sample with the percentages among survey respondents. Selected results of this analysis are presented Exhibit 16. As shown, the distribution of demographic characteristics among survey respondents were roughly in line with the full sample. The notable exceptions to this were age and application status. Younger respondents were much less likely to respond to the survey resulting in the percentage of respondents over the age of 55 being 10.8 percentage points higher than the percentage in the sample. Additionally, the percentage of enrollees was 9.5 percentage points higher among respondents than in the sample. The fact that the percentage of enrollees within the respondents was much higher is concerning because this could indicate that individuals who did not complete an application were underrepresented within the survey responses.

We also found that the preferred language distribution among survey respondents was skewed away from English and toward Chinese; however, this merely reflects the above average response rates among Chinese (36% compared to an average of 24%) and the below average response rates among the English respondents (22% compared to an average of 24%). The response rate among Spanish was roughly equal to the overall average response rate (25% compared to 24%), and this too was reflected in the distributions by language in Exhibit 16.

Exhibit 16. Comparison of selected demographics between full sample and survey respondents

Demographic	Percentage of Sample (n=10,524)	Percentage of Survey Respondents (n=2,554)
Sex		
Male	44.58%	41.2%
Female	54.35%	58.2%
Missing	1.1%	0.7%
Age		
18 - 32	26.3%	17.2%
33 - 44	24.5%	19.5%
45 – 54	23.3%	26.5%
55 or older	26.0%	36.8%
Citizenship Status		

AIR ®
AMERICAN INSTITUTES FOR RESEARCH'

Demographic	Percentage of Sample (n=10,524)	Percentage of Survey Respondents (n=2,554)
Citizen	70.6%	69.2%
Non-Citizen	29.5%	30.8%
Disability Status		
Has disability	5.6%	6.6%
No disability	94.4%	93.4%
Application Status		
Potential Applicant (PA)	15.9%	12.1%
Potential Enrollee (PE)	41.4%	35.7%
Enrollee (E)	42.7%	52.2%
Language Preference		
English	71.5%	63.9%
Spanish	14.3%	14.9%
Chinese	14.3%	21.1%
Race		
White	45.5%	42.9%
Black or African American	8.9%	6.0%
American Indian or Alaskan Native	0.9%	0.7%
Asian	14.1%	20.0%
Native Hawaiian or Pacific Islander	0.1%	0.1%
Multiple Races	2.4%	2.2%
Missing	28.1%	28.2%
Eligible for Medicaid		
Yes	9.9%	10.3%
No	88.5%	88.1%
Missing	1.6%	1.5%
Eligible for Advanced Premium Tax Credit (APTC)		
Yes	48.7%	56.6%
No	49.7%	41.9%
Missing	1.6%	1.5%

Given that the bivariate analysis indicated some significant differences in response by consumer characteristics, AIR utilized a multivariate logistic regression in order to determine which consumer characteristics were associated with returning the survey and to estimate the direction and size of the effect of these characteristics. The outcome for this regression was a dummy coded variable where a value of 1 indicates that the sampled consumer was a respondent and a value of zero indicates that the sampled consumer was a non-respondent, and the model was set up to estimate the propensity to respond.

Exhibit 17 shows the marginal effect (odds ratio) of each variable on the propensity to respond as well as the 95% confidence interval associated with the estimate. Estimates above 1.0 indicate that the variable was associated with an increased propensity to respond in comparison to the reference group, while estimates below 1.0 indicate that the variable was associated with a lower propensity to respond to the survey in comparison to the reference group. For example, males were slightly less likely to return the survey than females (OR = 0.915). All odds ratios shown in Exhibit 17 were statistically significant ($p < 0.05$). Non-significant consumer characteristics are

not shown but were retained in the model to control for their effect on the propensity to respond (see Exhibit 15 for a list of all variables included in the model).

Exhibit 17. Odds ratios from variables included in logistic regression modeling survey response

Effect	Odds Ratio Estimate	Lower 95% Wald Confidence Limit	Upper 95% Wald Confidence Limit
Sex *(Ref: Female)*			
Male	0.915	0.912	0.918
Missing	0.304	0.295	0.313
Age *(Ref: Over 55)*			
18 - 32	0.426	0.423	0.428
33 - 44	0.445	0.442	0.447
45 – 54	0.687	0.684	0.690
Application Status *(Ref: Enrollee)*			
Potential Applicant (PA)	0.699	0.694	0.704
Potential Enrollee (PE)	0.755	0.751	0.759
Language Preference *(Ref: English)*			
Spanish	1.805	1.733	1.879
Chinese	2.692	2.662	2.722
Race *(Ref: White)*			
Black or African American	0.806	0.802	0.811
American Indian or Alaskan Native	1.566	1.538	1.595
Asian	0.881	0.873	0.890
Native Hawaiian or Pacific Islander	1.358	1.204	1.532
Multiple Races	0.434	0.427	0.442
Missing	0.703	0.699	0.707
Eligibility for Advanced Premium Tax Credit (APTC) *(Ref: Ineligible for APTC)*			
Yes	1.086	1.081	1.092
Missing	1.755	1.727	1.783
Assigned Mode *(Ref: Mail-Mail-Mail Mode)*			
Mail with Phone Follow-Up	2.135	2.122	2.149
Mail with FedEx Follow-Up	1.652	1.643	1.662
Telephone Only	1.239	1.231	1.247
Web Only with Email Notification	0.250	0.247	0.253
Web Only with Mail Notification	0.399	0.395	0.403
Telephone Number Provided with Application *(Ref: Yes)*			
No Telephone Number Provided	0.896	0.891	0.901
Email Address Provided with Application *(Ref: Yes)*			
No Email Address Provided	1.204	1.198	1.211

As shown in Exhibit 17, the results from the multivariate logistic regression confirm that individuals who were potential applicants (PA) or potential enrollees (PE) were significantly less likely to return the survey compared to enrollees (E) or effectuated enrollees (EE). Additionally, we see that individuals who were eligible for the APTC were slightly more likely to return the survey. Given that both application status and eligibility for the APTC were correlated with the

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

measurement goals of the Marketplace Survey this indicates that there was a potential for nonresponse bias and therefore nonresponse adjustments are needed.

4.4. Nonresponse Weights

Survey weights were generated in three stages. First, sampling weights were generated to account for respondents' unequal probability of selection by using the inverse of their probability of selection. Since we used a stratified sampling approach for the English sample, all consumers in the same state have the same weight and weights differ by state. Both the Chinese and Spanish sampling frames were FFM-level strata from which we selected a systematic random sample. As a result, all consumers in the Spanish sample have the same weight and all consumers in the Chinese sample have the same weight, and these weights differ by language.

Second, a separate weight was created to adjust for unit nonresponse The propensity to respond was adjusted for non-response bias by calculating the predicted propensity to respond for each respondent using the multivariate logistic regression utilized in the nonresponse bias analysis discussed above. The logistic model adjusts the propensity to respond by the vector of predictors included in the model. For example, potential applicants (PAs) would have their propensity to respond adjusted up to compensate for their lower odds of responding to the survey, while Chinese respondents would have their propensity to respond adjusted down to compensate for their higher odds of responding. This non-response weight was calculated by taking the inverse of each respondent's adjusted propensity to respond.

Finally, the sampling and nonresponse weights were multiplied to create the final weight, thus each weight is equal to:

$$W_{Final} = W_{Sampling} \, x \, W_{Nonresponse}$$

where $W_{Sampling}$ is equal to the inverse of the individual's probability of selection and $W_{Nonresponse}$ is equal to the inverse of the individual's adjusted propensity to respond.

Exhibit 18 shows the means of the four global rating questions using the sampling weights and the final weights.

Exhibit 18. Comparison of global rating questions using sampling weights and final weights

Survey Estimate	Mean using Sampling Weights (Standard Error)	Mean using Final Weights (Standard Error)
Global Rating of Health Insurance Marketplace Website (Healthcare.gov)	5.12 (.073)	5.22 (.072)
Global Rating of Health Insurance Marketplace Help Line	6.44 (.083)	6.45 (.084)
Global Rating of In-Person Assistance	7.96 (.104)	8.25 (.097)
Overall Global Rating of the Health Insurance Marketplace	5.82 (.065)	5.82 (.065)

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

Ultimately, the impact of adding non-response weights to the sampling weights was minimal with respect to the four global ratings. Also, the additional weighting had little or no impact on the standard errors associated with these estimates. These findings suggest that while some key groups had differing response propensities, the underlying survey data, at least at the level of the entire FFM (36 states), do not suffer from nonresponse bias. On the other hand, when we analyzed variation in scores among states we found that the use of weights had a substantial effect on this variation, and the non-response weights contributed more to this variance when combined with the sampling weights compared to using the sampling weights alone. These results are discussed in more detail in Section 6.3.2.

5.0 Effect of Mode of Administration

5.1 Evaluation of Mode Experiments

This section examines the impact of the Marketplace Survey Field Test modes of administration and nonresponse follow-up on response rates and respondent characteristics. The section begins with an overview of response rates, followed by a discussion of cost impact, and then discussion of differences in respondent characteristics by mode of response and mode of nonresponse follow-up. The mode of response was the mode the sampled participant used to answer the survey (mail, phone, web); mode of follow-up was the mode by which nonrespondents were contacted for the third and final contact.

5.1.1 Overview of Response Rates

Exhibit 14 in Section 4.1 displayed the response rates by each mode of administration and nonresponse follow-up type. As described in that sections, excluding the Chinese and Spanish language groups, mail with phone follow-up for nonrespondents achieved the highest response rate (32 percent), followed by the group with FedEx follow-up (27 percent). The English group receiving first-class mail only had a response rate of 20 percent. The response rate for the web experiment group was very low, at 9 percent.

Around 75 percent of the over 8 million consumers in the sampling frame provided a phone number, and this rate was slightly greater among the sampled consumers (just under 79 percent). Only consumers who had provided a phone number were eligible to be assigned to the two modes that involved phone. Approximately 80 percent of phone numbers dialed during the were cell phone numbers.

5.1.2 Cost Impact

The cost of the FedEx mailing was expensive for the Marketplace field test. The overall cost per complete for cases in the mail with FedEx nonresponse follow-up group was approximately $75 compared to approximately $60 for cases in the mail with phone nonresponse follow-up group.

5.1.3 Response Characteristics by Mode of Response

This section examines differences in respondent characteristics by the mode in which they responded. It does not take into account when the participant responded or the mode of initial contact or follow-up. These analyses also exclude the Chinese and Spanish respondents since they were only given the opportunity to complete the survey by mail and would therefore skew

the results by mode. Examining mode of response provides and overall picture of the types of participants who respond by phone, mail, or web.

Exhibit 19 shows the distribution of characteristics for each mode of survey response and the frame. There were noticeable differences by mode. The phone respondent distribution was generally more comparable to the frame distribution. The phone captures a higher percentage of potential enrollees and participants who were younger, Black, have lower incomes, and who have a disability compared to mail and web. The mail captures a higher percentage of participants who were older and APTC or CSR eligible compared to phone and web. The web was not as successful as mail and phone in bringing in a diverse group of respondents.

Exhibit 19. Selected respondent characteristics from the sampling frame by mode of survey response

Frame characteristics	Sampling Frame	Mode of Response: Mail (n=968)	Mode of Response: Phone (n=440)	Mode of Response Web (n=225)
Applicant Status*				
Enrolled	38.4%	50.6%	38.9%	47.6%
Effectuated Enrollee	0.5%	0.6%	0.9%	0.4%
Potential Applicant	16.0%	10.1%	13.0%	12.9%
Potential Enrollee	45.1%	38.6%	47.3%	39.1%
Region*				
Northeast	30.6%	29.4%	27.3%	27.1%
Midwest	30.6%	34.0%	29.8%	31.6%
South	19.5%	17.6%	23.4%	13.3%
West	19.4%	19.0%	19.6%	28.0%
Male	43.6%	37.8%	39.3%	39.3%
Age*				
18-24	9.1%	4.8%	7.5%	2.7%
25-34	25.8%	14.8%	23.6%	22.2%
35-44	19.1%	14.1%	14.6%	15.6%
45-54	20.8%	24.5%	22.7%	18.2%
55-64	24.4%	40.4%	31.6%	40.0%
65-74	0.8%	1.3%	0.0%	0.9%
75+	0.2%	0.1%	0.0%	0.4%
Race*				
White	76.4%	81.6%	72.4%	93.1%
Black or African American	16.4%	11.8%	19.1%	3.9%
Other Specified	7.2%	6.6%	8.6%	2.9%
APTC or CSR Eligible*	45.0%	56.8%	43.8%	50.2%
Mean household income	29,358	30,286	25,900	31,799
Disability Status	7.7%	9.8%	11.7%	6.0%

*Indicates differences among survey response modes are statistically significant at p<=.05. Pairwise differences between each mode and the frame were not tested. Analysis includes English respondents only.

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Exhibit 20 shows self-reported characteristics of survey respondents by mode of survey response; this analysis only included respondent characteristics obtained via the survey and was limited to English respondents. Like the frame analysis, there were again noticeable differences by mode of response, particularly for web compared to mail and phone. Mail and phone respondents share similar characteristics while web respondents were quite different. Web respondents appeared to overall be in better health. Compared to mail and phone respondents, a smaller percentage of web respondents reported fair or poor health; serious difficulty concentrating, remembering, or making decisions; walking or climbing stairs; or dressing and bathing. Web respondents reported a higher level of education than mail or phone respondents and a smaller percentage were unemployed. A greater percentage of web respondents reported having health insurance in 2013.

Exhibit 20. Selected survey characteristics by mode of response

Respondent Characteristics	Mode of Response: Mail (n=968)	Mode of Response: Phone (n=440)	Mode of Response Web (n=225)
q68: Overall health rating*			
Excellent	16.1%	17.0%	15.9%
Very good	34.1%	34.8%	39.4%
Good	29.8%	24.0%	33.7%
Fair	15.3%	18.1%	7.7%
Poor	4.8%	6.1%	3.4%
q69: Overall mental or emotional health			
Excellent	34.0%	31.8%	31.7%
Very good	31.4%	31.8%	37.5%
Good	24.3%	25.4%	17.8%
Fair	8.0%	7.9%	11.1%
Poor	2.2%	3.2%	1.9%
q70: Health care 3 or more times for same condition	27.4%	27.7%	26.9%
q71: If yes, condition that has lasted for at least 3 months	81.9%	83.0%	89.3%
q72: Need or take medicine prescribed by a doctor	60.4%	53.5%	58.7%
q73: If yes, medicine to treat a condition that has lasted for at least 3 months	93.5%	92.3%	94.2%
q74: Deaf or has serious difficulty hearing	4.1%	2.9%	3.9%
q75: Blind or has serious difficulty seeing*	2.9%	7.6%	2.9%
q76: Serious difficulty concentrating, remembering, or making decisions*	10.3%	12.3%	4.8%
q77: Serious difficulty walking or climbing stairs*	12.9%	14.9%	6.3%
q78: Difficulty dressing or bathing*	4.2%	4.4%	0.5%
q79: Difficulty doing errands alone such as visiting a doctor's office or shopping	7.8%	7.9%	4.4%
q80: Age*			
18 to 24 years	4.1%	6.7%	2.4%
25 to 34	14.1%	23.6%	20.7%
35 to 44	14.9%	14.0%	14.9%

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

Respondent Characteristics	Mode of Response: Mail (n=968)	Mode of Response: Phone (n=440)	Mode of Response Web (n=225)
45 to 54	22.8%	23.3%	17.8%
55 to 64	41.3%	31.2%	43.3%
65 to 74	2.6%	1.2%	1.0%
75 or older	0.2%	0.0%	0.0%
Male	36.9%	37.6%	38.9%
q82: Education*			
8th grade or less	2.2%	2.0%	0.5%
Some high school, but did not graduate	5.2%	5.5%	1.5%
High school graduate or GED	26.0%	28.6%	17.9%
Some college or 2-year degree	38.2%	38.2%	33.8%
4-year college graduate	15.5%	16.0%	22.2%
More than 4-year college graduate	12.8%	9.6%	24.2%
q83: Employment*			
Employed full-time	34.1%	37.7%	39.3%
Employed part-time	18.1%	20.5%	21.8%
A homemaker	5.7%	6.1%	3.4%
A full-time student	2.0%	4.1%	3.9%
Retired	11.0%	5.6%	11.7%
Unable to work for health reasons	10.0%	12.3%	5.8%
Unemployed	9.8%	8.8%	5.3%
Other	9.3%	5.0%	8.7%
q84: Hispanic, Latino/a, Spanish origin	6.7%	8.2%	3.9%
q86: Race*			
White	80.4%	68.7%	89.5%
Black or African American	12.4%	18.0%	5.0%
Other Specified	7.2%	13.3%	5.5%
q87: Eligible for health services from Indian Health Service*	1.2%	1.2%	0.0%
q90: How well do you speak English*			
Very well	78.5%	44.4%	33.3%
Well	13.4%	55.6%	50.0%
Not well	6.9%	0.0%	16.7%
Not at all	1.2%	0.0%	0.0%
q91: Had health insurance in US between Jan 1 and Dec 31, 2013*	53.4%	49.1%	63.3%
q92: Confidence in understanding health insurance terms*			
Not at all confident	11.9%	5.6%	6.8%
Slightly confident	22.1%	22.3%	24.8%
Moderately confident	40.7%	43.3%	40.8%

Respondent Characteristics	Mode of Response: Mail (n=968)	Mode of Response: Phone (n=440)	Mode of Response Web (n=225)
Very confident	25.3%	28.8%	27.7%

*Indicates differences among survey response modes are statistically significant at p<=.05. Pairwise differences between each mode and the frame were not tested.

5.1.4 Respondent Characteristics by Mode of Nonresponse Follow-up

The response rate analysis showed that the two modes with the highest response rates were mail with phone follow-up mode (32%) and mail with FedEx follow-up (27%). To help determine what the implications of choosing one over the other might be in terms of response bias, we examined differences in respondent characteristics by mode of third contact: FedEx versus phone.

The characteristics of respondents in the FedEx and phone follow-up groups were somewhat similar compared to the observed differences by mode of response. Exhibits 21 and 22 show the distributions of frame-based and survey-based characteristics by follow-up mode. Phone respondents were younger than FedEx respondents, had lower incomes, and a greater percentage were Black. Compared to the frame, using phone as a follow-up appears to underrepresent Whites and individuals of 'other' races (i.e., not Black or White), and over represent Blacks. Since Blacks were underrepresented in the mail and web modes (see Exhibits 19 and 20), phone follow-up may be a good way to address any potential response bias related to race.

Similar to participants who responded by mail or phone as reported earlier, a smaller percentage of phone respondents compared to FedEx respondents were APTC or CRS eligible, although these estimates were not measurably different. One difference seen in the follow-up analysis but not seen in our other mode analyses, was that a greater percentage of phone follow-up respondents expressed confidence in understanding health insurance terms compared to FedEx respondents. This finding suggests perhaps phone respondents were less likely to feel comfortable reporting lack of confidence in their understanding of health insurance terms to an interviewer.

Exhibit 21. Selected respondent characteristics from the sampling frame by mode of survey response among late responders

Frame characteristics	Sampling Frame	Mode of Late Response: FedEx (n=131)	Mode of Late Response: Phone (n=145)
Applicant Status			
Enrolled	38.4%	41.2%	37.9%
Effectuated Enrollee	0.5%	0.0%	1.4%
Potential Applicant	16.0%	14.5%	13.8%
Potential Enrollee	45.1%	44.3%	46.9%
Region			
Northeast	30.6%	29.0%	26.2%
Midwest	30.6%	30.5%	33.8%

Development of an Enrollee Satisfaction Survey for Use in the Health Insurance Marketplace

Frame characteristics	Sampling Frame	Mode of Late Response: FedEx (n=131)	Mode of Late Response: Phone (n=145)
South	19.5%	22.9%	24.1%
West	19.4%	17.6%	15.9%
Male	43.6%	41.1%	41.7%
Age*			
18-24	9.1%	4.6%	8.3%
25-34	25.8%	17.6%	31.0%
35-44	19.1%	20.6%	17.2%
45-54	20.8%	22.9%	15.9%
55-64	24.4%	32.8%	27.6%
65-74	0.8%	1.5%	0.0%
75+	0.2%	0.0%	0.0%
Race*			
White	76.4%	84.8%	69.5%
Black or African American	16.4%	10.9%	19.0%
Other Specified	7.2%	4.4%	11.6%
APTC or CSR Eligible	45.0%	55.0%	46.5%
Mean household income	29,358	32,888	29,409
Person with a disability	7.7%	12.3%	8.2%

*Indicates differences among survey late response modes are statistically significant at p<=.05. Pairwise differences between each mode and the frame were not tested. Analysis includes English respondents only.

Exhibit 22. Selected survey characteristics by mode of survey response among late responders

Frame characteristics	Mode of Late Response: FedEx (n=131)	Mode of Late Response: Phone (n=145)
q68: Overall health rating		
Excellent	20.8%	16.6%
Very good	29.2%	40.0%
Good	26.7%	22.6%
Fair	18.3%	15.7%
Poor	5.0%	6.1%
q69: Overall mental or emotional health		
Excellent	30.8%	33.6%
Very good	30.8%	35.3%
Good	26.7%	18.1%
Fair	9.2%	6.9%
Poor	2.5%	6.0%
q70: Health care 3 or more times for same condition	26.9%	31.0%
q71: If yes, condition that has lasted for at least 3 months	78.1%	68.6%
q72: Need or take medicine prescribed by a doctor	58.0%	47.4%
q73: If yes, medicine to treat a condition that has lasted for at least 3 months	95.5%	90.9%

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

Frame characteristics	Mode of Late Response: FedEx (n=131)	Mode of Late Response: Phone (n=145)
q74: Deaf or has serious difficulty hearing	2.5%	2.6%
q75: Blind or has serious difficulty seeing	3.4%	6.9%
q76: Serious difficulty concentrating, remembering, or making decisions	6.7%	8.8%
q77: Serious difficulty walking or climbing stairs	11.8%	13.8%
q78: Difficulty dressing or bathing	3.4%	4.3%
q79: Difficulty doing errands alone such as visiting a doctor's office or shopping	10.2%	9.5%
q80: Age		
18 to 24 years	5.9%	8.6%
25 to 34	16.0%	31.0%
35 to 44	19.3%	15.5%
45 to 54	22.7%	17.2%
55 to 64	32.8%	26.7%
65 to 74	2.5%	0.9%
75 or older	0.8%	0.0%
Male	37.3%	40.5%
q82: Education		
8th grade or less	3.4%	1.7%
Some high school, but did not graduate	6.8%	6.0%
High school graduate or GED	22.0%	24.1%
Some college or 2-year degree	45.8%	43.1%
4-year college graduate	11.0%	13.8%
More than 4-year college graduate	11.0%	11.2%
q83: Employment*		
Employed full-time	39.3%	42.6%
Employed part-time	13.7%	24.4%
A homemaker	9.4%	5.2%
A full-time student	2.6%	5.2%
Retired	6.0%	1.7%
Unable to work for health reasons	7.7%	9.6%
Unemployed	8.6%	6.1%
Other	12.8%	5.2%
q84: Hispanic, Latino/a, Spanish origin	14.9%	9.6%
q86: Race*		
White	82.3%	66.7%
Black or African American	11.5%	18.9%
Other Specified	6.2%	14.4%
q87: Eligible for health services from Indian Health Service*	2.6%	0.9%
q90: How well do you speak English		

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Frame characteristics	Mode of Late Response: FedEx (n=131)	Mode of Late Response: Phone (n=145)
Very well	69.4%	40.0%
Well	13.9%	60.0%
Not well	13.9%	0.0%
Not at all	2.8%	0.0%
q91: Had health insurance in US between Jan 1 and Dec 31, 2013	45.8%	50.4%
q92: Confidence in understanding health insurance terms*		
Not at all confident	24.2%	4.3%
Slightly confident	17.5%	25.0%
Moderately confident	37.5%	44.8%
Very confident	20.8%	25.9%

*Indicates differences across columns are statistically significant at p<=0.05.

5.2 Recommendation

Based on the analysis above, we recommend using mail with phone follow-up for the Marketplace beta test and future surveys. The comparison of the respondents captured by the follow-up using these two methods does not show substantial differences. This finding suggests that follow-up by phone is sufficient for nonresponse follow-up. However, if CMS would like increase response rates and minimize any potential impact of non-coverage bias for nonrespondents who have not provided a phone number, FedEx could be used as the final follow-up for those cases.

6.0 Evaluation of Reliability and Validity

The Health Insurance Marketplace Survey includes 29 substantive report items that were hypothesized as observed indicators of 8 unique latent constructs.[7] This factor structure was developed based on the results of the formative research and was intended to produce a set of domains that reflect the most salient features of the experience of the health insurance marketplace by consumers—in other words, to produce measures that have high face validity.

The guiding principle in the specification of the initial factor structure is to create a set of domains that are meaningful to consumers and that capture the full span of relevant dimensions of Marketplace experience for consumers. The eight hypothesized domains included:

1. Application Process
2. Premium Tax Credit Eligibility
3. Seeking Information on the Marketplace Website
4. Seeking Information over the Phone

---

[7] We use the terms "construct" or "domain" to refer to latent, unobserved phenomena; the term "composite" refers to the concrete measures that are calculated by mathematically combining the observed indicators of a construct into a single measure such as when calculating the mean of the indicators. A construct, or factor, is a theoretical entity, whereas a composite is an empirical measure.

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

5. Seeking Information In-Person

6. Health Plan Enrollment Process

7. Specialized Services

8. Cultural Competence

The analytic data set used for this analysis was based on the data set produced after the cleaning described in Section 3.3 of this report. Some questions and responses from the survey were phrased so that a higher score indicates a more positive experience, whereas others were phrased so that a higher score indicates a more negative experience. To facilitate the interpretation of the results, we reverse-scored responses to some questions so that in each case a higher score indicated a better care experience.

This survey produces some structured missing data because of screener items that direct ineligible respondents to skip questions that were irrelevant to their experience. When a list-wise deletion of missing data is implemented in subsequent factor analysis models, very few respondents may remain available for the analysis given that most have a missing response to at least one observed variable being used in the analysis.

We used an approach that makes use of all available responses by implementing a pairwise deletion of missing data, as opposed to the list-wise deletion. Such an approach is possible using Mplus (Muthén & Muthén, 1998-2011). In addition, Mplus allows one to specify observed variables as categorical (either binary or ordinal) and thus produces tetrachoric and polychoric correlations among these variables to be used as input for structural equation models. Factor analyses were then based on these more appropriately modeled relationships among the observed variables.

In the psychometric analysis of the test data, composites were evaluated based on the following criteria:

1. Results of a confirmatory factor analysis (CFA) show a good fit between the hypothesized composites and the observed psychometric test data. The fit is considered acceptable based on the following criteria: root mean square error of approximation (RMSEA) < 0.05; comparative fit index ≥ 0.95, and the Tucker-Lewis index ≥ 0.95.

2. Standardized factor loadings between items and the composites to which they belong should be at least 0.40, and should not cross-load on other composites.

3. Observed unit-level reliability should be ≥ 0.70. It can be less than 0.70, but in that case sample size projections based on unit-level reliability should indicate that an effective sample size of 300 or less per reporting unit will be needed to obtain unit-level reliability ≥ 0.70.

4. Internal-consistency reliability (Cronbach's alpha) should be ≥ 0.70.

5. Scaling success should be 100%. Scaling success is an indicator of the extent to which items correlate more highly with their own composites than with competing composites.

AIR
AMERICAN INSTITUTES FOR RESEARCH®

6. Ceiling effects should be no higher than 75%, and preferably would be no higher than 50%. The ceiling effect indicates the proportion of respondents who give the highest possible rating for all items in a composite.

7. The items grouped together in a composite should make sense, substantively. This requirement is subjective rather than statistical, and relies on the mutual agreement of various stakeholders.

In some cases, where there may be a compelling reason to retain a composite (e.g., the composite measures a concept that is very important to consumers), some of these criteria may be relaxed in order to allow the retention of that composite. Alternatively, one or more items in such a composite may be retained as single item indicators.

6.1 Evaluating the Original Composite Design

We conducted a Confirmatory Factor Analysis (CFA) to determine whether the data fit our conceptual framework (shown in Exhibit 23) using structural equation modeling (SEM).[8,9] With large samples, such as that used here, even trivial departures from the specified model may be statistically rejected; therefore, it is customary to use practical fit indices to evaluate the hypothesized model. Specifically, the comparative fit index (CFI) and Taylor-Lewis Index (TLI), along with the standardized root mean square error of approximation (RMSEA), were used to evaluate fit.[10,11,12]

The CFI and TLI compare the fit of the specified model to that of a model that specifies no covariation (the null model). Both indexes run from a value of "0" (no relationship between the predicted and observed correlation matrix) to "1.0" (the predicted correlation matrix is identical to the observed). The TLI includes a greater correction for the number of parameters in the model (analogous to an adjusted R^2) than the CFI. The RMSEA is the amount of variance that is not predicted by the model and had associated confidence intervals (which the CFI and TLI do not). A CFI and TLI of less than 0.90 and an RMSEA greater than 0.10 indicate that the hypothesized model may not be the best description of the data. Excellent fit of the model to the data is considered if the CFI and TLI are equal to or greater than 0.95 and the RMSEA is equal to or less than 0.06, though that cutoff is sometimes seen as too high.[13,14]

---

[8] Hoyle RH, editor. Structural Equation Modeling. Thousand Oaks: SAGE Publications, Inc.; c1995.

[9] Keller SD, O'Malley AJ, Hays RD, Zaslavsky AM, Hepner KA, Clearly PD. Methods used to streamline the CAHPS® Hospital Survey. Journal of Health Services Research. 2005;40(6):2057-2077.

[10] Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling. 1999;6(1),1-55.

[11] Kenny DA. Measuring Model Fit [Internet]. David A. Kenny; 2014 January 5 [updated 2014 February 6; cited 2014 March 14]. Available from: http://davidakenny.net/cm/fit.htm.

[12] Suhr DD. Exploratory or confirmatory factor analysis? SUGI 31 Proceedings. 2006. Paper 200-31.

[13] Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling. 1999;6(1),1-55.

[14] Kenny DA. Measuring Model Fit [Internet]. David A. Kenny; 2014 January 5 [updated 2014 February 6; cited 2014 March 14]. Available from: http://davidakenny.net/cm/fit.htm.

Exhibit 23 displays the initial hypothesized factor structure.[15] The 29 substantive report items were hypothesized to belong to an eight-factor composite structure.

Exhibit 23. Initial hypothesized factor structure

Q#	Composites and Items 29 Items Mapped to 8 Composites
	Application Process (3 items)
3	Was it easy to give information about the people in your family, including yourself, who wanted health insurance?
4	Did giving information about the people in your family, including yourself, take longer than you expected?
17	Was it easy to understand how to update the {INSERT MARKETPLACE NAME} about changes to your household income or the number of people in your family?
	Premium Tax Credit Eligibility (3 items)
8	When you gave your household income information, was it easy to find out if you could get help paying for your health insurance?
9	Did giving your household income information take longer than you expected?
15	Was it easy to understand how to appeal the decision?
	Seeking Information on the Marketplace Website (4 items)
19	Since October 1st, how often did you have to wait to get what you needed because of problems on the {INSERT MARKETPLACE NAME}'s website?
20	Since October 1st, how often did you get the information you needed from the {INSERT MARKETPLACE NAME}'s website?
22	Since October 1st, how often was it easy to understand the information on the {INSERT MARKETPLACE NAME}'s website?
24	Since October 1st, how often was the information on the {INSERT MARKETPLACE NAME}'s website as helpful as you thought it should be?
	Seeking Information over the Phone (5 items)
27	Since October 1st, how often did you get the information or help you needed when you called the {INSERT MARKETPLACE NAME} customer service Help Line?
29	Since October 1st, how often was it easy to understand the information you got when you called the {INSERT MARKETPLACE NAME}'s Help Line?
31	Since October 1st, how often was the {INSERT MARKETPLACE NAME}'s Help Line staff as helpful as you thought they should be?
32	Since October 1st, how often did the {INSERT MARKETPLACE NAME}'s Help Line staff use words or phrases you did not understand when you called?
34	Since October 1st, how often did the {INSERT MARKETPLACE NAME}'s Help Line staff treat you with courtesy and respect when you called?
	Seeking Information In-Person (5 items)
38	Since October 1st, how often did you get the information or help you needed when you met in person with anyone about getting health insurance from the {INSERT MARKETPLACE NAME}?
40	Since October 1st, how often was it easy to understand the information you got when you met in person with anyone about getting health insurance from the {INSERT MARKETPLACE NAME}?
42	Since October 1st, how often were the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} as helpful as you thought they should be?
43	Since October 1st, how often did the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} use words or phrases you did not understand?
44	Since October 1st, how often did the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} treat you with courtesy and respect?
	Health Plan Enrollment Process (4 items)
49	Since October 1st, how often was it easy to understand the services covered by the health plans available to you and how much you would have to pay?
51	Since October 1st, how often was it easy to understand which health plans had the doctors or hospitals you wanted?

---

[15] Questions not evaluated as part of the factor structure include those designed to determine eligibility for response (i.e., screener questions); single-item questions, such as the global ratings; and "about you" questions included as potential case mix adjusters (e.g., age, sex, education, self-reported health ratings).

AIR ®
AMERICAN INSTITUTES FOR RESEARCH®

Q#	Composites and Items 29 Items Mapped to 8 Composites
53	Since October 1st, how often was it easy to understand which health plans covered the prescription medicines you needed?
59	Was it easy to choose a health plan?
	Specialized Services (2 items)
55	Since October 1st, was it easy to find out which health plans in the {INSERT MARKETPLACE NAME} offered the physical, occupational, or speech therapy services you needed
57	Since October 1st, was it easy to find out which health plans in the {INSERT MARKETPLACE NAME} offered home health care services you needed?
	Cultural Competence (3 items)
61	Since October 1st, when you needed an interpreter to help you speak with anyone about getting health insurance from the {INSERT MARKETPLACE NAME}, how often did you get one?
63	Since October 1st, how often were the forms that you had to fill out through the {INSERT MARKETPLACE NAME} available in the language you prefer?
65	Since October 1st, how often were the forms that you had to fill out available in the format you needed, such as large print or braille?

We attempted to run the initial CFA on the Marketplace Survey data using Mplus with the full dataset. We found that items 57, 15, 61 and 65 had covariance coverage values below 0.10 when paired with many other items. The covariance coverage value shows the proportion of cases that contribute a value to calculate the covariance between a set of two items. Mplus will not allow the model to run if any item pair had a value below 0.10 (10%), a cutoff point indicating that the coverage is too weak.[16] We removed these items from the model, as well as items 55 and 63, as they were the only items left in their composites (Specialized Services and Cultural Competence) and re-ran the model.

The model ran successfully, with excellent fit statistics (RSMEA=0.045, CFI = 0.973, TLI=0.968). However, the inter-factor correlation between the Application Process and Premium Tax Credit Eligibility factors was very high (correlation > 1.0), indicating a lack of discriminant validity. We combined the items from these two factors into a single new factor ("Application Process/Premium Tax Credit Eligibility") and re-ran the model. The fit statistics showed that this model was a good fit (see Exhibit 24), and all but one item (32) had factor loadings ≥ 0.40.

Exhibit 24. CFA fit statistics for five-factor structure

Fit statistic (target)	Value
Root Mean Square Error Of Approximation (RMSEA ≤ 0.05)	0.045
RMSEA 95% Confidence Interval (Upper Limit ≤ 0.05)	0.042 - 0.047
Probability RMSEA ≤ 0.05 (p = 1.00)	1.000
Comparative Fit Index (CFI ≥ 0.95)	0.973
Taylor-Lewis Index (TLI ≥ 0.95)	0.969

---

[16] Geiser, Christian. 2013. *Data Analysis with MPlus*. The Guilford Press. New York, NY.

AIR
AMERICAN INSTITUTES FOR RESEARCH

6.2   Exploring Alternate Composite Structures

6.2.1 CFAs Using Subsets of Respondents and Items

Factor analysis is generally used to develop composites when respondents answer most of the questions in a survey. The Marketplace Survey was designed to have three distinct sections each dedicated to asking about a specific mode of interacting with the Marketplace (Web, phone, and in-person). Respondents can skip out of entire sections associated with a single composite because they did not have relevant experience. For example, 64 percent of respondents reported visiting healthcare.gov (Web), 54 percent reported calling the customer service Help Line (phone), and 30 percent reported meeting in person with a person from an organization that helps people get health insurance through the Health Insurance Marketplace (in-person). Thirty-nine percent used two modes (in any combination), and only 10 percent reported using all three modes of application.

Because there was little overlap among respondents across these three modes, we split the full data set into three subsets of respondents: 1) those who answered survey items in the *Seeking Information on the Marketplace Website* composite, 2) those who answered survey items in the *Seeking Information Over the Phone* composite, and 3) those who answered survey items in the *Seeking Information In-Person* composite. We modified the hypothesized structure to only include the information seeking composite that matched the subsample's mode of interaction. For example, a CFA that included the Web composite along with Apply and HP Enroll, but not the phone or in-person composites, was run using only those respondents who reported visiting healthcare.gov. Likewise, a CFA that included the phone composite along with Apply and HP Enroll, but not the Web or in-person composites, was run using only those who used the customer service Help Line, and so on. Additional items had to be dropped from each of the models due to low covariance coverage. Results are shown in Exhibit 25.

Exhibit 25. CFA results for mode of information seeking models – mode subsamples

Fit statistic (target)	Phone Assistance Model	Web Assistance Model	In-Person Assistance Model	Five-Factor Structure
RMSEA (≤0.05)	0.094	0.101	0.061	0.045
RMSEA 95 CI (U95 ≤0.05)	0.089 – 0.100	0.095 – 0.107	0.063 – 0.068	0.042 - 0.047
Probability RMSEA ≤0.05 (1.00)	0.000	0.000	0.008	1.000
CFI (≥0.95)	0.963	0.943	0.976	0.973
TLI (≥0.95)	0.954	0.927	0.971	0.969

In general, the fit statistics for each of the three information seeking models were worse than the initial hypothesized composite structure. These models do not indicate an improved (or even equivalent fit) compared to the original structure. However, it was not clear from this finding alone whether subsetting the full data set or running separate models for each mode of application was the driving force behind the decline in model fit. In addition, this approach reduced the sample size, created covariance coverage problems, and required us to drop additional variables.

We thus re-ran the modified composite models using the full dataset so as to compare the original hypothesized composites to three new models without the confounding factor of subsetting the data file. Fit statistics, in general, were worse for each of the specific information seeking models, compared to the initial model, even when using the full dataset for both analyses; however, the fit was better than when using subsets of respondents for each model (see Exhibit 26).

Exhibit 26. CFA results for mode of information seeking models – full sample

Fit statistic (target)	Phone Assistance Model	Web Assistance Model	In-Person Assistance Model	Five-Factor Structure
RMSEA (≤0.05)	0.055	0.063	0.051	0.045
RMSEA 95 CI (U95 ≤0.05)	0.051 - 0.060	0.059 - 0.068	0.047 - 0.056	0.042 - 0.047
Probability RMSEA ≤0.05 (1.00)	0.030	0.000	0.269	1.000
CFI (≥0.95)	0.984	0.974	0.976	0.973
TLI (≥0.95)	0.980	0.967	0.970	0.969

6.2.3 Exploratory Factor Analysis by Mode of Information Seeking

We conducted an exploratory factor analysis (EFA) on the full dataset to see if there was evidence supporting alternative factor structures. EFAs were conducted on both subsets of items and on subsets of respondents. The number of factors was determined by the eigenvalues, interpretability of the rotated factor pattern matrix and scree plot analysis.[17] None of the EFA models showed a better fit than our initial CFA.

Given these results, we established a tentative revised factor structure shown in Exhibit 27.[18] This structure was used in subsequent analyses described below, including:

1. Invariance testing by mode and language (Section 6.2.4)

2. Evaluation of a possible health insurance literacy composite (Section 6.2.5)

3. Evaluation of the addition of item 63 ("How often were the forms that you had to fill out through the Health Insurance Marketplace available in the language you prefer?") to the item pool (Section 6.2.6)

4. Evaluation of reliability, variability, and validity of the composites (Section 6.3)

Exhibit 27. Tentative revised factor structure

Q# Field Test[†]	Composites and Items 23 Items Mapped to 5 composites
	Application and Premium Tax Credit Eligibility Process (5 items) - Apply
3	Was it easy to give or update information about yourself or the people in your family who wanted health insurance?
4	Did giving or updating information about yourself or the people in your family take longer than you expected?

---

[17] Muthén, L.K. and Muthén, B.O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén

[18] Note: this tentative revised structure is the same as the five-factor structure shown in earlier sections. It has merely been upgraded based on the additional CFAs and EFAs.

Q# Field Test[†]	Composites and Items 23 Items Mapped to 5 composites
8	When you gave or updated your household income information, was it easy to find out if you or the people in your family could get help paying for health insurance?
9	Did giving or updating your household income information take longer than you expected?
17	Was it easy to understand how to update {INSERT MARKETPLACE NAME} about changes to your household income or the number of people in your family?
Seeking Information on the Marketplace Website (4 items) - Web	
19	Since November 15th, how often did you have to wait to get what you needed because of problems on {INSERT MARKETPLACE NAME}'s website?
20	Since November 15th, how often did you get the information you needed from {INSERT MARKETPLACE NAME}'s website?
22	Since November 15th, how often was it easy to understand the information on {INSERT MARKETPLACE NAME}'s website?
24	Since November 15th, how often was the information on {INSERT MARKETPLACE NAME}'s website as helpful as you thought it should be?
Seeking Information over the Phone (5 items) - Phone	
27	Since November 15th, how often did you get the information or help you needed when you called {INSERT MARKETPLACE NAME}'s customer service Call Center?
29	Since November 15th, how often was it easy to understand the information you got when you called {INSERT MARKETPLACE NAME}'s customer service Call Center?
31	Since November 15th, how often was {INSERT MARKETPLACE NAME}'s customer service Call Center as helpful as you thought it should be?
32	Since October 1st, how often did the {INSERT MARKETPLACE NAME}'s Help Line staff use words or phrases you did not understand when you called?
34	Since November 15th, how often did {INSERT MARKETPLACE NAME}'s customer service Call Center staff treat you with courtesy and respect when you called?
Seeking Information In-Person (5 items) – In Person	
38	Since November 15th, how often did you get the information or help you needed when you met in person with someone about getting health insurance from {INSERT MARKETPLACE NAME}?
40	Since November 15th, how often was it easy to understand the information you got when you met in person with someone about getting health insurance from {INSERT MARKETPLACE NAME}?
42	Since November 15th, how often were the persons you met with about getting health insurance from {INSERT MARKETPLACE NAME} as helpful as you thought they should be?
43	Since October 1st, how often did the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} use words or phrases you did not understand?
44	Since November 15th, how often did the persons you met with about getting health insurance from {INSERT MARKETPLACE NAME} treat you with courtesy and respect?
Health Plan Enrollment Process (4 items) – HP Enroll	
49	Since November 15th, how often was it easy to understand the services covered by the health plans available to you and how much you would have to pay?
51	Since November 15th, how often was it easy to understand which health plans had the doctors or hospitals you wanted?
53	Since November 15th, how often was it easy to understand which health plans covered the prescription medicines you needed?
59	Since November 15th, was it easy to choose a health plan?

6.2.4 Invariance Testing by Mode and Language of Survey Administration

The CFA in section 6.2.1 shows that the initial set hypothesized composites fits the combined dataset well. However, one of our analysis goals was to test the invariance of the factor structure – that is, whether or not the structure of the factor model is the same across various languages (English, Spanish, and Chinese) and modes (Mail, Phone and Web). We want the initial hypothesized composite structure found in the full dataset to be the same across all languages

AIR ®
AMERICAN INSTITUTES FOR RESEARCH®

and modes. If our factor structure is non-invariant then it means our items may be measuring different things across language or mode. This could suggest potential translation issues. We might not want to combine all of the data for analyses if our factor structure was not invariant across mode and language.

There are different levels of measurement invariance to test, which range from least restrictive to most restrictive:

1. Configural – restricts the factor structures to be the same, but allows factor loadings and intercepts to be different
2. Metric– restricts factor loadings to be the same but allows intercepts to differ
3. Scalar – restricts both factor loadings and intercepts to be the same

Invariance needs to be established at the configural level before moving on to the metric level, then established at the metric level before moving on to scalar. We divided the dataset into three groups based on survey language and ran a CFA to test the tentative revised structure within each group. We followed the same protocol for the three modes of survey completion groups. In both cases, the reduction in the number of respondents due to splitting the full sample resulted covariance coverage problems and some items that were supposed to be included in each model had to be dropped. Invariance testing at the configural level involves restricting the overall fit to be equal across groups (language or mode). The fit statistics indicate how good the fit is given such restriction.

Fit statistics for the for the mail-only dataset used to test invariance by language appear in Exhibit 28, which shows a good fitting model although the RMSEA was slightly higher than the cutoff point.

Exhibit 28. CFA results by language of survey completion

Fit statistic (target)	Mail Dataset	Five-Factor Structure
RMSEA (≤0.05)	0.056	0.045
RMSEA 95 CI (U95 ≤0.05)	0.052 - 0.061	0.042 - 0.047
Probability RMSEA ≤0.05 (1.00)	0.008	1.000
CFI (≥0.95)	0.977	0.973
TLI (≥0.95)	0.972	0.969

Fit statistics for the English-only dataset used to test invariance by mode appear in Exhibit 29 and show a good fitting model although the RMSEA was slightly higher than the cutoff point.

Exhibit 29. CFA results by mode of survey completion

Fit statistic (target)	English Dataset	Five-Factor Structure
RMSEA (≤0.05)	0.056	0.045
RMSEA 95 CI (U95 ≤0.05)	0.050 - 0.061	0.042 - 0.047
Probability RMSEA ≤0.05 (1.00)	0.040	1.000
CFI (≥0.95)	0.984	0.973

AIR AMERICAN INSTITUTES FOR RESEARCH

Fit statistic (target)	English Dataset	Five-Factor Structure
TLI (≥0.95)	0.980	0.969

In our full CFA on the tentative revised factor structure, we found a good fit of the factor structure to the data. When the revised factor structure was compared across language and mode, the goodness of fit varied. Our weakest evidence in favor of configural invariance was for Chinese language group (532 completed surveys) and the Web mode group (120 completed surveys). The Web mode was one group for which we did not obtain an adequate number of useable survey responses (as specified in our sampling design, which indicated a minimum of 450 completed surveys per group). In addition, both the phone group (259 completed surveys) and the Spanish language group (372 completed surveys) fell short of our target as well.

Furthermore, similar to what we encountered with the original CFAs, several items had to be dropped out of the configural models due to the low number of usable responses to items that occurs when due to low screen-in rates in combination with subsetting the full data set. Moreover, the item sets that had to be dropped varied across both mode and language, and thus the models were not really comparable across these groupings. Given these limitations, we cannot make final decisions about language or mode invariance at this time, due to sample size issues. The beta test sampling design has been revised to take into consideration both sample-level and item-level completions rates so as to obtain the required number of completed surveys and items to allow us to conduct invariance testing with greater confidence in the results. Thus, invariance testing will be revisited using beta test data.

For our purposes, meeting the first level of invariance may be enough. Composite scoring using the CAHPS macro does not take factor weights into consideration, so invariant factor loadings and intercepts are less important in that context. Furthermore, since composite scoring can be done using full weights (sampling weights in combination with non-response weights adjusted for non-response bias) and case mix adjusters (including language and mode), having evidence that the factor *structure* is invariant across language and mode (as opposed to loadings and intercepts) may be sufficient.

6.2.5 Testing a Health Insurance Literacy Composite

Based on interest in the October 2014 TEP meeting, we modified the original hypothesized composite model to create a new hypothesized composite intended to measure consumers' health insurance literacy. This factor included item 92 ("How confident are you that you understand health insurance terms?") as well as items 32 and 43, which both ask the respondent about the use of unfamiliar terms or phrases when receiving help on the phone or in person.

We ran a CFA, which showed adequate fit statistics shown in Exhibit 30. In addition, all items from the Health Insurance Literacy Factor had factor loadings greater than 0.40.

Exhibit 30. CFA fit statistics for model with health insurance literacy factor

Fit statistic (target)	Health Insurance Literacy Factor Structure
RMSEA (≤0.05)	0.040
RMSEA 95 CI (U95 ≤0.05)	0.038 0.042
Probability RMSEA ≤0.05 (1.00)	1.000
CFI (≥0.95)	0.977

Fit statistic (target)	Health Insurance Literacy Factor Structure
TLI (≥0.95)	0.973

However, the Cronbach's Alpha for the Health Insurance Literacy composite of 0.44 is below the standard of 0.70, which indicates that the items have low internal consistency and should not be grouped together into a composite.

Next, we ran an EFA model using all items from the interim revised factor structure plus item 92 to see if: 1) items 92, 32, and 43 would load on the same factor, and 2) what other items would load with them. Based on scree plot analysis, we found that 4, 5 or 6 factor solutions were possible fits for the data. All models had adequate fit statistics, as summarized in Exhibit 31. However, in all the models, items 92, 32, and 43 never loaded onto the same factor, confirming that the three items should not be grouped together into a composite.

Exhibit 31. Health insurance literacy EFA fit statistics

Fit statistic (target)	Health Insurance Literacy 4-Factor Model	Health Insurance Literacy 5-Factor Model	Health Insurance Literacy 6-Factor Model	Five-Factor Structure
RMSEA (≤0.05)	0.041	0.034	0.026	0.045
RMSEA 95 CI (U95 ≤0.05)	0.039 - 0.044	0.031 - 0.037	0.023 - 0.029	0.042 - 0.047
Probability RMSEA ≤0.05 (1.00)	1.000	1.000	1.000	1.000
CFI (≥0.95)	0.981	0.988	0.994	0.973
TLI (≥0.95)	0.971	0.980	0.989	0.969

During an internal team meeting, senior level advisors hypothesized that the Health Insurance Literacy composite may fit better if we test the hypothesized HIL composite on three datasets divided by language of survey completion. We started by running a CFA using the three-item HIL composite structure described above on just the English survey data. Exhibit 32 details the fit statistics, which were all adequate.

Exhibit 32. CFA fit statistics for model with health insurance literacy factor, English only

Fit statistic (target)	Health Insurance Literacy Factor Structure, English Dataset	Five-Factor Structure
RMSEA (≤0.05)	0.039	0.045
RMSEA 95 CI (U95 ≤0.05)	0.036 0.043	0.042 - 0.047
Probability RMSEA ≤0.05 (1.00)	1.000	1.000
CFI (≥0.95)	0.980	0.973
TLI (≥0.95)	0.977	0.969

Similar to the original model, the fit statistics indicated that the model was a good fit among English respondents; however, the Cronbach's Alpha for this composite was 0.46, which indicates low internal consistency. Instead of continuing to run additional CFAs for the Spanish and Chinese datasets, we ran the Alpha levels for the HIL items for each language. The alpha level for Spanish was 0.38 and the alpha level for Chinese could not be calculated. These results indicate that the scales in all language datasets have low internal consistency. We concluded that the data does not support the measurement of a Health Insurance Literacy construct.

6.2.6 Adding a Cultural Competence Item

Several members at the October 2014 TEP meeting expressed interest in retaining item 63 (how often were the forms that you had to fill out through the Health Insurance Marketplace available in the language you prefer?) from the *Cultural Competence* composite. This item was dropped because all other items in the *Cultural Competence* factor had low covariance coverage issues, even though item 63 did not have this problem. Members of the TEP were interested to see if it would load onto other factors, such as the *Seeking Information In-Person* and *Seeking Information over the Phone* composites, under the assumption that it might be easier to obtain forms in another language when using the help line or in-person assistance.

We first conducted an EFA to evaluate any potential changes to the underlying factor structure with this item added to the pool of items from the tentative revised set shown in Exhibit 27. The results suggested a range of structures with four, five, or six factors. In general, fit statistics for all three alternative structures were good (see Exhibit 33).

AIR ®
AMERICAN INSTITUTES FOR RESEARCH

Exhibit 33. Item 63 EFA fit statistics

Fit statistic (target)	Item 63 4-Factor Model	Item 63 5-Factor Model	Item 63 6-Factor Model	Five-Factor Structure
RMSEA (≤0.05)	0.041	0.031	0.021	0.045
RMSEA 95 CI (U95 ≤0.05)	0.039 - 0.044	0.028 - 0.034	0.018 - 0.025	0.042 - 0.047
Probability RMSEA ≤0.05 (1.00)	1.000	1.000	1.000	1.000
CFI (≥0.95)	0.981	0.990	0.996	0.973
TLI (≥0.95)	0.972	0.984	0.992	0.969

Looking across the factor structures, we found that that item 63 loads onto factors that include items from the *Seeking Information In-Person* and *Seeking Information over the Phone* composites. This knowledge, combined with the knowledge that individuals most likely had more access to forms in their preferred language when they sought help in-person led us to hypothesize that item 63 should be in the *Seeking Information In Person* composite. We ran a CFA on this model and found a good fit (Exhibit 34).

Exhibit 34. CFA with Q63 added to seeking information in person

Fit statistic (target)	Item 63 Model	Original CFA Model
RMSEA (≤0.05)	0.044	0.045
RMSEA 95 CI (U95 ≤0.05)	0.042 0.047	0.042 - 0.047
Probability RMSEA ≤0.05 (1.00)	1.000	1.000
CFI (≥0.95)	0.971	0.973
TLI (≥0.95)	0.967	0.969

Cronbach's Alpha for this composite was 0.79, which indicates that the items have a good internal consistency. This indicates that we may want to add item 63 to the *Seeking Information In-Person* composite. A question remains on whether or not it makes sense substantively to group this item in this composite, which deals with in-person help. Forms were available in other languages via the web, so this was not exclusive to the in-person help mode. This item will be retained on the beta test version of the survey and we will reevaluate with the beta test data.

6.3 Evaluation of Reliability, Variability, and Validity of Composite Scores

This section reports the results from further evaluations of the measurement properties of the proposed composites. Each is explained in more detail below.

6.3.1 Unit-Level Reliability

Unit-level reliability indicates the extent to which the respondents within a unit of interest (e.g., health plans, clinician groups, or hospitals) agree with one another in terms of their reported experiences within that unit compared to the amount that reported experiences differs among units. As such, it reflects the ratio of the within-unit variation in scores (agreement among respondents within a unit) to the between-unit variation in scores (variation in scores across units). One of the primary purposes of this survey is to be able to detect difference among states, and thus, this ratio is a good indicator of the extent to which the composites and other survey items accomplish this goal. It also tells us how reliable a measure is across different respondents. Inter-rater reliability is implied here in the sense that if each respondent is considered a "rater,"

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

then a measure with good unit-level reliability will yield consistent results across respondents, or raters, who have interacted with the same state Marketplace.

In CAHPS, there are two statistics used to assess this reliability.[19] One is a measure of inter-unit reliability (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.[20] The other measure is the intra-class correlation (ICC), which is also calculated using statistics produced by an ANOVA. The ICC used in calculating unit-level reliability is calculated as the between-unit variance minus the within-unit variance over the total variance adjusted for the average number of respondents per reporting unit.[21] 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. Since unit-level reliability is partly a function of sample size, both the IUR and the ICC allow estimation of the reliability associated with a particular number of respondents as well as making it possible to calculate the number of respondents needed per reporting unit to obtain a particular level of reliability. 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.

When calculated as described above, the IUR and ICC pool information across all units into a single survey-wide scalar summary for each item or composite. However, if the number of respondents varies across units (as they do for this survey), units will have different sampling variances. Thus, an alternative is to estimate reliability for *each* unit. This approach uses standard CAHPS macro outputs to calculate the statistics needed to calculate reliability.

If the unit mean, standard error, and total number of responders to an item or composite for unit i = 1, …, D are m_i, s_i, and n_i respectively, these statistics include:

1. The total number of respondents across all units (N)

2. The overall mean rating ($\overline{m}$)

3. The sample variance estimate for each unit (v_i)

4. The within-unit variance ($\overline{v}$)

5. The between-unit variance (b)[22]

6. Reliability for a specific unit (R_i ¿

---

[19] 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 from: https://cahps.ahrq.gov/surveys-guidance/docs/2015_instructions_for_analyzing_data.pdf . Much of the text in this section is based on information provided in that document.

[20] Winer BJ. Statistical principles in experimental design. New York: McGraw-Hill; c1970.
Zaslavsky AM, Buntin MJB. Using survey measures to assess risk selection among Medicare Managed care plans. Inquiry. 2002;39(2):138-151.

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

[22] The equations for calculating b and $\overline{v}$ can be found on 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: https://cahps.ahrq.gov/surveys-uidance/docs/2015_instructions_for_analyzing_data.pdf

7. The projected reliability for a future survey with r respondents per unit (R)

8. The projected number of respondents per unit needed to obtain a reliability of R (usually $R = 0.70$).

We can calculate the reliability for a specific unit (#6) as:

$$R_i = \frac{b}{b + s_i^2}$$

For a future survey with r respondents per unit (#7), the projected reliability is calculated as:

$$R = \frac{b}{b + \frac{\bar{v}}{r}}$$

Finally, to project the number of respondents per unit[23] needed to obtain a desired reliability (#8), the formula for #7 can be rewritten to solve for r, resulting in:

$$r = \frac{R\bar{v}}{b - Rb}$$

Scales with reliability coefficients above 0.70 provide adequate precision for use in statistical analysis of unit-level comparisons[24] though it has been argued that IURs should be at least 0.90.[25]

Exhibits 35a to 35e display the IUR analysis results respectively for each composite along with the items that belong to the composite. As mentioned above, the observed reliability is partly a function of sample size, or more specifically, the number of respondents used in the analysis (taking into consideration the item-level or composite-level response rate). Therefore, statistic of primary interest that is produced with this analysis is the average number of *usable* responses to a particular item or composite per state needed (the 'effective sample size') to obtain a target IUR. Tables 35a to 35e display this number at the composite and item level for a target IUR of 0.70, which is the minimum acceptable level of unit reliability. As shown in the Exhibits, at the composite level, this number ranges from a low of 40 per state for the Apply composite (Exhibit 35a) to a high of 88 per state for the HP Enroll composite (Exhibit 35e). For individual items, the effective sample size ranges from a low of 40 per state for Q42 (Exhibit 35d) to a high of 186 per state for Q53 (Exhibit 35d).

Exhibit 35a. IUR results for application and subsidy eligibility process

Q#	Item Text	ICC	Observed State R	N	Average N per State	N needed for State R >= 0.70	Item RR	Effective Sample Size Needed per State for R >=0.70
Apply	Application and Subsidy Eligibility	0.0581	0.80	2,281	63	38	95%	40

---

[23] Not all respondents are eligible to answer all report items in CAHPS surveys, and thus, the estimate of r must be calculated separately for each composite, global rating, and any single-item measures so as to take into consideration the impact of item nonresponse.

[24] Nunnally JC. Psychometric theory. 2nd ed. New York: McGraw-Hill Book Company; c1978.

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

Q#	Item Text	ICC	Observed State R	N	Average N per State	N needed for State R >= 0.70	Item RR	Effective Sample Size Needed per State for R >=0.70
	Process							
rq03	Was it easy to give information about the people in your family, including yourself, who wanted health insurance?	0.0329	0.66	2,050	57	69	85%	81
q04	Did giving information about the people in your family, including yourself, take longer than you expected?	0.0222	0.55	1,974	55	103	82%	126
rq17	Was it easy to understand how to update the Marketplace about changes to your household income or the number of people in your family?	0.0953	0.76	1,112	31	22	46%	48
rq08	When you gave your household income information, was it easy to find out if you could get help paying for your health insurance?	0.0599	0.78	2,032	56	37	84%	43
q09	Did giving your household income information take longer than you expected?	0.0659	0.79	1,964	55	33	81%	41

ICC = intra class correlation; R = state reliability; N = number of responses used; RR = response rate

Exhibit 35b. IUR results for seeking information on the marketplace website

Q#	Item Text	ICC	Observed State R	N	Average N per State	N needed for State R >= 0.70	Item RR	Effective Sample Size Needed per State for R >=0.70
Infoweb	**Seeking Information on the Marketplace Website**	**0.0454**	**0.67**	**1,511**	**42**	**49**	**63%**	**78**
rq19	How often did you have to wait to get what you needed because of problems on the Marketplace's website?	0.0392	0.63	1,492	41	57	62%	93
q20	How often did you get the information you needed from the Marketplace's website?	0.0333	0.59	1,499	42	68	62%	109
q22	How often was it easy to understand the information on the Marketplace's website?	0.0228	0.49	1,491	41	100	62%	162
q24	How often was the information on the Marketplace's website as helpful as you thought it should be?	0.0569	0.71	1,482	41	39	61%	63

ICC = intra class correlation; R = state reliability; N = number of responses used; RR = response rate

Exhibit 35c. IUR results for seeking information on the phone

Q#	Item Text	ICC	Observed State R	N	Average N per State	N needed for State R >= 0.70	Item RR	Effective Sample Size Needed per State for R >=0.70
InfoPhon	**Seeking Information on the Phone**	**0.0600**	**0.69**	**1,277**	**35**	**37**	**53%**	**69**
q27	How often did you get the information or help you needed when you called the Marketplace customer service Help Line?	0.0626	0.70	1,257	35	35	52%	67
q29	How often was it easy to understand the information you got when you called the Marketplace's Help Line?	0.0738	0.73	1,248	35	29	52%	57
q31	How often was the Marketplace's Help Line staff as helpful as you thought they should be?	0.0654	0.71	1,249	35	33	52%	64
rq32	How often did the	0.0520	0.65	1,230	34	43	51%	83

AIR® AMERICAN INSTITUTES FOR RESEARCH®

Q#	Item Text	ICC	Observed State R	N	Average N per State	N needed for State R >= 0.70	Item RR	Effective Sample Size Needed per State for R >=0.70
	Marketplace's Help Line staff use words or phrases you did not understand when you called?							
q34	How often did the Marketplace's Help Line staff treat you with courtesy and respect when you called?	0.0768	0.72	1,130	31	28	47%	60
InfoPhon	**Seeking Information on the Phone (w/o q32)**	**0.0756**	**0.74**	**1,277**	**35**	**29**	**53%**	**54**

ICC = intra class correlation; R = state reliability; N = number of responses used; RR = response rate

Exhibit 35d. IUR results for seeking information in-person

Q#	Item Text	ICC	Observed State R	N	Average N per State	N needed for State R >= 0.70	Item RR	Effective Sample Size Needed per State for R >=0.70
InfoPers	**Seeking Information In-Person**	**0.1502**	**0.77**	**671**	**19**	**13**	**28%**	**47**
q38	How often did you get the information or help you needed when you met in person with anyone about getting health insurance from the Marketplace?	0.1118	0.70	654	18	19	27%	68
q40	How often was it easy to understand the information you got when you met in person with anyone about getting health insurance from the Marketplace?	0.1369	0.74	652	18	15	27%	54
q42	How often were the persons you met with about getting health insurance from the Marketplace as helpful as you thought they should be?	0.1775	0.80	648	18	11	27%	40
rq43	How often did the persons you met with about getting health insurance from the Marketplace use words or phrases you did not understand?	0.1638	0.78	650	18	12	27%	44
q44	How often did the persons you met with about getting health insurance from the Marketplace treat you with courtesy and respect?	0.0629	0.55	656	18	35	27%	128
InfoPers	**Seeking Information In-Person (w/o Q43**	**0.1515**	**0.77**	**670**	**19**	**13**	**28%**	**47**

ICC = intra class correlation; R = state reliability; N = number of responses used; RR = response rate

Exhibit 35e. IUR results for enrollment process

Q#	Item Text	ICC	Observed State R	N	Average N per State	N needed for State R >= 0.70	Item RR	Effective Sample Size Needed per State for R >=0.70
Enroll	**Enrollment Process**	**0.0303**	**0.64**	**2,054**	**57**	**75**	**85%**	**88**
q49	How often was it easy to understand the services covered by the health plans available to you and how much you would have to pay?	0.0449	0.70	1,757	49	50	73%	68
q51	How often was it easy to understand which health plans had the doctors or hospitals you wanted?	0.0461	0.61	1,185	33	48	49%	98
q53	How often was it easy to understand which health plans covered the prescription medicines you needed?	0.0370	0.46	788	22	61	33%	186
rq59	Was it easy to choose a health plan?	0.0301	0.55	1,429	40	75	59%	127
HP Enroll	**Enrollment Process (w/o q53)**	**0.0427**	**0.72**	**2,039**	**57**	**52**	**85%**	**62**

ICC = intra class correlation; R = state reliability; N = number of responses used; RR = response rate

Although not shown in these Exhibits, we also calculated, for each item and composite, the average effective sample size needed to obtain and IUR of 0.90. This number ranged from 154 for the Apply composite to a high of 338 for the HP Enroll composite. At the item level, this number ranged from a low of 155 for Q42 to a high of 717 for Q53.

Given that we plan to obtain 1,200 completes per state in the beta test phase of data collection, we should have enough usable responses to obtain state reliabilities of at least 0.70 for all items and composites.

6.3.2 Multitrait Analysis

The multitrait analysis approach extends the logic of multitrait-multimethod analysis to scale construction and validation when only one method of evaluation is used, as is the case with the single instrument approach taken here.[26] This technique is used to calculate Cronbach's Alpha, which is an indicator of internal consistency reliability, the scaling success of the proposed

---

[26] Hays RD & Hayashi T. (1990). Beyond internal consistency reliability: Rationale and user's guide for multitrait scaling analysis program on the microcomputer. *Behavior Research Methods, Instruments, and Computers, 22(2)*, 167-175.

composites, which is an indicator of discriminant validity (i.e., the degree to which items correlate more highly with their own composites than with competing composites), and the floor/ceiling effects of the composites.

Internal Consistency Reliability. Internal consistency reliability is a traditional method used to evaluate the amount of systematic variance among the items in a composite. Internal consistency reliability may be thought of as an estimate of repeatability in that each item in the composite measures the same thing, and so the strength of the relationship among items provides an estimation of the repeatability (reliability) of measurement for that composite.

Variability. Lack of variability in scores can attenuate validity coefficients and reduce the amount of information provided by the survey. To evaluate variability, AIR will examine the distributional properties of item and composite scores. For a composite, a respondent who gives the highest possible score for each item comprising the composite is said to be at the ceiling while a respondent who gives the lowest possible score for each item comprising the composite is at the floor. AIR will calculate, for each composite, the percentage of respondents with the highest (ceiling effect) and lowest (floor effect) possible scores. Ceiling and floor effects indicate the percentage of people for whom it would be impossible to assess improvement or decrement, respectively, over time. Composites or single items with high ceiling and/or floor effects should be considered for modification or deletion.

Validity. We consider three types of validity in evaluating survey measures: content, construct, and criterion validity. Content validity (establishing that the questions are representative of the concepts they are supposed to reflect) was established via the formative research—review of existing instruments, focus groups with people similar to those who would be responding to the survey, public comment in response to two Federal Register Notices, input from the TEP and other stakeholders, and the cognitive testing with people similar to those who would complete the survey.

Construct validity was assessed using factor analysis (see section 6.2); it was further evaluated using multi-trait analysis.[27] Valid constructs should have statistically significant loadings of a moderate to large magnitude (e.g., loadings ≥0.40). The tentative revised factor structure (see Exhibit 27) appears to have good construct validity as all items assessed meet this criteria, with the exception of one report item (q32).

In contrast, the multi-trait analysis compares the correlations of items with their composite total (correcting for overlap[28]) to the correlations of those items with competing composites. The scaling success statistic (an indicator of discriminant validity) is one of a number of pieces of evidence that bears on the construct validity of the proposed composites—scaling success of 100 percent indicates that all items correlate more highly (at least 1 standard error higher) with their own composites than with competing composites.[29]

---

[27] Hays RD, Hayashi T. Beyond internal consistency reliability: rationale and user's guide for multitrait scaling analysis program on the microcomputer. Behav Res Meth Ins C. 1990;22:1678-175.

[28] 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.

[29] Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull. 1959;56:81-105.

6.3.5 Results of Multitrait Analysis

The multitrait analysis was conducted using the Multitrait Analysis Program (MAP) version 2.0. The results are shown in Exhibits 36a to 36e. The MAP analysis was run using the default imputation approach in which respondents were retained in the analysis if they answer at least one item for every composite in the analysis. Typically, the MAP analysis would be run with all composites (or scales) included in a single model. However, since only a portion of respondents used any given mode for seeking information (web, phone, or in-person), and a very small number of respondents used all three modes, only 198 (8%) respondents can be retained in the multitrait analysis if all five composites were analyzed in a single MAP run. This lack of coverage might negatively impact the generalizability of the MAP analysis results to the entire population.

In order to mitigate this problem, the MAP analysis was conducted on five separate subsets of respondents formed by creating five unique combinations of the five composites. Each composite was thus evaluated in combination with either all other composites (Model 1) or in combination with two or three other composites (Models 2 to 5). The Enrollment composite and the Application composites had the largest number of useable responses, followed by the Web, Phone, and In-person composites respectively. Thus, each of the information composites was evaluated with either: 1) all other composites (Model 1); or 2) the Apply and Enroll composites only (Models 2 to 4). In addition, the Web and Phone composites were evaluated together with the Apply and Enroll composites (Model 5).

The models included:

- Model 1: multitrait analysis across all five domains (N=198)
    - Apply (application and subsidy eligibility process)
    - Web (seeking information on the Marketplace website)
    - Phone (seeking information over the phone)
    - In Person (seeking information in person)
    - Enroll (enrollment process)
- Model 2: multitrait analysis with (N=1,336)
    - Web
    - Apply
    - Enroll
- Model 3: multitrait analysis across three domains (N=1,122)
    - Phone
    - Apply
    - Enroll
- Model 4: multitrait analysis across three domains (N=580)
    - In person
    - Apply
    - Enroll
- Model 5: multitrait analysis across four domains (N=830)
    - Apply
    - Web

AIR
AMERICAN INSTITUTES FOR RESEARCH®

- o Phone
- o Enroll

Five Exhibits (36a to 36e) show the results for the Apply, Web, Phone, In-person, and Enroll composites, respectively, and each Exhibit includes all combinations in which the evaluated composite appears.

Exhibit 36a. Multitrait analysis results for application and subsidy eligibility process

Model	Subjects Included	Subjects Omitted	Mean	StDev	Range	Scaling success	Floor / Ceiling	Alpha
Apply with Web, Phone, In Person and Enroll	198	2,215 (92%)	11.27	2.91	5 to 15	100%	6% / 16%	0.85
Apply with Web and Enroll	1,336	1,077 (45%)	11.12	2.93	5 to 15	100%	5% / 17%	0.84
Apply with Phone and Enroll	1,122	1,291 (54%)	11.17	2.87	5 to 15	100%	5% / 15%	0.83
Apply with In Person and Enroll	580	1,833 (76%)	11.87	2.66	5 to 15	100%	3% / 22%	0.83
Apply with Web, Phone, and Enroll	830	1,583 (66%)	10.93	2.94	5 to 15	100%	5% / 13%	0.83

Exhibit 36b. Multitrait analysis results for seeking information on the web

Model	Subjects Included	Subjects Omitted	Mean	StDev	Range	Scaling success	Floor / Ceiling	Alpha
Web with Apply, Phone, In Person, and Enroll	198	2,215 (92%)	9.39	3.13	4 to 16	100%	6% / 7%	0.81
Web with Apply and Enroll	1,336	1,077 (45%)	9.77	3.11	4 to 16	100%	4% / 6%	0.81
Web with Apply, Phone, and Enroll	830	1,583 (66%)	9.25	2.99	4 to 16	100%	6% / 4%	0.79

Exhibit 36c. Multitrait analysis results for seeking information on the phone

Model	Subjects Included	Subjects Omitted	Mean	StDev	Range	Scaling success	Floor / Ceiling	Alpha
Phone with Apply, Web, In Person, and Enroll	198	2,215 (92%)	14.68	4.02	5 to 20	90%	2% / 19%	0.87
Phone with Apply and Enroll	1,122	1,291 (54%)	14.94	3.76	5 to 20	100%	1% / 18%	0.83
Phone with Apply, Web, and Enroll	830	1,583 (66%)	14.86	3.79	5 to 20	100%	1% / 18%	0.84

Exhibit 36d. Multitrait analysis results for seeking information in-person

Model	Subjects Included	Subjects Omitted	Mean	StDev	Range	Scaling success	Floor / Ceiling	Alpha
In Person with Apply, Web, Phone, and Enroll	198	2,215 (92%)	16.85	3.43	6 to 20	100%	0% / 36%	0.82
In Person with Apply and Enroll	580	1,833 (76%)	16.96	3.24	5 to 20	100%	0% / 35%	0.81

Exhibit 36e. Multitrait analysis results for enrollment process

Model	Subjects Included	Subjects Omitted	Mean	StDev	Range	Scaling success	Floor / Ceiling	Alpha
Enroll with Apply, Web, Phone, and In Person	198	2,215 (92%)	10.14	3.18	4 to 15	100%	7% / 11%	0.91
Enroll with Apply and Web	1,336	1,077 (45%)	9.69	3.15	4 to 15	100%	8% / 8%	0.92
Enroll with Apply and Phone	1,122	1,291 (54%)	9.79	3.15	4 to 15	100%	9% / 8%	0.92
Enroll with Apply and In Person	580	1,833 (76%)	10.37	3.04	4 to 15	100%	5% / 11%	0.91

Model	Subjects Included	Subjects Omitted	Mean	StDev	Range	Scaling success	Floor / Ceiling	Alpha
Enroll with Apply, Web, and Phone	830	1,583 (66%)	9.68	3.13	4 to 15	100%	8% / 7%	0.91

Results were fairly consistent across the five models. All five composites have acceptable reliability and validity, with the following cautions:

- In model 1 where all five composites were included in the analysis, the scaling success was less than 100 percent for the Phone composite (one item in Phone, Q32, has a slightly higher correlation with the In Person and Enroll composites, which results in the scaling success of 90 percent for Phone).

- The ceiling effect for the In Person composite was a little high at 35-36 percent; however, it is still possible to assess improvement over time for about 65 percent of the respondents, so modification to items in this composite might not be needed.

6.3.6 Criterion Validity

Criterion validity refers to the extent to which the Marketplace Survey composites agrees with some criterion of the "true" value of the measure, and can be predictive or concurrent. To evaluate the latter, we estimated correlation coefficients between each overall rating and the each composite. If the composites have good concurrent validity, then we would expect that those composites that tap into experiences that are conceptually related to a given overall rating would have a moderate to strong correlation ($r > 0.30$). For example, we would expect that the overall rating of the Website (q25) would be strongly correlated with the *Seeking Information on the Website* composite; however, that composite may not correlate as strongly with the overall rating of in-person assistance. Correlations of at least 0.40 indicate good criterion validity between ratings and composites or items that are related to similar topics.

In addition, the 'Apply' composite, which taps into the application and eligibility determination aspects of Marketplace experiences, had its strongest correlation with the overall rating of the Website. The 'HP Enroll' composite, which taps into the actual health plan selection aspect of Marketplace experiences, had its strongest correlation with the overall rating of the Marketplace, though its correlation with the rating of the Website was almost as large.

Exhibit 37. Pearson correlations between composites and overall ratings

Composite	Overall Rating of Website	Overall Rating of Help Line	Overall Rating of In-Person Help	Overall Rating of HIM	Recommend to Family and Friends
Apply	0.57	0.41	0.37	0.51	0.45
Web	0.75*	0.50	0.30	0.61	0.51
Phone	0.46	0.75*	0.31	0.51	0.40
In-Person	0.25	0.28	0.69*	0.36	0.31
HP Enroll	0.45	0.41	0.36	0.47	0.40

*Composite and overall rating align

6.4 Summary of Results

In Sections 6.1 and 6.2, we presented the results of the confirmatory and exploratory factor analyses. These analyses allowed us to:

1. Identify survey items to exclude from the hypothesized factor structure due to low covariance coverage.

2. Identify the underlying structure of the remaining items and test the validity of that structure.

3. Establish construct validity for the tentative revised factor structure with 23 items loading on five factors (see Exhibit 27): the CFA on that structure indicated statistically significant standardized factor loadings ≥ 0.40 for 22 out of 23 items (q32 had a factor loading of 0.36).

4. Show that all interfactor correlations were below 0.80, which indicates that these five factors, while related, do not overlap to the point of being redundant (see Exhibit 38).

5. Provide empirical support for the invariance of the factor structure across survey mode and language (configural invariance); however, we had insufficient data to test for invariant factor loadings (metric invariance) or invariant factor loadings *and* intercepts (scalar invariance).

Exhibit 38. Inter-factor correlations among final composites

Final Composites	Apply	Web	Phone	In Person	HP Enroll
Apply	1.00				
Web	0.71	1.00			
Phone	0.54	0.69	1.00		
In Person	0.49	0.47	0.48	1.00	
HP Enroll	0.65	0.71	0.64	0.65	1.00

In Section 6.3 we presented the results of various tests of reliability, variability, and validity. Our analyses demonstrated that the measurement properties of the tentative revised factor structure are all at or above target. We evaluated:

1. Internal consistency reliability,

2. State-level reliability,

3. Variability (ceiling and floor effects),

4. Discriminant validity (scaling success), and

5. Criterion (concurrent) validity.

The results of this work is summarized in Exhibit 39. As shown, each composite in the tentative revised factor structure met most, if not all, of the established criteria: an 85% overall success rate (41 'yes' cells out of 48 total cells in Exhibit 39).

Exhibit 39. Summary of measurement properties of final core composites

Measurement Property	Apply	Web	Phone	In Person	HP Enroll
All factor loadings ≥ 0.40	Yes	Yes	No	Yes	Yes
Internal consistency reliability >0.70	Yes	Yes	Yes	Yes	Yes
Observed State-level reliability >0.70	Yes	No	No	Yes	No
Effective sample size for IUR ≥ 0.70 is ≤ 300*	Yes	Yes	Yes	Yes	Yes
Effective sample size for IUR ≥ 0.90 is ≤ 300*	Yes	No	Yes	Yes	No
Ceiling effect < 50%	Yes	Yes	Yes	Yes	Yes
Ceiling effect < 25%	Yes	Yes	Yes	No	Yes
All correlations with other 4 scales < 0.80	Yes	Yes	Yes	Yes	Yes
Scaling success (100%)	Yes	Yes	Yes	Yes	Yes
Correlation with Associated Overall Rating > Correlation with other Ratings	n/a	Yes	Yes	Yes	n/a
Percent of criteria met	100%	80%	80%	90%	78%

*The effective sample size is the number of completed surveys required to reach an IUR of 0.70, adjusted for the composite level response rate.

6.5 Final Composite and Item Recommendations

Based on the factor analysis and other reliability and validity results, we decided to drop two additional items from the composite structure: q32 and q43, which both ask about Marketplace representatives using words or phrases the consumer did not understand, either by phone (q32) or in-person (q43). The reasons for dropping these items are detailed in Exhibit XX in Section 8.1.2, which provides a comprehensive summary of the disposition of all survey questions from the survey instrument.

Our final recommendation based on our analysis is to retain 21 items that map to five core composites. This structure is shown in Exhibit 40, which includes a crosswalk to the question number in the beta test version of the Marketplace Survey.

Exhibit 40. Final recommended composites

Q# Beta Test*	Q# Field Test†	Composites and Items 21 Items Mapped to 5 composites
		Application and Premium Tax Credit Eligibility Process (5 items)
3	3	Was it easy to give or update information about yourself or the people in your family who wanted health insurance?
4	4	Did giving or updating information about yourself or the people in your family take longer than you expected?
7	8	When you gave or updated your household income information, was it easy to find out if you or the people in your family could get help paying for health insurance?
8	9	Did giving or updating your household income information take longer than you expected?
13	17	Was it easy to understand how to update {INSERT MARKETPLACE NAME} about changes to your household income or the number of people in your family?
		Seeking Information on the Marketplace Website (4 items)
15	19	Since November 15th, how often did you have to wait to get what you needed because of problems on {INSERT MARKETPLACE NAME}'s website?
16	20	Since November 15th, how often did you get the information you needed from {INSERT MARKETPLACE NAME}'s website?
18	22	Since November 15th, how often was it easy to understand the information on {INSERT MARKETPLACE NAME}'s website?
20	24	Since November 15th, how often was the information on {INSERT MARKETPLACE NAME}'s website as helpful as you thought it should be?
		Seeking Information over the Phone (4 items)
23	27	Since November 15th, how often did you get the information or help you needed when you called {INSERT MARKETPLACE NAME}'s customer service Call Center?
25	29	Since November 15th, how often was it easy to understand the information you got when you called {INSERT MARKETPLACE NAME}'s customer service Call Center?
27	31	Since November 15th, how often was {INSERT MARKETPLACE NAME}'s customer service Call Center as helpful as you thought it should be?
29	34	Since November 15th, how often did {INSERT MARKETPLACE NAME}'s customer service Call Center staff treat you with courtesy and respect when you called?
		Seeking Information In-Person (4 items)
32	38	Since November 15th, how often did you get the information or help you needed when you met in person with someone about getting health insurance from {INSERT MARKETPLACE NAME}?
34	40	Since November 15th, how often was it easy to understand the information you got when you met in person with someone about getting health insurance from {INSERT MARKETPLACE NAME}?
36	42	Since November 15th, how often were the persons you met with about getting health insurance from {INSERT MARKETPLACE NAME} as helpful as you thought they should be?
37	44	Since November 15th, how often did the persons you met with about getting health insurance from {INSERT MARKETPLACE NAME} treat you with courtesy and respect?
		Health Plan Enrollment Process (4 items)
42	49	Since November 15th, how often was it easy to understand the services covered by the health plans available to you and how much you would have to pay?
44	51	Since November 15th, how often was it easy to understand which health plans had the doctors or hospitals you wanted?
46	53	Since November 15th, how often was it easy to understand which health plans covered the prescription medicines you needed?
53	59	Since November 15th, was it easy to choose a health plan?

*Health Insurance Marketplace Survey for 2015 Beta Test: December 11, 2014
†Health Insurance Marketplace Survey for 2014 Field Test: July 7, 2014

AIR
AMERICAN INSTITUTES FOR RESEARCH®

7.0 Case Mix Analysis

7.1 Background

One of the primary purposes CAHPS surveys is to be able to compare providers, treatment centers, or health plans—more generally referred to as "reporting units"—to some benchmark, typically the mean of all reporting units in a particular universe. For the Marketplace Survey, the reporting unit is the state, and ultimately the benchmark for any score will be the national average of that score. Other benchmarks, such as regional or marketplace type (FFM, SPM, SBM, etc.), may be considered in the future.

Past research using Hospital CAHPS data has shown that some types of respondents, such as older respondents or respondents in better health, tend to give higher ratings of their hospital care than respondents who are younger or in poor health.[30] Conversely, those respondents with more education tend to give lower ratings of their health care experiences. These are characteristics of the respondents that are related to the CAHPS scores but are not within the control of the service provider, nor are they believed to reflect true differences in the quality of the service that is delivered.

In the context of the Marketplace Survey data, when comparing states to a benchmark, the differences reported to any audience should derive as much as possible from differences in the quality of service provided to consumers. If the differences derive in part from differences in the respondent populations across states, then it will be important to remove (i.e., adjust for) the portion of the scores that come from individual characteristics so that the Marketplaces are not held accountable for factors that are beyond their control.[31] Thus, the three goals of case-mix adjustment are to:

1. Help remove the effects of individual characteristics that can affect scores and ratings;

2. Remove effects that might be considered spurious (i.e., that reflect something other than quality of the Marketplace experience); and

3. Remove incentives for states to avoid groups of consumers that are likely to provide low ratings.[32]

Three conditions must be met in the selection of variables for case-mix adjustment of Marketplace Survey scores:

1. Within states, the case-mix variables must be related to the outcome measures. That is, the variables must have sufficient predictive power in relation to the outcomes (e.g., older respondents give higher ratings of their overall Marketplace experience). These variables are referred to as "predictors" of the outcome being examined.

---

[30] O'Malley AJ, Zaslavsky AM, Elliott MN, Zaborski L, Cleary PD. Case-mix adjustment of the CAHPS Hospital Survey. Health Services Research. 2005;40(6, part 2):2162-81.

[31] Zaslavsky AM, Zaborski LB, Ding L, Shaul JA, Cioffi MJ, Cleary PD. Adjusting performance measures to ensure equitable plan comparisons. Health Care Financing Review. 2001;22(3):109-26.

[32] Zaslavsky AM. Issues in case-mix adjustment of measures of the quality of health plans. Proc Gov Soc Stat. American Statistical Association, 1998, 56-64.

2. The distributions of these predictor variables must vary among states. For example, some states are likely to have younger Marketplace consumer populations than other states. This condition is the heterogeneity factor of the predictor.

3. The case-mix variables must be appropriate for adjustment because they are not themselves determined by the actions of the states. That is, they must be characteristics that are brought to the Marketplace by the consumer (e.g., age or education), not characteristics that might be consequences of the consumer's satisfaction with, or assessment of, the state Marketplace. Predictors that are consequences of the consumer's satisfaction with the state Marketplace are endogenous.

The case-mix analysis follows four steps:

1. Selection of potential case-mix adjusters;

2. Estimation of predictive power of the selected adjusters;

3. Estimation of heterogeneity; and

4. Estimation of the impact of each adjuster.

Predictive power, heterogeneity, and impact are necessary conditions for choosing a case-mix adjuster.

7.2 Variable Recoding

Exhibit 41 displays the variables we evaluated as potential case mix adjusters. We chose variables as potential case mix adjusters if they were standard CAHPS case mix adjusters, or if we thought they might affect any of the five global ratings and vary across states. For simplicity, when describing the case mix analyses below, we use the shortened variable names provided in Exhibit 41.

Exhibit 41. Variables evaluated as potential case mix adjusters

Question #	Shortened Variable Name	Questionnaire Item or Definition
Q11	Medicaid Eligibility	Since October 1st, did you qualify for Medicaid, the program in your state that provides health plan coverage for some low-income people, families and children, pregnant women, and persons with disabilities? *(Yes, No, Don't Know)*
Q12	Subsidy Eligibility	Since October 1st, did the {INSERT MARKETPLACE NAME} help you pay for your health insurance? *(Yes, No, Don't Know)*
Q68	Health Rating: General	In general, how would you rate your overall health? *(Excellent, Very Good, Good, Fair, Poor)*
Q69	Health Rating: Mental	In general, how would you rate your overall mental or emotional health *(Excellent, Very Good, Good, Fair, Poor)*
Q70-73	Comorbid Conditions	Q70: Since October 1st, did you get health care 3 or more times for the same condition or problem? *(Yes, No)* Q71: Is this a condition or problem that has lasted for at least 3 months? Do not include pregnancy or menopause. *(Yes, No)* Q72: Do you now need or take medicine prescribed by a doctor? Do not include birth control. *(Yes, No)* Q73: Is this medicine to treat a condition that has lasted for at least 3 months? Do not include pregnancy or menopause *(Yes, No.)*
Q74	Disability: Deaf	Are you deaf or do you have serious difficulty hearing? *(Yes, No)*
Q75	Disability: Blind	Are you blind or do you have serious difficulty seeing, even when wearing glasses? *(Yes, No)*

Question #	Shortened Variable Name	Questionnaire Item or Definition
Q76	Difficulty Concentrating/Remembering	Because of a physical, mental, or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions? *(Yes, No)*
Q77	Difficulty Walking/Climbing stairs	Do you have serious difficulty walking or climbing stairs? *(Yes, No)*
Q78	Difficulty Dressing/Bathing	Because of a physical, mental, or emotional condition, do you have difficulty dressing or bathing? *(Yes, No)*
Q79	Difficulty Errands	Because of a physical, mental, or emotional condition, do you have difficulty doing errands alone such as visiting a doctor's office or shopping? *(Yes, No)*
Q80	Age	What is your age? *(18 to 24 years, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65 to 74, 75 or older)*
Q81	Sex	What is your sex? *(Male, Female)*
Q82	Education	What is the highest grade or level of school that you have completed? *(8th grade or less, Some high school but did not graduate, High school graduate or GED, Some college or 2-year degree, 4-year college graduate, More than 4-year college degree)*
Q83	Employment Status	What best describes your employment status? Mark only ONE. *(Employed full-time, Employed part-time, A homemaker, A full-time student, Retired, Unable to work for health reasons, Unemployed, Other)*
Q84	Hispanic Ethnicity	Are you Hispanic, Latino/a, or Spanish origin? *(Yes, No)*
Q86	Race	What is your race? *(White, Black or African American, American Indian or Alaska Native, Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian, Native Hawaiian, Guamanian or Chamorro, Samoan, Other Pacific Islander)*
Q87	Eligibility for Indian Health Services	Are you eligible to get health services from an Indian Health Service, tribal, or urban Indian health program? *(Yes, No, Don't Know)*
Q91	Health Insurance Status	Did you have health insurance in the United States at any time between January 1st and December 31st, 2013? *(Yes, No)*
Q92	Health Insurance Confidence	How confident are you that you understand health insurance terms? *(Not at all confident, Slightly confident, Moderately confident, Very confident)*
Q93	Comfort with Computers/Internet	Do you feel comfortable using the internet through a computer, tablet, or smart phone? *(Yes definitely, Yes somewhat, No)*
Q94	Assistance Filling out Survey	Did someone help you complete this survey? *(Yes, No)*
Frame	Household Size	Number of applicants associated with the application.
Frame	Language	Language in which the survey was administered *(English, Spanish, or Chinese)*
Frame	Survey mode	Mode of survey administration *(Online, Telephone, In person)*

Due to the structure of the survey items, several potential case mix variables required recoding before running the variable selection models. We recoded categorical variables as a series of k-1 dummy variables (where k is the number of response options associated with the categorical variable). For example, we coded language (which had values of English, Spanish, or Chinese) as two dummy variables for Spanish and Chinese. Both dummy variables had values of 0 or 1 to indicate that a person did or did not complete a survey in that language. For each dummy coded variable, we established a referent category, which we excluded from the models. For example, for language, English was the referent category and was left out of the models. We also dummy

AIR
AMERICAN INSTITUTES FOR RESEARCH®

coded age (which we treated as a categorical variable), sex, education, employment status, eligibility for Indian Health Services, eligibility for Medicaid, and survey mode.

In a few cases, we re-coded items that had "1: yes," "2: no" response categories in a similar fashion (i.e., to 0/1 dummy codes), thus creating indicator variables equal to 1 for the positive response. We coded the following variables in this way: health insurance status, Hispanic ethnicity, assistance filling out survey, subsidy eligibility, disability: deaf, disability: blind, Activities of Daily Living (ADL): concentrating/remembering, Difficulty walking/climbing stairs, Difficulty climbing stairs; and Difficulty errands.

We reverse-coded several variables (i.e., "flipped" the values, for example such that 1 became 7, 2 became 6, etc.) in order to make the interpretation of parameter estimates and the relationships between variables more meaningful. For example, the general health rating had values of "1: excellent," "2: very good," "3: good," "4: fair," and "5: poor." We reverse coded this variable, flipping the values such that increasing values represented increasing good health. We also reverse-coded the potential adjuster variables mental health rating and comfort with computers/internet, and the global rating item "Would you recommend the Health Insurance Marketplace to your friends and family? (Yes definitely, Yes somewhat, No)".

A different recoding logic was required for race, which had a code-all-that-apply format. The race item asked respondents to "please mark one or more," implying that applicants should leave inapplicable response options blank. To code race, we considered a value to be missing only if all six response options were missing. For each non-missing response, we set a series of dummy variables to either 1 or 0, depending on the responses to the race choices. For example, if respondents marked only "White," the dummy variable White was coded as "1"; if they marked both "White" and "Asian," then a new dummy variable captured this as multiple races (i.e., "multi") by coding it as "1." Other race categories included Black; American Indian or Alaskan Native; and Asian, Hawaiian, or Pacific Islander.

Finally, we created a variable to capture use of acute and chronic care and medications . The variable "comorbid conditions" had four values: 0 if the applicant did not receive health care three times or more for the same condition and did not take medicine prescribed by a doctor; 1 if the applicant received acute care (i.e., less than three times) or acute medication use (i.e., used medication to treat a condition for less than three months); 2 if the applicant received chronic care (i.e., for more than three months) or chronic medication use (i.e., for longer than three months); and 3 if the applicant received both chronic care and had chronic medication use. We dummy coded this variable as described above.

7.2 Variable Selection

Potential case mix adjusters included applicant characteristics (e.g., age, sex, education) and several design variables (e.g., survey mode, language). The complete list of variables assessed as potential case mix adjusters is shown in Exhibit 41.

We used stepwise regression to select a subset of the potential case-mix adjusters for further analysis. Stepwise regression analyses evaluated the strength of the relationship of each potential adjuster to five rating variables (see Exhibit 42) in separate models in which each rating variable was regressed on all of the potential adjusters.

Exhibit 42. Global rating (outcome) variables

Question #	Shortened Name	Questionnaire Item
Q25	Website rating	We want to know your rating of the {INSERT MARKETPLACE NAME} website, {INSERT MARKETPLACE URL}, that you visited since October 1, 2013. Using any number from 0 to 10, where 0 is the worst website possible and 10 is the best website possible, what number would you use to rate the {INSERT MARKETPLACE NAME} website?
Q35	Help line rating	We want to know your rating of the {INSERT MARKETPLACE NAME} customer service Help Line that you called since October 1, 2013. Using any number from 0 to 10, where 0 is the worst customer service Help Line possible and 10 is the best customer service Help Line possible, what number would you use to rate the {INSERT MARKETPLACE NAME} customer service Help Line?
Q45	In-person assistance rating	We want to know your rating of the in-person assistance you got to help you use the {INSERT MARKETPLACE NAME} since October 1, 2013. Using any number from 0 to 10, where 0 is the worst in-person assistance possible and 10 is the best in-person assistance possible, what number would you use to rate the assistance you got when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}?
Q66	Global marketplace rating	Using any number from 0 to 10, where 0 is the worst health insurance marketplace possible and 10 is the best health insurance marketplace possible, what number would you use to rate your {INSERT MARKETPLACE NAME} since October 1st?
Q67	Recommend marketplace	Would you recommend the {INSERT MARKETPLACE NAME} to your friends and family?

In our stepwise regression models, we added variables one-by-one to the model. For a variable to remain in the model, its F-statistic had to be significant at $p<.05$. Upon addition of a new variable to the model, each variable already in the model was re-assessed and variables that no longer retained an F-statistic significant at the retention p-level ($p<.05$) were excluded from the model. Only after this check was made and the necessary deletions accomplished was another variable added to the model. The stepwise process was complete for a given model when none of the variables outside the model had an F statistic significant at $p<.05$ and every variable in the model was statistically significant at $p<.05$. Adjuster variables selected in any of the models formed a core set of potential case mix adjusters eligible for final selection (see Exhibit 43).

Exhibit 43. Selection status for variable selection models

Predictors	Web Site Rating n=926[a]	Help Line Rating n=738[a]	In-person Assistance Rating n=326[b]	Global Rating of Marketplace Experience n=1278[a]	Recommend Marketplace n=1291[a]
Age	√	√		√	√
Comorbid Conditions	√				
Disability: Deaf		√	√		
Education				√	
Eligibility for Indian Health Services			√		
Employment Status			√		
Health Insurance Confidence				√	√
Health Rating: General	√			√	
Household Size	√	√		√	√
Language	√	√		√	√
Medicaid Eligibility			√		
Race	√	√		√	√
Subsidy Eligibility	√	√	√	√	√
Survey mode	√	√	√	√	√
Difficulty Concentrating/Remembering					
Difficulty Dressing/Bathing					
Difficulty Errands					
Difficulty Walking/Climbing stairs					
Assistance Filling out Survey					
Comfort with Computers/Internet					
Disability: Blind					
Health Insurance Status					
Health Rating: Mental					
Hispanic					
Sex					
Adjusted R^2	0.10	0.10	0.12	0.18	0.19

√ Selected in stepwise regression at *p*<0.05.

[a] Model excludes Eligibility for Indian Health Services, Computer Literacy, and Medicaid Eligibility to increase N as these variables were not significant in preliminary models. Sample size not equal to full sample available due to deletion of cases with missing values on case mix variables.

[b] Model excludes Health insurance status and Health insurance literacy to increase N as these variables were not significant in preliminary models. Sample size not equal to full sample available due to deletion of cases with missing values on case mix variables.

AIR ®
AMERICAN INSTITUTES FOR RESEARCH®

7.4 Estimating Heterogeneity, Predictive and Explanatory Power, and Impact Factors

We estimated the heterogeneity factor, predictive power, explanatory power, and impact factor for each potential case mix variable selected in the regression models. We measured the heterogeneity of the predictor variables across states as the ratio of between-state to within-state variance of the residuals when the variable was regressed on all other potential case mix adjusters in a random effects model. We measured the heterogeneity of outcome variables across states as the ratio of between-state to within-state variance of the residuals when the variable was regressed on state in a random effects model. We measured predictive power as the incremental amount of variance explained by the predictor (represented as the partial r^2 x 1000) in the stepwise regression analyses, controlling for the other potential case mix adjusters. To measure explanatory power, which considers both the predictive power of each potential adjuster and the heterogeneity of the adjusters across states, we calculated the predictive power x the adjuster heterogeneity factor. Finally, we calculated the impact factor, which standardizes explanatory power with respect to the overall variance in the outcome being assessed, as explanatory power / outcome heterogeneity. We considered variables that had an impact factor > 1.0 as candidates for case mix adjusters (O'Malley et al, 2005).

Results are shown in Exhibit 44. Variables that had an impact factor > 1.0, and were therefore eligible to be considered as case mix adjusters, included: *age*, *education*, *household size*, *language*, *race*, and *subsidy eligibility*.

Exhibit 44. Parameter estimates and selection status for variable selection models

Case Adjustment Variables	Adjuster Heterogeneity	Website Rating Outcome Heterogeneity=0.0041		Help Line Rating Outcome Heterogeneity<0.0001		In-person Assistance Rating Outcome Heterogeneity=0.0032		Global Marketplace Rating Outcome Heterogeneity=0.0192		Recommend Marketplace Outcome Heterogeneity=0.0151	
		Partial r²	*Impact Factor* > 1.0*	*Partial r²*	*Impact Factor* > 1.0*	*Partial r²*	*Impact Factor* > 1.0*	*Partial r²*	*Impact Factor* > 1.0*	*Partial r²*	*Impact Factor* > 1.0*
Age	0.031	0.008	√	0.008	√	0.005		0.005	√	0.005	√
Education: High School Graduate or GED[1]	0.011	0.000		0.002		0.002		0.003	√	0.000	
Household Size	0.009	0.005	√	0.006	√	0.002		0.004	√	0.003	√
Language: Chinese[2]	0.031	0.004	√	0.003		0.005		0.000		0.000	
Language: Spanish[2]	0.350	0.002		0.006	√	0.010		0.014	√	0.014	√
Race: Asian, Hawaiian, or Pacific Islander[3]	0.004	0.005	√	0.009		0.005		0.021	√	0.018	√
Race: Black[3]	0.154	0.014	√	0.004		0.006		0.016	√	0.016	√
Subsidy Eligibility: Yes[4]	0.001	0.034	√	0.049	√	0.046	√	0.089	√	0.113	√

* Impact Factor = (Adjuster Heterogeneity * (R-square*1000)) / (Outcome heterogeneity)

[1] Reference category = More than a 4-year college degree

[2] Reference category = English

[3] Reference category = White

[4] Reference category = No subsidy eligibility

7.5 Case Mix Recommendations

While the statistical evidence for choosing case-mix adjusters is fairly straightforward, the judgment regarding the endogeneity of adjusters is more controversial and ultimately more challenging to make. The published literature on CAHPS surveys makes various recommendations regarding what constitutes "standard" CAHPS case-mix adjusters. For example, standard Medicare CAHPS case-mix adjusters include age, self-rated overall health, self-rated mental health, education, assistance with survey (i.e., whether or not a proxy helped in completing the survey), and a Medicaid eligibility indicator.[33,34]

There can be some controversy regarding the choice of adjusters. For example, in terms of health plan experiences, people who have been enrolled in Medicaid at some point are typically happier with a private health plan once they are able to enroll in one. If this experience is reflected across all plans, then health plans with a greater number of former Medicaid enrollees will tend to get higher scores if no case mix adjustment is used. If scores are adjusted, those plans that enroll a greater share of former Medicaid enrollees will thus have their scores down-adjusted. Health plan issuers object that this down-adjustment is unfair, even though it appears to meet the requirements of inclusion as a case-mix adjuster.

The publicly available document that provides instructions for analyzing data from CAHPS surveys recommends only three adjusters: age, education, and self-rated overall health.[35] Even with this smaller set of recommended adjusters, there are complaints that the adjustment for self-rated overall health in effect punishes health plans that improve the behavioral and physical health of their Medicare members by decreasing (i.e., down-adjusting) their Medicare CAHPS scores over time.[36] These are issues to consider when evaluating potential case mix adjusters for the Marketplace Survey scoring approach.

Another consideration relates to the need or desire to hold states accountable for the quality of services they provide to particular vulnerable populations. Case mix adjustment for characteristics such as race, ethnicity, and disability status would mask variations in consumer experience by controlling for these characteristics (i.e., holding them constant). On one hand, the decision to case mix adjust for a particular characteristic is a decision to not hold states accountable for the quality of services they provide to certain populations. For example, if people with disabilities report more negative experiences in some states, case mix adjustment for disability status will up-adjust scores in those states and down-adjust scores in states that provide better services. On the other hand, the decision to exclude certain characteristics from case mix adjustment could incentivize states to avoid those consumers for whom they are unable to provide quality service. Thus, assuming the statistical evidence supports the use of certain consumer characteristics in case mix adjustment, the decision to use those characteristics hinges

---

[33] Elliott MN, Hambarsoomians K, Edwards CA. Analysis of case-mix strategies and recommendations for Medicare fee-for-service CAHPS. Case-mix adjustment report. Santa Monica, CA: The RAND Corporation; 2005

[34] Elliott MN, Beckett MK, Chong K, Hambarsoomians K, Hays RD. How do proxy responses and proxy-assisted responses differ from what Medicare beneficiaries might have reported about their health care? Health Services Research. 2008;4(3):833-48.

[35] See p. 1 and p. 5 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 from: https://cahps.ahrq.gov/surveys-guidance/docs/2015_instructions_for_analyzing_data.pdf.

[36] DSS Research. Medicare Case-mix Adjustments Can Penalize Plans with Healthy Members [Internet]. Looking Beyond the Expected blog. 2011 August 8. Available from: http://blog.dssresearch.com/?p=97.

on weighing the relative risks of removing accountability for *underserving* vulnerable populations versus creating an incentive to provide *no services at all* to those populations.

We suggest age, education, the number of people on the application, and language, as case mix adjusters for the Health Insurance Marketplace Survey based on the case mix analyses described herein. We also recommend including assistance completing the survey, as well as the general and mental health ratings, as they are considered standard CAHPS case mix adjusters. Although subsidy eligibility had an impact, we are concerned that it is endogenous and thus not appropriate as an adjuster. It is not possible to disentangle whether a respondent reported that they were able to get help paying for health insurance because they were eligible for a subsidy or because they were able to get far enough along in the application process to get this information, the latter being an indicator of Marketplace quality. In addition, although race qualifies as a case mix adjuster, as mentioned above, race is typically recommended as a stratification variable rather than a case mix adjuster.

The next step will be to assess the impact of each case mix adjuster separately by comparing unadjusted results to adjusted results for the overall rating and the five composites.

8.0 Survey Revisions

8.1 Instrument

8.1.2 Overview of Marketplace Survey Revisions for Beta Test

Overall, we dropped 15 questions to reduce the length of the survey and added 2 new questions on re-enrollees and 1 new question on multiple chronic conditions. There was a net decrease of 12 questions which brought the total number of survey questions from 95 to 83. The most substantial changes to the Marketplace Survey for the Beta Test had to do with revisions or restructuring to address the experiences of re-enrollees. We defined re-enrollees as those who had health insurance through the Marketplace last year rather than just interacted with the Marketplace. We used this definition because the process of updating family and income information and re-selecting a health plan only applies to people who already enrolled in a plan. A summary of all the changes to the Marketplace Survey for the Beta Test are listed below followed by more detailed descriptions of the most substantive changes.

- Changed the reference period from October 1, 2013 to November 15, 2014 to align with the start of open enrollment in 2014

- Dropped 15 questions to reduce length (more detail below)

- Added a new question that measures multiple chronic conditions (more detail below)

- Added a new first question in the survey that identifies re-enrollees (more detail below)

- Added a question in the 'Choosing a Health Plan' section about whether re-enrollees chose the same health plan they had in 2014 through the Marketplace (more detail below)

- Added a new response option about "not finding the same health plan you had in 2014" in the information seeking questions that ask about the reasons why someone did not get

the information they needed from the website, phone, or in-person to account for re-enrollee experiences (more detail below).

- Added 'or update' into questions that asked about giving information about the people in your family and giving household income information to the Marketplace to account for re-enrollee experiences (more detail below).

- Reworded 'the people in your family, including yourself' to say 'yourself or the people in your family' to avoid the problem of individuals skipping out of questions because they were not giving information about other family members to the Marketplace.

- Reworded 'What kind of information was not easy to understand' to 'What kind of information was hard to understand' to make the question cognitively easier to comprehend, especially during a telephone interview.

- Reworded 'customer service Help Line' to 'customer service Call Center' to align more closely to the terminology used by Marketplaces.

8.1.2 Detailed Descriptions of Marketplace Survey Revisions for Beta Test

As mentioned above, we dropped 15 questions in an effort to shorten the survey:

1. Q2: Were any of the following a reason why you did not give information about the people in your family, including yourself, who wanted health insurance? Mark one or more.

2. Q7: Were any of the following a reason why you did not give your household income information? Mark one or more.

3. Q10: How did you give your household income information?

4. Q14: Since October 1st, were you told by the {INSERT MARKETPLACE NAME} how to appeal the decision?

5. Q15: Was it easy to understand how to appeal the decision?

6. Q32: Since October 1st, how often did the {INSERT MARKETPLACE NAME} customer service Help Line use words or phrases you did not understand when you called?

7. Q37: Since October 1st, did you want in-person help but were unable to get it because the building was not accessible for persons with disabilities?

8. Q43: Since October 1st, how often did the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} use words or phrases you did not understand?

9. Q70: Since October 1st, did you get health care 3 or more times for the same condition or problem?

10. Q71: Is this a condition or problem that has lasted for at least 3 months? Do not include pregnancy or menopause.

11. Q72: Do you now need or take medicine prescribed by a doctor? Do not include birth control.

12. Q73: Is this medicine to treat a condition that has lasted for at least 3 months? Do not include pregnancy or menopause.

13. Q87: Are you eligible to get health services from an Indian Health Service, tribal, or urban Indian health program?

14. Q88: Did you ever get health services from an Indian Health Service, tribal, or urban Indian health program?

15. Q93: Do you feel comfortable using the internet through a computer, tablet, or smart phone?

Questions 2 and 7 were included in the list because they were designed to capture tourists or people who were exploring the Marketplace but never intended to purchase health insurance, which we believe will not be as common after the first year of open enrollment. These questions created very complicated skip patterns in the mail survey (more than 30% did not follow the skip pattern correctly). Also, only 11-12% of people actually answered these questions.

Question 10 was included in the list because the information was redundant with Q5, since most people gave information about the people in their family and their income information using the same mode (web, mail, phone, or in-person).

Questions 14 and 15 were included in the list because of low item response. Q15 was an assessment item that had to be dropped in our psychometrics due to low covariance coverage with other items. Q14 is the screener for Q15 so it was removed as well. We decided to keep Q13 which asks if they were told they could appeal a decision about how much they had to pay for their health insurance because we think it is important to track the percentage of enrollees who received this information over time.

Questions 32 and 43 performed poorly in the psychometric analyses. They had low correlations with their scales and lacked discriminant validity (they cross-loaded with other scales). The internal reliability (Cronbach's Alpha) and inter-unit reliability for their scales both improved when these questions were dropped.

Q37 was included in the list because it had both complicated skip patterns and a low screen-in rate of 2%.

Q70-Q73 were dropped in favor of one question intended to identify multiple chronic conditions. Q70-Q73 measure chronic condition status without identifying how many chronic conditions the respondent had. In an effort to reduce the length of the survey and focus on multiple chronic conditions, which is a more important issue for policy and oversight purposes, we dropped the four CAHPS questions that measure chronic condition status and wrote a new question that measures the presence of multiple chronic conditions.

Q87-88 measure eligibility and utilization of Indian Health Services. Less than 1% of psychometric test respondents screened into these questions. We do not believe these questions provide useful information regarding Native American experiences with the Marketplace. With a large enough sample size we could still measure Native American experiences with the Marketplace by using the self-identification as Native American from the race question.

The question about comfort using the internet (Q93) was dropped because it was correlated with age and the relationship between being comfortable using the internet and the website global rating disappeared when age was added to regression model.

Exhibit 45. Summary of changes to the Marketplace Survey

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
New Question	N/A	N/A	Did you have health insurance through the {INSERT MARKETPLACE NAME} at any time in 2014?	This question was added to identify new enrollees from re-enrollees.
Q1	Since October 1st, did you give information about the people in your family, including yourself, who wanted health insurance through the {INSERT MARKETPLACE NAME}?	Keep	Since November 15th, did you give or update information about yourself or the people in your family who wanted health insurance through {INSERT MARKETPLACE NAME}?	"Or update" was added to make this question applicable to re-enrollees. "Yourself or the people in your family" was added to reduce confusion about the meaning of "people in your family, including yourself," which was used in the original question.
Q2	Were any of the following a reason why you did **not** give information about the people in your family, including yourself, who wanted health insurance?	Drop	N/A	Questions 2 and 7 were dropped because they were designed to capture tourists or people who were exploring the Marketplace but never intended to purchase health insurance. We believe this will not be as common after the first year of open enrollment. These questions created very complicated skip patterns in the mail survey (more than 30% did not follow the skip pattern correctly). Also, only 11-12% of people actually answered these questions.
Q3	Was it easy to give information about the people in your family, including yourself, who wanted health insurance?	Keep	Was it easy to give or update information about yourself or the people in your family who wanted health insurance?	"Or update" was added to make this question applicable to re-enrollees. "Yourself or the people in your family" was added to reduce confusion about the meaning of "people in your family, including yourself," which was used in the original question.
Q4	Did giving information about the people in your family, including yourself, take longer than you expected?	Keep	Did giving or updating information about yourself or the people in your family take longer than you expected?	"Or update" was added to make this question applicable to re-enrollees. "Yourself or the people in your family" was added to reduce confusion about the meaning of "people in your family, including yourself," which was used in the original question.
Q5	How did you give information about the people in your family, including yourself?	Keep	How did you give or update information about yourself or the people in your family?	"Or update" was added to make this question applicable to re-enrollees. "Yourself or the people in your family" was added to reduce confusion about the meaning of "people in your family, including yourself," which was used in the original question.

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q6	Since October 1st, did you give the {INSERT MARKETPLACE NAME} information about your household income to see if you could get help paying for your health insurance?	Keep	Household income can be your income or the income from people in your family. Since November 15th, did you give or update information about your household income to see if you or the people in your family could get help paying for health insurance through {INSERT MARKETPLACE NAME}?	The definition of "household income" and "you or the people in your family" was added to make this question applicable to family members as well as individuals. "Or update" was added to make this question applicable to re-enrollees.
Q7	Were any of the following a reason why you did **not** give your household income information?	Drop	N/A	Questions 2 and 7 were dropped because they were designed to capture tourists or people who were exploring the Marketplace but never intended to purchase health insurance. We believe this will not be as common after the first year of open enrollment. These questions created very complicated skip patterns in the mail survey (more than 30% did not follow the skip pattern correctly). Also, only 11-12% of people actually answered these questions.
Q8	When you gave your household income information, was it easy to find out if you could get help paying for your health insurance?	Keep	When you gave or updated your household income information, was it easy to find out if you or the people in your family could get help paying for health insurance?	"Or update" was added to make this question applicable to re-enrollees. "You or the people in your family" was added to make this question applicable to family members as well as individuals.
Q9	Did giving your household income information take longer than you expected?	Keep	Did giving or updating your household income information take longer than you expected?	"Or updating" was added to make this question applicable to re-enrollees.
Q10	How did you give your household income information ?	Drop	N/A	Question 10 was dropped because the information was redundant with Question 5. Most people gave information about the people in their family and their income information using the same mode (web, mail, phone, or in-person).

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q11	Since October 1st, did you qualify for *Medicaid*, the program in your state that provides health plan coverage for some low-income people, families and children, pregnant women, and persons with disabilities?	Keep	Since November 15th, did you or the people in your family qualify for {INSERT MEDICAID NAME}, the program in your state that provides health plan coverage for some low-income persons, families and children, pregnant women, and persons with disabilities?	"You or the people in your family" was added to make this question applicable to family members as well as individuals. As a result of TEP feedback, we will use the state specific program name for Medicaid since SBM states may not be using the term "Medicaid."
Q12	Since October 1st, did the {INSERT MARKETPLACE NAME} help you pay for your health insurance?	Keep	Since November 15th, did {INSERT MARKETPLACE NAME} help you or the people in your family pay for your health insurance?	"You or the people in your family" was added to make this question applicable to family members as well as individuals.
Q13	To appeal means to tell someone at the {INSERT MARKETPLACE NAME} that you think the decision is wrong, and ask for a fair review of the decision. Since October 1st, were you told by the {INSERT MARKETPLACE NAME} that you could appeal if you disagreed with the decision about how much you would have to pay for your health insurance?	Keep	N/A	Analyses did not identify any problems with this question.
Q14	Since October 1st, were you told by the {INSERT MARKETPLACE NAME} how to appeal the decision?	Drop	N/A	This question has low applicability and screen in rates. Questions 14 and 15 were dropped because Question 15 was an assessment item that had to be dropped in our psychometrics, due to low covariance coverage with other items. Question 14 is the screener for Question 15 so it was removed as well.
Q15	Was it easy to understand how to appeal the decision?	Drop	N/A	This question has low applicability and screen in rates. Questions 14 and 15 were dropped because Question 15 was an assessment item that had to be dropped in our psychometrics, due to low covariance coverage with other items. Question 14 is the screener for Question 15 so it was removed as well.

Development of an Enrollee Satisfaction Survey for Use in the Health Insurance Marketplace

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q16	Since October 1st, were you told by the {INSERT MARKETPLACE NAME} that you should update them about changes to your household income or the number of people in your family?	Keep	N/A	Analyses did not identify any problems with this question.
Q17	Was it easy to understand how to update the {INSERT MARKETPLACE NAME} about changes to your household income or the number of people in your family?	Keep	N/A	Analyses did not identify any problems with this question.
Q18	Since October 1st, did you visit the {INSERT MARKETPLACE NAME} website {INSERT MARKETPLACE URL}?	Keep	N/A	Analyses did not identify any problems with this question.
Q19	Since October 1st, how often did you have to wait to get what you needed because of problems on the {INSERT MARKETPLACE NAME} website?	Keep	N/A	This was an item that we could drop to increase the Cronbach's alpha, but the alpha is not significantly changed when the question is dropped. Also, website problems may still exist this year, so the question is still applicable.
Q20	Since October 1st, how often did you get the information you needed from the {INSERT MARKETPLACE NAME} website?	Keep	N/A	Analyses did not identify any problems with this question.
Q21	Were any of the following a reason why you did **not** get the information you needed from the {INSERT MARKETPLACE NAME} website?	Keep	N/A	Analyses did not identify any problems with this question.
Q22	Since October 1st, how often was it easy to understand the information on the {INSERT MARKETPLACE NAME} website?	Keep	N/A	Analyses did not identify any problems with this question.
Q23	What kind of information on the {INSERT MARKETPLACE NAME} website was **not** easy to understand?	Keep	What information on {INSERT MARKETPLACE NAME}'s website was hard to understand?	The language "was not easy" was confusing, according to TEP feedback and CATI Behavior Coding.

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q24	Since October 1st, how often was the information on the {INSERT MARKETPLACE NAME} website as helpful as you thought it should be?	Keep	N/A	Analyses did not identify any problems with this question.
Q25	We want to know your rating of the {INSERT MARKETPLACE NAME} website, {INSERT MARKETPLACE URL}, that you visited since October 1, 2013. Using any number from 0 to 10, where 0 is the worst website possible and 10 is the best website possible, what number would you use to rate the {INSERT MARKETPLACE NAME} website?	Keep	N/A	Analyses did not identify any problems with this question.
Q26	Since October 1st, did you call the {INSERT MARKETPLACE NAME} customer service Help Line?	Keep	Since November 15, did you call {INSERT MARKETPLACE NAME}'s customer service Call Center?	"Call Center" is a term more consistent with Marketplace language
Q27	Since October 1st, how often did you get the information or help you needed when you called the {INSERT MARKETPLACE NAME} customer service Help Line?	Keep	Since November 15th, how often did you get the information or help you needed when you called {INSERT MARKETPLACE NAME}'s customer service Call Center?	"Call Center" is a term more consistent with Marketplace language
Q28	Were any of the following a reason why you did **not** get the information or help you needed when you called the {INSERT MARKETPLACE NAME} customer service Help Line?	Keep	Were any of the following a reason why you did **not** get the information or help you needed when you called {INSERT MARKETPLACE NAME}'s customer service Call Center?	"Call Center" is a term more consistent with Marketplace language
Q29	Since October 1st, how often was it easy to understand the information you got when you called the {INSERT MARKETPLACE NAME} customer service Help Line?	Keep	Since November 15th, how often was it easy to understand the information you got when you called {INSERT MARKETPLACE NAME}'s customer service Call Center?	"Call Center" is a term more consistent with Marketplace language

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q30	What kind of information was **not** easy to understand when you called the {INSERT MARKETPLACE NAME} customer service Help Line?	Keep	What information was hard to understand when you called {INSERT MARKETPLACE NAME}'s customer service Call Center?	The language "was not easy" was confusing, according to TEP feedback and CATI Behavior Coding.
Q31	Since October 1st, how often was the {INSERT MARKETPLACE NAME} customer service Help Line as helpful as you thought it should be?	Keep	Since November 15th, how often was {INSERT MARKETPLACE NAME}'s customer service Call Center as helpful as you thought it should be?	"Call Center" is a term more consistent with Marketplace language
Q32	Since October 1st, how often did the {INSERT MARKETPLACE NAME} customer service Help Line use words or phrases you did not understand when you called?	Drop	N/A	This question has low factor loading in the full model with full data (0.357). Dropping question 32 from the Info Seek on the Phone composite would raise Cronbach's alpha from 0.79 to 0.86. It has low correlations with its scales and a lack of discriminant validity (cross-loads with other scales). Preliminary results show that this item is not an important driver of global rating for this information seeking section. The item may not be measuring information seeking or anything about the customer service, but rather language barriers or HIL of the respondent, which is not what we want. The TEP liked this item because they thought it measured HIL but it does not fit in a HIL composite. The one benefit is that the IUR is good. Question 32 has an observed IUR of 0.65 and would only require 83 completed surveys to get an IUR of 0.70. However, when Question 32 was dropped the IUR for the composite improved.
Q33	Since October 1st, did you speak to a person when you called the {INSERT MARKETPLACE NAME} customer service Help Line?	Keep	Since November 15th, did you speak to a person when you called {INSERT MARKETPLACE NAME}'s customer service Call Center?	"Call Center" is a term more consistent with Marketplace language

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q34	Since October 1st, how often did the {INSERT MARKETPLACE NAME} customer service Help Line staff treat you with courtesy and respect when you called?	Keep	Since November 15th, how often did {INSERT MARKETPLACE NAME}'s customer service Call Center staff treat you with courtesy and respect when you called?	"Call Center" is a term more consistent with Marketplace language
Q35	We want to know your rating of the {INSERT MARKETPLACE NAME} customer service Help Line that you called since October 1, 2013. Using any number from 0 to 10, where 0 is the worst customer service Help Line possible and 10 is the best customer service Help Line possible, what number would you use to rate the {INSERT MARKETPLACE NAME} customer service Help Line?	Keep	We want to know your rating of {INSERT MARKETPLACE NAME}'s customer service Call Center that you called since November 15th, 2014. Using any number from 0 to 10, where 0 is the worst customer service Call Center possible and 10 is the best customer service Call Center possible, what number would you use to rate {INSERT MARKETPLACE NAME}'s customer service Call Center?	"Call Center" is a term more consistent with Marketplace language
Q36	Since October 1st, did you meet in person with anyone from an organization that helps people get health insurance through the {INSERT MARKETPLACE NAME}?	Keep	N/A	Analyses did not identify any problems with this question.
Q37	Since October 1st, did you want in-person help but were unable to get it because the building was not accessible for persons with disabilities?	Drop	N/A	This question was difficult for consumers to understand. It is double barreled and produces complicated skip patterns. Also, 98% of respondents answered "no."
Q38	Since October 1st, how often did you get the information or help you needed when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}?	Keep	N/A	Analyses did not identify any problems with this question. We addressed TEP feedback about sentence structure and decided, in order to prevent repetitiveness, we should keep the section of the sentence that varies in the beginning of the sentence.
Q39	Were any of the following a reason why you did **not** get the information or help you needed when you met in person with someone about getting health insurance from the	Keep	N/A	Analyses did not identify any problems with this question.

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
	{INSERT MARKETPLACE NAME}?			
Q40	Since October 1st, how often was it easy to understand the information you got when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}?	Keep	N/A	Analyses did not identify any problems with this question. We addressed TEP feedback about sentence structure and decided, in order to prevent repetitiveness, we should keep the section of the sentence that varies in the beginning of the sentence.
Q41	What kind of information was **not** easy to understand when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}?	Keep	What information was hard to understand when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}?	The language "was not easy" was confusing, according to TEP feedback and CATI Behavior Coding.
Q42	Since October 1st, how often were the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} as helpful as you thought they should be?	Keep	N/A	Analyses did not identify any problems with this question. We addressed TEP feedback about sentence structure and decided, in order to prevent repetitiveness, we should keep the section of the sentence that varies in the beginning of the sentence.
Q43	Since October 1st, how often did the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} use words or phrases you did not understand?	Drop	N/A	Dropping question 43 from the Info Seek In-Person composite would raise Cronbach's alpha from 0.79 to 0.83. It has low correlations with its scales and a lack of discriminant validity (cross-loads with other scales). Preliminary results show that this item is not an important driver of global rating for this information seeking section. This item also may not be measuring information seeking or anything about the customer service, but rather language barriers or HIL of the respondent, which is not what we want. The TEP liked this item because they thought it measured HIL, but it does not fit in a HIL composite. The one benefit is that the IUR is good. Question 43 has an observed IUR of 0.78 and would only require 44 completed surveys to get an IUR of 0.70. However, when question 43 was dropped, the IUR for the composite improved.

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q44	Since October 1st, how often did the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} treat you with courtesy and respect?	Keep	N/A	Analyses did not identify any problems with this question. We addressed TEP feedback about sentence structure and decided, in order to prevent repetitiveness, we should keep the section of the sentence that varies in the beginning of the sentence.
Q45	We want to know your rating of the in-person assistance you got to help you use the {INSERT MARKETPLACE NAME} since October 1, 2013. Using any number from 0 to 10, where 0 is the worst in-person assistance possible and 10 is the best in-person assistance possible, what number would you use to rate the assistance you got when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}?	Keep	N/A	Analyses did not identify any problems with this question.
Q46	Since October 1st, were you looking for health insurance for yourself through the {INSERT MARKETPLACE NAME}?	Keep	N/A	Analyses did not identify any problems with this question.
Q47	Since October 1st, were you looking for health insurance for another family member, such as a spouse or child, through the {INSERT MARKETPLACE NAME}?	Keep	N/A	Analyses did not identify any problems with this question.
Q48	Since October 1st, did you consider the services covered by the health plans available to you in the {INSERT MARKETPLACE NAME} and how much you would have to pay?	Keep	Since November 15th, did you consider the services covered by the health plans available to you through {INSERT MARKETPLACE NAME} and how much you would have to pay?	Analyses did not identify any problems with this question. The wording was adjusted to ensure that the question was grammatically correct when both FFM and SBM Marketplace names were inserted into the fill.

Development of an Enrollee Satisfaction Survey for Use in the Health Insurance Marketplace

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q49	Since October 1st, how often was it easy to understand the services covered by the health plans available to you and how much you would have to pay?	Keep	N/A	Analyses did not identify any problems with this question.
Q50	Since October 1st, did you try to find out which health plans in the {INSERT MARKETPLACE NAME} had the doctors or hospitals you wanted?	Keep	Since November 15th, did you try to find out which health plans available to you through {INSERT MARKETPLACE NAME} had the doctors or hospitals you wanted?	Analyses did not identify any problems with this question. The wording was adjusted to ensure that the question was grammatically correct when both FFM and SBM Marketplace names were inserted into the fill.
Q51	Since October 1st, how often was it easy to understand which health plans had the doctors or hospitals you wanted?	Keep	N/A	Analyses did not identify any problems with this question.
Q52	Since October 1st, did you try to find out which health plans in the {INSERT MARKETPLACE NAME} covered the prescription medicines you needed?	Keep	Since November 15th, did you try to find out which health plans available to you through {INSERT MARKETPLACE NAME} covered the prescription medicines you needed?	Analyses did not identify any problems with this question. The wording was adjusted to ensure that the question was grammatically correct when both FFM and SBM Marketplace names were inserted into the fill.
Q53	Since October 1st, how often was it easy to understand which health plans covered the prescription medicines you needed?	Keep	N/A	Question 53 has an observed IUR of 0.46 and would require 186 completed surveys to get an IUR of 0.70, which is a good IUR. Since the field test only included FFMs, we may want to keep the question for the beta test to see how it does with some SBMs, since there will be more Marketplace variation. Also, from an oversight perspective we think this question is important to keep because comparing drug formularies across health plans is important for consumer choice.
Q54	Special therapy includes physical, occupational, or speech therapy. Since October 1st, did you need any special therapy?	Keep, tentatively	N/A	This is a potential item to cut. It could be a useful measure with a larger sample size, but we will re-assess during the beta test. Specialized services are important benefits and CMS may want to know about the consumer's experience trying to find health plans with these services. However, 92.7%(2197/2370) said "No" to this question.

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q55	Since October 1st, was it easy to find out which health plans in the {INSERT MARKETPLACE NAME} offered the physical, occupational, or speech therapy services you needed?	Keep, tentatively	Since November 15th, was it easy to find out which health plans available to you through {INSERT MARKETPLACE NAME} offered the physical, occupational, or speech therapy services you needed?	This is a potential item to drop. It could be a useful measure with a larger sample size, but we will re-assess during beta test. It has a low screen-in rate of 7%. We dropped item 55 in the factor analysis because it was the only item left in the Specialized Services composite after question 57 was dropped in the factor analysis due to low covariance coverage. The wording was adjusted to ensure that the question was grammatically correct when both FFM and SBM Marketplace names were inserted into the fill.
Q56	Home health care or assistance means home nursing, help with bathing or dressing, and help with basic household tasks. Since October 1st, did you need someone to come into your home to give you home health care or assistance?	Keep, tentatively	N/A	This is a potential item to drop. It could be a useful measure with a larger sample size, but we will re-assess during the beta test. Specialized services are important benefits and CMS may want to know about the consumer's experience trying to find health plans with these services. However, 98.35% (2326/2365) of respondents said "No" to this item.
Q57	Since October 1st, was it easy to find out which health plans in the {INSERT MARKETPLACE NAME} offered home health care services you needed?	Keep, tentatively	Since November 15th, was it easy to find out which health plans available to you through {INSERT MARKETPLACE NAME} offered home health care services you needed?	This is a potential item to drop. It could be a useful measure with a larger sample size, but we will re-assess during the beta test. It has a low screen-in rate of 2%. We dropped this item in the factor analysis due to low covariance coverage. The wording was adjusted to ensure that the question was grammatically correct when both FFM and SBM Marketplace names were inserted into the fill.
Q58	Did you choose a health plan through the {INSERT MARKETPLACE NAME}?	Keep	Since November 15th, did you choose a health plan through {INSERT MARKETPLACE NAME}?	"Since November 15th" was added to make the question applicable to the beta test year.
New Question	N/A	N/A	Did you choose a health plan through the {INSERT MARKETPLACE NAME}?	This question was added to determine if re-enrollees stayed in the same health plan as they had the previous year.
Q59	Was it easy to choose a health plan?	Keep	Since November 15th, was it easy to choose a health plan?	"Since November 15th" was added to make the question applicable to the beta test year.
Q60	An interpreter is someone who helps you talk with others who do not speak your language. Since October 1st, did you need an interpreter to help you speak with	Keep	N/A	Cultural competence is important to measure. 16% of the sample said 'Yes' to this question, which is more than we were expecting. We will re-assess with a larger sample size in the beta test.

Development of an Enrollee Satisfaction Survey for Use in the Health Insurance Marketplace

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
	anyone about getting health insurance from the {INSERT MARKETPLACE NAME}?			
Q61	Since October 1st, when you needed an interpreter to help you speak with anyone about getting health insurance from the {INSERT MARKETPLACE NAME}, how often did you get one?	Keep	N/A	Cultural competence is important to measure. We dropped this item in the factor analysis due to low covariance coverage. We will re-assess with a larger sample size in the beta test.
Q62	Since October 1st, did you fill out any forms for the {INSERT MARKETPLACE NAME}?	Keep	N/A	Cultural competence is important to measure. We will re-assess with a larger sample size in the beta test. 46% of the sample said 'Yes' to this question.
Q63	Since October 1st, how often were the forms that you had to fill out through the {INSERT MARKETPLACE NAME} available in the language you prefer?	Keep	N/A	Cultural competence is important to measure. We dropped item 63 in the factor analysis because it was the only item left in the Cultural Competence composite after question 61 and question 65 were dropped from the factor analysis. We will re-assess with a larger sample size in the beta test.
Q64	Since October 1st, did you need the forms in a different format, such as large print or braille?	Keep, tentatively	N/A	Cultural competence is important to measure. We will re-assess with a larger sample size in the beta test. 97.49%(1010/1036) said "No" to this question.
Q65	Since October 1st, how often were the forms that you had to fill out available in the format you needed, such as large print or braille?	Keep, tentatively	N/A	Cultural competence is important to measure. We will re-assess with a larger sample size in the beta test. We dropped this item in the factor analysis due to low covariance coverage.
Q66	Using any number from 0 to 10, where 0 is the worst health insurance marketplace possible and 10 is the best health insurance marketplace possible, what number would you use to rate your {INSERT MARKETPLACE NAME} since October 1st?	Keep	N/A	This question has significantly contributed to analyses.

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q67	Would you recommend the {INSERT MARKETPLACE NAME} to your friends and family?	Keep	N/A	This question has significantly contributed to analyses.
Q68	In general, how would you rate your overall health?	Keep	N/A	This question is a potential case mix adjuster.
Q69	In general, how would you rate your overall **mental or emotional** health?	Keep	N/A	This question is a potential case mix adjuster.
New Question	N/A	N/A	In the last 12 months, did you get care for 2 or more health problems or conditions that each lasted for at least a year?	CMS expressed interest in measuring the experience of respondents who believe they have chronic conditions. We revised this question by combining aspects of Q70 and Q71, allowing us to use only one question to identify those who believe they have chronic conditions.
Q70	Since October 1st, did you get health care 3 or more times for the same condition or problem?	Drop	N/A	Although we believe measuring chronic conditions is important and could affect Marketplace experiences, we have not found sufficient evidence from the Marketplace Survey Field Test to justify keeping all (Q70-Q73) chronic condition questions. For example, overall Marketplace experiences does not vary by chronic condition status. In addition, chronic condition status is not a significant case-mix adjuster. However, CMS expressed interest in measuring the experience of respondents with multiple chronic conditions so we wrote a new question on multiple chronic conditions.
Q71	Is this a condition or problem that has lasted for at least 3 months? Do **not** include pregnancy or menopause.	Drop	N/A	Although we believe measuring chronic conditions is important and could affect Marketplace experiences, we have not found sufficient evidence from the Marketplace Survey Field Test to justify keeping all (Q70-Q73) chronic condition questions. For example, overall Marketplace experiences does not vary by chronic condition status. In addition, chronic condition status is not a significant case-mix adjuster. However, CMS expressed interest in

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
				measuring the experience of respondents with multiple chronic conditions so we wrote a new question on multiple chronic conditions.
Q72	Do you now need or take medicine prescribed by a doctor? Do **not** include birth control.	Drop	N/A	Although we believe measuring chronic conditions is important and could affect Marketplace experiences, we have not found sufficient evidence from the Marketplace Survey Field Test to justify keeping all (Q70-Q73) chronic condition questions. For example, overall Marketplace experiences does not vary by chronic condition status. In addition, chronic condition status is not a significant case-mix adjuster. However, CMS expressed interest in measuring the experience of respondents with multiple chronic conditions so we wrote a new question on multiple chronic conditions.
Q73	Is this medicine to treat a condition that has lasted for at least 3 months? Do **not** include pregnancy or menopause.	Drop	N/A	Although we believe measuring chronic conditions is important and could affect Marketplace experiences, we have not found sufficient evidence from the Marketplace Survey Field Test to justify keeping all (Q70-Q73) chronic condition questions. For example, overall Marketplace experiences does not vary by chronic condition status. In addition, chronic condition status is not a significant case-mix adjuster. However, CMS expressed interest in measuring the experience of respondents with multiple chronic conditions so we wrote a new question on multiple chronic conditions.
Q74	Are you deaf or do you have serious difficulty hearing?	Keep	N/A	This question is a potential case mix adjuster and could be used for subgroup analysis.
Q75	Are you blind or do you have serious difficulty seeing, even when wearing glasses?	Keep	N/A	This question is a potential case mix adjuster and could be used for subgroup analysis.
Q76	Because of a physical, mental, or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions?	Keep	N/A	This question is a potential case mix adjuster and could be used for subgroup analysis.
Q77	Do you have serious difficulty walking or climbing stairs?	Keep	N/A	This question is a potential case mix adjuster and could be used for subgroup analysis.
Q78	Because of a physical, mental, or emotional condition, do you have difficulty dressing or bathing?	Keep	N/A	This question is a potential case mix adjuster and could be used for subgroup analysis.
Q79	Because of a physical, mental, or emotional condition, do you have difficulty doing errands alone such as visiting a doctor's office or shopping?	Keep	N/A	This question is a potential case mix adjuster and could be used for subgroup analysis.

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q80	What is your age?	Keep	N/A	This question is a potential case mix adjuster.
Q81	What is your sex?	Keep	N/A	This question is a potential case mix adjuster.
Q82	What is the highest grade or level of school that you have completed?	Keep	N/A	This question is a potential case mix adjuster.
Q83	What best describes your employment status?	Keep	N/A	This question is a potential case mix adjuster.
Q84	Are you Hispanic, Latino/a, or Spanish origin?	Keep	N/A	This question is a potential case mix adjuster.
Q85	Which group best describes you?	Keep	N/A	This question is a potential case mix adjuster.
Q86	What is your race?	Keep	N/A	This question is a potential case mix adjuster.
Q87	Are you eligible to get health services from an Indian Health Service, tribal, or urban Indian health program?	Drop	N/A	Less than 1% of field test respondents screened into these questions. We do not believe these questions provide useful information regarding Native American experiences with the Marketplace. With a large enough sample size we could still measure Native American experiences with the Marketplace by using the self-identification as Native American from the race question.
Q88	Did you ever get health services from an Indian Health Service, tribal, or urban Indian health program?	Drop	N/A	Less than 1% of field test respondents screened into these questions. We do not believe these questions provide useful information regarding Native American experiences with the Marketplace. With a large enough sample size we could still measure Native American experiences with the Marketplace by using the self-identification as Native American from the race question.
Q89	What is your preferred language?	Keep	N/A	This question is a potential case mix adjuster.
Q90	How well do you speak English?	Keep	N/A	This question is a potential case mix adjuster.
Q91	Did you have health insurance in the United States at any time between January 1st and December 31st, 2013?	Keep	N/A	This question assesses if respondents had any health insurance last year, which is important because limited experience with health insurance may have an impact on Marketplace experiences. This question is distinct from other questions in the survey that ask about having the same health plan from the Marketplace last year.

AIR ®
AMERICAN INSTITUTES FOR RESEARCH®

Q#	Original Question	Keep/Drop	Revised Question Wording	Reason or Comments
Q92	How confident are you that you understand health insurance terms?	Keep	N/A	This question contributes to health insurance literacy analysis and is important for sub-group analyses. Marketplace experiences vary by health insurance literacy.
Q93	Do you feel comfortable using the internet through a computer, tablet, or smart phone?	Drop	N/A	Website rating does vary by this item, but when included in the driver analyses it does not hold up after age is included. Age and using the internet correlate such that people who are older are less comfortable using the internet. Relationship between comfortable using the internet and website global rating disappears when age is added to regression model.
Q94	Did someone help you complete this survey?	Keep	N/A	This question is a potential case mix adjuster.
Q95	How did that person help you?	Keep	N/A	This question is a potential case mix adjuster.

New Question on Multiple Chronic Conditions. We believe measuring chronic conditions is important and could affect Marketplace experiences. In an effort to reduce the length of the survey and focus on multiple chronic conditions for policy and oversight purposes, we dropped the four CAHPS questions that measure chronic condition status and wrote a new question that measures multiple chronic conditions. The new question is, "In the last 12 months, did you get care for 2 or more health problems or conditions that each lasted for at least a year?"

New Question to Identify Re-enrollees. We wanted to add a question in the survey that could distinguish re-enrollees from new enrollees in order to do analyses where we compare their experiences. The new question is: "Did you have health insurance through the {INSERT MARKETPLACE NAME} at any time in 2014?" Yes/No. We defined re-enrollees as those who had health insurance through the Marketplace last year rather than just interacted with the Marketplace. This is because the process of updating family and income information and re-selecting a health plan only applies to people who already enrolled in a plan.

New Questions about Difficulty Finding Same Plan as Last Year. Our Technical Expert Panel suggested adding a new question about how easy it was for a re-enrollee to find their same health plan from last year. The assumption is that it may not be that easy to do. Re-enrollees may not remember the marketing name of their health plan or they may have trouble entering in the long ID number associated with their plan. The benefit or cost structure may have changed so re-enrollees may not think it is the same health plan. The health plan name may have changed or may not exist anymore.

We already have a question that measures difficulties with choosing a plan, "was it easy to choose a health plan" that can apply to re-enrollees and new enrollees. In addition, we decided to add a new response option within the questions that ask about "reasons why someone did not get the information they needed from the website, phone, in-person" to address this issue more specifically. The new response option would be: "You/They could not find the same health plan you had in 2014." To measure whether someone does not know if they were enrolled in the same plan last year we ask a follow-up question to those who say they chose a plan during open enrollment for 2015 coverage, "Were you enrolled in that health plan in 2014?"

Modify Existing Questions for Re-enrollees. We modified existing questions to address the experiences of re-enrollees who went back into the Marketplace to update their family and income information. For example we changed "did you *give information* about the people in your family, including yourself, who wanted health insurance through the {INSERT MARKETPLACE NAME}?" to "did you *give or update information* about yourself or the people in your family who wanted health insurance through the {INSERT MARKETPLACE NAME}?" We know some people will only verify their information and not make any changes, but it seems that 'update' is the word being used with consumer facing materials from the Marketplace.

8.2 Next Steps

We plan to conduct additional cognitive testing on the Marketplace Survey. We would be testing changes to the survey since our second round of testing in Oct 2013, which includes quite a few changes from round 2 cognitive testing, TEP input, and the field test analyses. The Marketplaces themselves have changed a lot since Oct 2013 and that will affect consumer experiences and how they answer the survey questions. We also have the new re-enrollee population that we would like to

include and test to make sure the questions we have apply to them. We plan to do 9 interviews in English with consumers interacting with the VA, MA, or DC Marketplaces.

Appendix A:
Health Insurance Marketplace Survey (English)

Health Insurance Marketplace Survey

Language: English
Reference Period: Since October 1, 2013

Each item has been labeled to indicate the domain, construct source, and CAHPS or other survey indicator for this review process; the lists below provide the abbreviations used. For example, if a question is labeled: **(IS/F,T/HP5-AM-m1)**, it means this question is from the Information Seeking domain, the construct came from the Focus Groups and Technical Expert Panel, and the question wording is a modified version of the CAHPS Health Plan 5.0 Adult Medicaid Question #1. The headings in this survey are meant for respondent navigation, not domain headings.

Marketplace Domain Name
AP=Application Process
TC=Premium Tax Credit Eligibility
IS=Information Seeking
CuC=Cultural Competence
EP=Health Plan Enrollment Process
GR=Global Ratings
CM=Case Mix Adjusters
RC=Respondent Characteristics
SP=Specialized Services
All the questions have a domain label.

Construct Source
L=Lit Review
F=Focus Groups
S=Stakeholder Interviews
T=Technical Expert Panel
C=Centers for Medicare & Medicaid Services
CI1=Cognitive Interview Round 1
CI2=Cognitive Interview Round 2
OMB60 = OMB 60 Day Comment Period
OMB30 = OMB 30 Day Comment Period
Questions that don't have a construct source were included because they came from the CAHPS Health Plan 5.0 survey. For example, we included global ratings and case mix adjuster questions because they are a CAHPS convention.

Survey Indicator
HP5-AM-Q# = <u>CAHPS Health Plan 5.0, Adult Medicaid</u>, Question #
HP5-AM-mQ# = <u>CAHPS Health Plan 5.0, Adult Medicaid</u>, modified Question #
HP4-AS-mQ# = <u>CAHPS Health Plan 4.0, Adult Supplemental</u>, modified Question #
HP5-AS-mQ# = CAHPS Health Plan 5.0, Adult Supplemental, modified Question #
      These are new CAHPS questions that are not in public documentation yet.
CG2-AS-mQ# = <u>CAHPS Clinician & Group 2.0, Adult Supplemental</u>, modified Question #

H-mQ = Hospital CAHPS , Modified Question #

OMH-4302-Q# = HHS Office of Minority Health ACA Section 4302 Data Collection Standards, Question #

ACS-P-Q# = American Community Survey (ACS) – Person Section - Question #

NHBS-Q# = 2010 National HIV Behavioral Surveillance System – Question #

M-ACO-Q# = 2014 Medicare Provider Satisfaction Survey – Items for ACOs Participating in Medicare Initiatives – Question #

Questions that don't have a survey indicator are new questions written for the Marketplace Survey.

OVERVIEW MARKETPLACE SURVEY DOMAINS

I.   APPLICATION PROCESS

- Gave information about the people in your family who wanted health insurance
- Reason why you did not give information about the people in your family
- Easy to give information about the people in your family
- Giving information about the people in your family took longer than expected
- Mode used to give information about the people in your family
- Told should update Marketplace about changes to income or family size
- Easy to understand how to update Marketplace about changes to income or family size

II.   PREMIUM TAX CREDIT ELIGIBILITY

- Gave information about household income
- Reason why you did not give information about household income
- Easy to find out if could get help paying for health insurance
- Giving information about household income took longer than expected
- Mode used to give information about household income
- Qualify for Medicaid
- Marketplace help paying for health insurance
- Told could appeal decision about how much have to pay for health insurance
- Told how to appeal
- Easy to understand how to appeal

III. INFORMATION SEEKING ON THE WEBSITE

- Visited the Marketplace website
- Had to wait to get what you needed because of problems on website
- Got information you needed
- Why did not get information needed
- Easy to understand the information
- What kind of information not easy to understand
- Information as helpful as you thought it should be

IV.   INFORMATION SEEKING OVER THE PHONE

- Called the Marketplace Help Line
- Got information or help you needed
- Why did not get information or help needed
- Easy to understand the information
- What kind of information not easy to understand
- As helpful as you thought they should be
- Used words or phrases you did not understand
- Spoke to a person
- Treat you with courtesy and respect

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

V. INFORMATION SEEKING IN-PERSON

- Met in person with anyone from an organization that helps people get health insurance through Marketplace
- Unable to meet in person because building was not accessible for persons with disabilities
- Got information or help you needed
- Why did not get information or help needed
- Easy to understand the information
- What kind of information not easy to understand
- As helpful as you thought they should be
- Used words or phrases you did not understand
- Treat you with courtesy and respect

VI. HEALTH PLAN ENROLLMENT

- Who is covered in health plan
- Considered services covered and how much you have to pay
- Easy to understand services covered and how much you have to pay
- Try to find out which health plans had doctors or hospitals you wanted
- Easy to understand which health plans had doctors or hospitals you wanted
- Try to find out which health plans covered prescription medicines you needed
- Easy to understand which health plans covered prescription medicines you needed
- Chose a health plan through Marketplace
- Easy to choose a health plan

VII. SPECIALIZED SERVICES

- Easy to find out which health plans offer physical, occupational therapy you needed
- Easy to find out which health plans offer home health care services you needed

VIII. CULTURAL COMPETENCE

- Need interpreter
- How often got an interpreter
- Forms available in preferred language
- Forms available in preferred format, such as large print or braille

GLOBAL RATINGS

- Rating of information–Web
- Rating of information–Phone
- Rating of information–In-Person
- Rating of health insurance marketplace
- Recommend marketplace to friends and family

CASE MIX ADJUSTERS

- Rating of overall health
- Age
- Sex

RESPONDENT CHARACTERISTICS

- Rating of overall mental or emotional health
- Got health care 3 or more times for same condition
- Got health care 3 or more times for condition lasted for at least 3 months

- Take medicine prescribed by a doctor
- Take medicine for condition lasted for at least 3 months
- Are you deaf
- Are you blind
- Difficulty concentrating, remembering, or making decisions because of a physical, mental, or emotional condition
- Difficulty walking or climbing stairs
- Difficulty dressing or bathing because of a physical, mental, or emotional condition
- Education status
- Employment status
- Ethnicity
- Race
- Eligibility to get health services from Indian Health Service
- Received care at an Indian Health Service facility
- Preferred Language
- Rating of English language skills
- Covered by health insurance at any time in 2013
- Knowledge of health insurance terms
- Comfortable using the Internet
- Someone help you complete this survey
- How did someone help you complete this survey

Domain Overview Note: The Domain Overview is meant to provide a quick overview of what is measured in this survey. It is NOT meant to list hypothesized composite items. There are a mix of screener, assessment/composite, and single items listed under each domain. It also does NOT list out every item but rather is meant to cover unique constructs. For example, if there is a screener item and an assessment item that measure the same construct, then the assessment item is listed.

Introduction

We are asking you to complete this survey because you contacted the {INSERT MARKETPLACE NAME} to learn about your health insurance options since October 1, 2013. You might have used the website, sent an application by mail, ca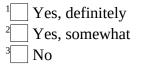lled the toll free Help Line, or met with someone in person. This survey asks about your experiences with the {INSERT MARKETPLACE NAME}, also known as Obamacare or Healthcare.gov, which was created by the Affordable Care Act.

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid Office of Management and Budget (OMB) control number. The valid OMB control number for this information collection is 0938-1221. The time required to complete this information collection is estimated to average 25 minutes per response, including the time to review instructions, search existing data resources, gather the data needed, and complete and review the information collection. If you have comments concerning the accuracy of the time estimate(s) or suggestions for improving this form, please write to: CMS, 7500 Security Boulevard, Attn: PRA Reports Clearance Officer, Mail Stop C4-26-05, Baltimore, Maryland 21244-1850.

Survey Instructions

Answer each question by marking the box next to your answer.

You are sometimes told to skip over some questions in this survey. When this happens you will see an arrow with a note that tells you what question to answer next, like this:

[1] ☐ Yes
[2] ☒ No → **If No, go to #1**

Development of an Enrollee Satisfaction Survey for Use in the Health Insurance Marketplace

Giving Information to Learn About Your Health Insurance Options

The following questions ask about your experiences giving information to learn about your health insurance options through the {INSERT MARKETPLACE NAME} since October 1, 2013. You might have used the website, sent an application by mail, called the toll free Help Line, or met with someone in person.

1. Since October 1ˢᵗ, did you give information about the people in your family, including yourself, who wanted health insurance through the {INSERT MARKETPLACE NAME}? **(AP/T,CI2)**

   ¹☐ Yes → **If Yes, go to #3**
   ²☐ No

2. Were any of the following a reason why you did **not** give information about the people in your family, including yourself, who wanted health insurance? *Mark one or more.* **(AP/CI2/HP4-AS-mCS1)**

	Did **not** give your family's information because
a) You did not have all the information they asked for	¹☐ (**Go to #6**)
b) You changed your mind and did not want to give your information	¹☐ (**Go to #6**)
c) You never intended to give your information	¹☐ (**Go to #6**)
d) There was a problem with the website	¹☐ (**Go to #6**)
e) Some other reason	¹☐ (**Go to #6**)

   Please specify: _____

   _____

3. Was it easy to give information about the people in your family, including yourself, who wanted health insurance? *If you did not give this information, go to #6*. **(AP/T,CI2)**

   ¹☐ Yes, definitely
   ²☐ Yes, somewhat
   ³☐ No

4. Did giving information about the people in your family, including yourself, take longer than you expected? **(AP/L,S,T, CI2)**

AIR
AMERICAN INSTITUTES FOR RESEARCH®

¹ ☐ Yes, definitely
² ☐ Yes, somewhat
³ ☐ No

5. How did you give information about the people in your family, including yourself? **(AP/T,CI1,CI2)**

¹ ☐ On the {INSERT MARKETPLACE NAME} website
² ☐ By mail
³ ☐ On the phone
⁴ ☐ In person

6. Since October 1ˢᵗ, did you give the {INSERT MARKETPLACE NAME} information about your household income to see if you could get help paying for your health insurance? **(TC/T)**

¹ ☐ Yes → **If Yes, go to #8**
² ☐ No

7. Were any of the following a reason why you did **not** give your household income information? *Mark one or more.* **(TC/CI2/HP4-AS-mCS1)**

<u>Did **not** give your information because</u>

a)	You did not have all the information they asked for	¹ ☐ **(Go to #16)**
b)	You changed your mind and did not want to give your information	¹ ☐ **(Go to #16)**
c)	You never intended to give your information	¹ ☐ **(Go to #16)**
d)	There was a problem with the website	¹ ☐ **(Go to #16)**
e)	Some other reason	¹ ☐ **(Go to #16)**

Please specify: _____

_____

8. When you gave your household income information, was it easy to find out if you could get help paying for your health insurance? *If you did not give this information, go to #16.* **(TC/T)**

¹ ☐ Yes, definitely
² ☐ Yes, somewhat
³ ☐ No

9. Did giving your household income information take longer than you expected? **(TC/L,S,T)**

¹☐ Yes, definitely
²☐ Yes, somewhat
³☐ No

10. How did you give your household income information? **(TC/T,CI1)**

¹☐ On the {INSERT MARKETPLACE NAME} website
²☐ By mail
³☐ On the phone
⁴☐ In person

11. Since October 1st, did you qualify for *Medicaid*, the program in your state that provides health plan coverage for some low-income people, families and children, pregnant women, and persons with disabilities? **(TC/T)**

¹☐ Yes → **If Yes, go to #13**
²☐ No
³☐ Don't know

12. Since October 1st, did the {INSERT MARKETPLACE NAME} help you pay for your health insurance? **(TC/T)**

¹☐ Yes
²☐ No
³☐ Don't know

13. To appeal means to tell someone at the {INSERT MARKETPLACE NAME} that you think the decision is wrong, and ask for a fair review of the decision. Since October 1st, were you told by the {INSERT MARKETPLACE NAME} that you could appeal if you disagreed with the decision about how much you would have to pay for your health insurance? **(TC/L,T)**

1 ☐ Yes
2 ☐ No → **If No, go to #16**

14. Since October 1ˢᵗ, were you told by the {INSERT MARKETPLACE NAME} how to appeal the decision? **(TC/CI1)**

1 ☐ Yes
2 ☐ No → **If No, go to #16**

15. Was it easy to understand how to appeal the decision? **(TC/L,T)**

1 ☐ Yes, definitely
2 ☐ Yes, somewhat
3 ☐ No

16. Since October 1ˢᵗ, were you told by the {INSERT MARKETPLACE NAME} that you should update them about changes to your household income or the number of people in your family? **(AP/CI1)**

1 ☐ Yes
2 ☐ No → **If No, go to #18**

17. Was it easy to understand how to update the {INSERT MARKETPLACE NAME} about changes to your household income or the number of people in your family? **(AP/CI1)**

1 ☐ Yes, definitely
2 ☐ Yes, somewhat
3 ☐ No

Looking for Information on the Marketplace Website

The following questions ask about your experiences when you visited the {INSERT MARKETPLACE NAME} website since October 1, 2013.

18. Since October 1ˢᵗ, did you visit the {INSERT MARKETPLACE NAME} website {INSERT MARKETPLACE URL}? **(IS/T)**

1 ☐ Yes
2 ☐ No → **If No, go to #26**

19. Since October 1ˢᵗ, how often did you have to wait to get what you needed because of problems on the {INSERT MARKETPLACE NAME} website? **(IS/OMB60)**

¹☐ Never
²☐ Sometimes
³☐ Usually
⁴☐ Always

20. Since October 1ˢᵗ, how often did you get the information you needed from the {INSERT MARKETPLACE NAME} website? **(IS/F,T/HP4-AS-mPW2)**

¹☐ Never
²☐ Sometimes
³☐ Usually
⁴☐ Always→ **If Always, go to #22**

21. Were any of the following a reason why you did **not** get the information you needed from the {INSERT MARKETPLACE NAME} website? *Mark one or more.* **(IS/F,T/HP4-AS-mCS1)**

<div align="right">

Did **not** get the
information because

</div>

a) You could not find the information you needed   ¹☐
b) The information was hard to understand   ¹☐
c) The website was confusing   ¹☐
d) It was hard to find out how to get help   ¹☐
e) The website was too complicated   ¹☐
f) The information the website gave you was wrong   ¹☐
g) The information was not in the language you prefer   ¹☐
h) The website did not work well with the special equipment
or software you use because of a disability   ¹☐
i) Some other reason   ¹☐
Please specify: _____

_____

22. Since October 1ˢᵗ, how often was it easy to understand the information on the {INSERT MARKETPLACE NAME} website? **(IS/L,S,T/HP4-AS-mPW3)**

¹☐ Never
²☐ Sometimes
³☐ Usually
⁴☐ Always → **If Always, go to #24**

▥ AIR®
AMERICAN INSTITUTES FOR RESEARCH®

23. What kind of information on the {INSERT MARKETPLACE NAME} website was **not** easy to understand? *Mark one or more.* **(IS/L,S,T/HP4-AS-mPW4)**

<u>**Not** easy to understand</u>

a) How to get help paying for your health insurance [1] ☐
b) Important deadlines [1] ☐
c) Benefits and coverage for doctor or specialist visits [1] ☐
d) Benefits and coverage for prescription drugs [1] ☐
e) Benefits and coverage for prenatal care or childbirth [1] ☐
f) How much you would have to pay for each health plan [1] ☐
g) How much you would have to pay out-of-pocket for health care services in each health plan [1] ☐
h) Which doctors are in each health plan [1] ☐
i) What you would have to pay if you used a doctor outside of the health plan [1] ☐
j) How to figure out your family size or income [1] ☐
k) Which doctors in each health plan have offices that are accessible for people with disabilities [1] ☐
l) How to find a health plan that meets your family's needs [1] ☐
m) Something else [1] ☐
   Please specify: _____

_____

24. Since October 1st, how often was the information on the {INSERT MARKETPLACE NAME} website as helpful as you thought it should be? **(IS/F,T/CG2-AC-m24)**

[1] ☐ Never
[2] ☐ Sometimes
[3] ☐ Usually
[4] ☐ Always

25.
We want to know your rating of the {INSERT MARKETPLACE NAME} website, {INSERT MARKETPLACE URL}, that you visited since October 1, 2013. Using any number from 0 to 10, where 0 is the worst website possible and 10 is the best website possible, what number would you use to rate the {INSERT MARKETPLACE NAME} website? **(GR/HP5-AM-m26)**

☐ 0 Worst website possible
☐ 1
☐ 2
☐ 3
☐ 4
☐ 5
☐ 6
☐ 7
☐ 8
☐ 9
☐ 10 Best website possible

Getting Information over the Phone

The following questions ask about your experiences when you called the {INSERT MARKETPLACE NAME} customer service Help Line since October 1, 2013.

26. Since October 1st, did you call the {INSERT MARKETPLACE NAME} customer service Help Line? **(IS/T)**

    1 ☐ Yes
    2 ☐ No → **If No, go to #36**

27. Since October 1st, how often did you get the information or help you needed when you called the {INSERT MARKETPLACE NAME} customer service Help Line? **(IS/F,T/HP5-AM-m22)**

    1 ☐ Never
    2 ☐ Sometimes
    3 ☐ Usually
    4 ☐ Always → **If Always, go to #29**

28. Were any of the following a reason why you did **not** get the information or help you needed when you called the {INSERT MARKETPLACE NAME} customer service Help Line? *Mark one or more.* **(IS/F,T/HP4-AS-mCS1)**

<u>Did **not** get the information or help needed because</u>

a) They were unable to answer your questions [1]
b) Was on hold too long [1]
c) You had to call several times before you could speak with someone [1]
d) You waited too long for someone to call you back [1]
e) No one called you back [1]
f) The information they gave you was wrong [1]
g) They did not have the information you needed [1]
h) The information they gave you was hard to understand [1]
i) You could not talk to someone in the language you prefer [1]
j)There was no video relay service available for persons who are deaf [1]
k) Some other reason [1]
     Please specify: _____

     _____

AIR
AMERICAN INSTITUTES FOR RESEARCH

29. Since October 1st, how often was it easy to understand the information you got when you called the {INSERT MARKETPLACE NAME} customer service Help Line? **(IS/L,S,T/HP4-AS-mPW3)**

1 ☐ Never
2 ☐ Sometimes
3 ☐ Usually
4 ☐ Always → **If Always, go to #31**

30. What kind of information was **not** easy to understand when you called the {INSERT MARKETPLACE NAME} customer service Help Line? *Mark one or more.* **(IS/L,S,T/HP4-AS- mPW4)**

<u>**Not** easy to understand</u>

a) How to get help paying for your health insurance   1 ☐
b) Important deadlines   1 ☐
c) Benefits and coverage for doctor or specialist visits   1 ☐
d) Benefits and coverage for prescription drugs   1 ☐
e) Benefits and coverage for prenatal care or childbirth   1 ☐
f) How much you would have to pay for each health plan   1 ☐
g) How much you would have to pay out-of-pocket for
   health care services in each health plan   1 ☐
h) Which doctors are in each health plan   1 ☐
i) What you would have to pay if you used a doctor outside
   of the health plan   1 ☐
j) How to figure out your family size or income   1 ☐
k) Which doctors in each health plan have offices that are
   accessible for people with disabilities   1 ☐
l) How to find a health plan that meets your family's needs   1 ☐
m) Something else   1 ☐
   Please specify: _____

_____

31. Since October 1st, how often was the {INSERT MARKETPLACE NAME} customer service Help Line as helpful as you thought it should be? **(IS/F,T/CG2-AC-m24)**
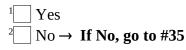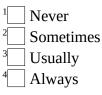
1 ☐ Never
2 ☐ Sometimes
3 ☐ Usually
4 ☐ Always

AIR
AMERICAN INSTITUTES FOR RESEARCH®

32. Since October 1ˢᵗ, how often did the {INSERT MARKETPLACE NAME} customer service Help Line use words or phrases you did not understand when you called? **(IS/L,T/CG2-AS-mCU2)**

¹ ☐ Never
² ☐ Sometimes
³ ☐ Usually
⁴ ☐ Always

33. Since October 1ˢᵗ, did you speak to a person when you called the {INSERT MARKETPLACE NAME} customer service Help Line? **(IS/CI1)**
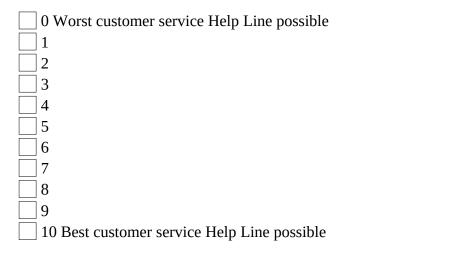
¹ ☐ Yes
² ☐ No → **If No, go to #35**

34. Since October 1ˢᵗ, how often did the {INSERT MARKETPLACE NAME} customer service Help Line staff treat you with courtesy and respect when you called? **(IS/L,F/HP5-AM-m23)**

¹ ☐ Never
² ☐ Sometimes
³ ☐ Usually
⁴ ☐ Always

35. We want to know your rating of the {INSERT MARKETPLACE NAME} customer

service Help Line that you called since October 1, 2013. Using any number from 0 to 10, where 0 is the worst customer service Help Line possible and 10 is the best customer service Help Line possible, what number would you use to rate the {INSERT MARKETPLACE NAME} customer service Help Line? **(GR/HP5-AM-m26)**

☐ 0 Worst customer service Help Line possible
☐ 1
☐ 2
☐ 3
☐ 4
☐ 5
☐ 6
☐ 7
☐ 8
☐ 9
☐ 10 Best customer service Help Line possible

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Getting Information In Person

The following questions ask about your experiences when you met in person with anyone from an organization that helps people get health insurance through the {INSERT MARKETPLACE NAME}, since October 1, 2013.
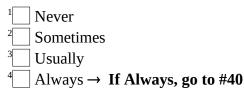
36. Since October 1st, did you meet in person with anyone from an organization that helps people get health insurance through the {INSERT MARKETPLACE NAME}? **(IS/T)**

 1 ☐ Yes → **If Yes, go to #38**
 2 ☐ No

37. Since October 1st, did you want in-person help but were unable to get it because the building was not accessible for persons with disabilities? **(IS/OMB60)**

 1 ☐ Yes→ **If Yes, go to #46**
 2 ☐ No → **If No, go to #46**

38. Since October 1st, how often did you get the information or help you needed when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}? **(IS/F,T/HP5-AM-m22)**

 1 ☐ Never
 2 ☐ Sometimes
 3 ☐ Usually
 4 ☐ Always → **If Always, go to #40**

39. Were any of the following a reason why you did **not** get the information or help you needed when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}? *Mark one or more.* **(IS/F,T/HP4-AS-mCS1)**

	Did **not** get the information or help because
a) There was not enough time	1 ☐
b) They did not have the information you needed	1 ☐
c) The information they gave you was hard to understand	1 ☐
d) The information they gave you was wrong	1 ☐
e) You could not talk or sign to someone in the language you prefer	1 ☐
f)Some other reason	1 ☐

      Please specify: _____

_____

40. Since October 1ˢᵗ, how often was it easy to understand the information you got when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}? **(IS/L,S,T/HP4-AS-mPW3)**

- 1 ☐ Never
- 2 ☐ Sometimes
- 3 ☐ Usually
- 4 ☐ Always → **If Always, go to #42**

41. What kind of information was **not** easy to understand when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}? *Mark one or more.* **(IS/L,S,T/HP4-AS-mPW4)**

<u>**Not** easy to understand</u>

- a) How to get help paying for your health insurance   1 ☐
- b) Important deadlines   1 ☐
- c) Benefits and coverage for doctor or specialist visits   1 ☐
- d) Benefits and coverage for prescription drugs   1 ☐
- e) Benefits and coverage for prenatal care or childbirth   1 ☐
- f) How much you would have to pay for each health plan   1 ☐
- g) How much you would have to pay out-of-pocket for health care services in each health plan   1 ☐
- h) Which doctors are in each health plan   1 ☐
- i) What you would have to pay if you used a doctor outside of the health plan   1 ☐
- j) How to figure out your family size or income   1 ☐
- k) Which doctors in each health plan have offices that are accessible for people with disabilities   1 ☐
- l) How to find a health plan that meets your family's needs   1 ☐

m)

Something else 1 ☐

Please specify: _____

_____

42. Since October 1ˢᵗ, how often were the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} as helpful as you thought they should be? **(IS/F,T/CG2-AC-m24)**
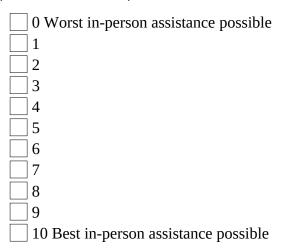
- 1 ☐ Never
- 2 ☐ Sometimes
- 3 ☐ Usually
- 4 ☐ Always

43. Since October 1st, how often did the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} use words or phrases you did not understand? **(IS/L,T/CG2-AS-mCU2)**

1 ▢ Never
2 ▢ Sometimes
3 ▢ Usually
4 ▢ Always

44. Since October 1st, how often did the persons you met with about getting health insurance from the {INSERT MARKETPLACE NAME} treat you with courtesy and respect? **(IS/L,F/HP5-AM-m23)**

1 ▢ Never
2 ▢ Sometimes
3 ▢ Usually
4 ▢ Always

45. We want to know your rating of the in-person assistance you got to help you use the {INSERT MARKETPLACE NAME} since October 1, 2013. Using any number from 0 to 10, where 0 is the worst in-person assistance possible and 10 is the best in-person assistance possible, what number would you use to rate the assistance you got when you met in person with someone about getting health insurance from the {INSERT MARKETPLACE NAME}? **(GR/HP5-AM-m26)**

▢ 0 Worst in-person assistance possible
▢ 1
▢ 2
▢ 3
▢ 4
▢ 5
▢ 6
▢ 7
▢ 8
▢ 9
▢ 10 Best in-person assistance possible

AIR
AMERICAN INSTITUTES FOR RESEARCH®

Choosing a Health Plan

The following questions ask about your experience choosing a health plan through the {INSERT MARKETPLACE NAME} since October 1, 2013.

46. Since October 1ˢᵗ, were you looking for health insurance for yourself through the {INSERT MARKETPLACE NAME}? **(EP/C)**
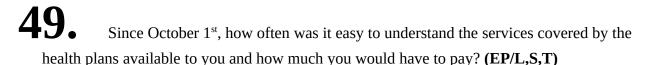
¹ ☐ Yes
² ☐ No

47. Since October 1ˢᵗ, were you looking for health insurance for another family member, such as a spouse or child, through the {INSERT MARKETPLACE NAME}? **(EP/C)**
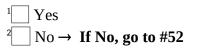
¹ ☐ Yes
² ☐ No

48. Since October 1ˢᵗ, did you consider the services covered by the health plans available to you in the {INSERT MARKETPLACE NAME} and how much you would have to pay? **(EP/L,S,T)**

¹ ☐ Yes
² ☐ No → **If No, go to #50**

49. Since October 1ˢᵗ, how often was it easy to understand the services covered by the health plans available to you and how much you would have to pay? **(EP/L,S,T)**

¹ ☐ Never
² ☐ Sometimes
³ ☐ Usually
⁴ ☐ Always

50. Since October 1ˢᵗ, did you try to find out which health plans in the {INSERT MARKETPLACE NAME} had the doctors or hospitals you wanted? **(EP/L,S,T)**

¹ ☐ Yes
² ☐ No → **If No, go to #52**

51. Since October 1ˢᵗ, how often was it easy to understand which health plans had the doctors or hospitals you wanted? **(EP/L,S,T)**

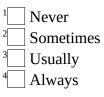¹☐ Never
²☐ Sometimes
³☐ Usually
⁴☐ Always

52. Since October 1ˢᵗ, did you try to find out which health plans in the {INSERT MARKETPLACE NAME} covered the prescription medicines you needed? **(EP/OMB30)**

¹☐ Yes
²☐ No → **If No, go to #54**

53. Since October 1ˢᵗ, how often was it easy to understand which health plans covered the prescription medicines you needed? **(EP/OMB30)**

¹☐ Never
²☐ Sometimes
³☐ Usually
⁴☐ Always

54. Special therapy includes physical, occupational, or speech therapy. Since October 1ˢᵗ, did you need any special therapy? **(SP/C/HP5-AS-CC11)**

¹☐ Yes
²☐ No → **If No, go to #56**

55. Since October 1ˢᵗ, was it easy to find out which health plans in the {INSERT MARKETPLACE NAME} offered the physical, occupational, or speech therapy services you needed? **(SP/C/ HP5-AS-mCC12)**

¹☐ Yes, definitely
²☐ Yes, somewhat
³☐ No

56. Home health care or assistance means home nursing, help with bathing or dressing, and help with basic household tasks. Since October 1ˢᵗ, did you need someone to come into your home to give you home health care or assistance? **(SP/C/ HP5-AS-CC13)**

¹☐ Yes
²☐ No → **If No, go to #58**

57.
Since October 1ˢᵗ, was it easy to find out which health plans in the {INSERT MARKETPLACE NAME} offered home health care services you needed? **(SP/C/ HP5-AS-mCC14)**

¹ ☐ Yes, definitely
² ☐ Yes, somewhat
³ ☐ No

58.
Did you choose a health plan through the {INSERT MARKETPLACE NAME}? (EP/T)

¹ ☐ Yes
² ☐ No → **If No, go to #60**

59. Was it easy to choose a health plan? **(EP/L,S,T/HP5-AM-m25)**

¹ ☐ Yes, definitely
² ☐ Yes, somewhat
³ ☐ No

Language Services

The following questions ask about language services, such as using an interpreter when you needed one, through the {INSERT MARKETPLACE NAME} since October 1, 2013.

60. An interpreter is someone who helps you talk with others who do not speak your language. Since October 1ˢᵗ, did you need an interpreter to help you speak with anyone about getting health insurance from the {INSERT MARKETPLACE NAME}? **(CuC/S,T/ HP5-AS-mNew_Q#)**

¹ ☐ Yes
² ☐ No → **If No, go to #62**

61.
Since October 1ˢᵗ, when you needed an interpreter to help you speak with anyone about getting health insurance from the {INSERT MARKETPLACE NAME}, how often did you get one? **(CuC/S,T/ HP5-AS-mNew_Q#)**

¹☐ Never
²☐ Sometimes
³☐ Usually
⁴☐ Always

62. Since October 1st, did you fill out any forms for the {INSERT MARKETPLACE NAME}? **(CUC/CI2)**

¹☐ Yes
²☐ No → **If No, go to #66**

63.   Since October 1st, how often were the forms that you had to fill out through the {INSERT MARKETPLACE NAME} available in the language you prefer? (**CuC/S,T/CG2-AS-mHL32)**

¹☐ Never
²☐ Sometimes
³☐ Usually
⁴☐ Always

64. Since October 1st, did you need the forms in a different format, such as large print or braille? **(CuC/OMB30/HP5-AM-m24)**

¹☐ Yes
²☐ No → **If No, go to #66**

65. Since October 1st, how often were the forms that you had to fill out available in the format you needed, such as large print or braille? (**CuC/OMB30/CG2-AS-mHL32)**
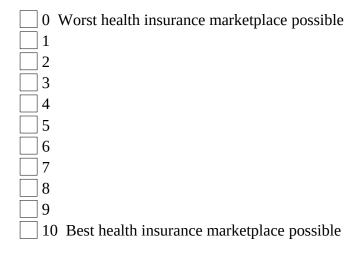
¹☐ Never
²☐ Sometimes
³☐ Usually
⁴☐ Always
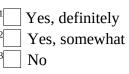
Overall Rating of Your Health Insurance Marketplace

66. Using any number from 0 to 10, where 0 is the worst health insurance marketplace possible and 10 is the best health insurance marketplace possible, what number would you use to rate your {INSERT MARKETPLACE NAME} since October 1st? **(GR/HP5-AM-m26)**

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

☐ 0 Worst health insurance marketplace possible
☐ 1
☐ 2
☐ 3
☐ 4
☐ 5
☐ 6
☐ 7
☐ 8
☐ 9
☐ 10 Best health insurance marketplace possible

67. Would you recommend the {INSERT MARKETPLACE NAME} to your friends and family? **(GR/CI1/H-m22)**

1 ☐ Yes, definitely
2 ☐ Yes, somewhat
3 ☐ No

About You

68. In general, how would you rate your overall health? **(CM/HP5-AM-27)**

1 ☐ Excellent
2 ☐ Very good
3 ☐ Good
4 ☐ Fair
5 ☐ Poor

69. In general, how would you rate your overall **mental or emotional** health? **(RC/HP5-AM-28)**

1 ☐ Excellent
2 ☐ Very good
3 ☐ Good
4 ☐ Fair
5 ☐ Poor

70.
Since October 1ˢᵗ, did you get health care 3 or more times for the same condition or problem? **(RC/HP5-AM-29)**

¹ ☐ Yes
² ☐ No → **If No, go to #72**

71.
Is this a condition or problem that has lasted for at least 3 months? Do **not** include pregnancy or menopause. **(RC/HP5-AM-30)**

¹ ☐ Yes
² ☐ No

72.
Do you now need or take medicine prescribed by a doctor? Do **not** include birth control. **(RC/HP5-AM-31)**

¹ ☐ Yes
² ☐ No → **If No, go to #74**

73.
Is this medicine to treat a condition that has lasted for at least 3 months? Do **not** include pregnancy or menopause. **(RC/HP5-AM-32)**

¹ ☐ Yes
² ☐ No

74. Are you deaf or do you have serious difficulty hearing? **(RC/OMB60/ACS-P-17a, OMH-4302-5)**

¹ ☐ Yes
² ☐ No

75. Are you blind or do you have serious difficulty seeing, even when wearing glasses? **(RC/OMB60/ACS-P-17b, OMH-4302-5)**

¹ ☐ Yes
² ☐ No

76. Because of a physical, mental, or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions? **(RC/OMB60/ACS-P-18a, OMH-4302-5)**

¹ ☐ Yes
² ☐ No

77. Do you have serious difficulty walking or climbing stairs? **(RC/OMB60/ACS-P-18b, OMH-4302-5)**

¹ ☐ Yes
² ☐ No

78. Because of a physical, mental, or emotional condition, do you have difficulty dressing or bathing? **(RC/OMB60/ACS-P-18c, OMH-4302-5)**
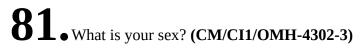
¹ ☐ Yes
² ☐ No

79. Because of a physical, mental, or emotional condition, do you have difficulty doing errands alone such as visiting a doctor's office or shopping? **(RC/OMB60/ACS-P-19, OMH-4302-5)**

¹ ☐ Yes
² ☐ No

80. What is your age? **(CM/HP5-AM-33)**

¹ ☐ 18 to 24 years
² ☐ 25 to 34
³ ☐ 35 to 44
⁴ ☐ 45 to 54
⁵ ☐ 55 to 64
⁶ ☐ 65 to 74
⁷ ☐ 75 or older

81. What is your sex? **(CM/CI1/OMH-4302-3)**

¹ ☐ Male
² ☐ Female

82.

What is the highest grade or level of school that you have completed? **(CM/HP5-AM-35)**

¹ ☐ 8th grade or less
² ☐ Some high school, but did not graduate
³ ☐ High school graduate or GED
⁴ ☐ Some college or 2-year degree
⁵ ☐ 4-year college graduate
⁶ ☐ More than 4-year college degree

83. What best describes your employment status? Mark only ONE. **(RC/OMB60/NHBS-DM6)**

¹ ☐ Employed full-time
² ☐ Employed part-time
³ ☐ A homemaker
⁴ ☐ A full-time student
⁵ ☐ Retired
⁶ ☐ Unable to work for health reasons
⁷ ☐ Unemployed
⁸ ☐ Other

84. Are you Hispanic, Latino/a, or Spanish origin? **(RC/OMB60/M-ACO-77)**

¹ ☐ Yes, Hispanic, Latino/a, or Spanish origin
² ☐ No, not of Hispanic, Latino/a, or Spanish origin → **If No, go to #86**

85. Which group best describes you? **(RC/OMB60/M-ACO-78)**

¹ ☐ Mexican, Mexican American, Chicano
² ☐ Puerto Rican
³ ☐ Cuban
⁴ ☐ Another Hispanic, Latino, or Spanish Origin

86. What is your race? *Mark one or more*. **(RC/CI1/OMH-4302-2)**

¹ ☐ White
² ☐ Black or African American
³ ☐ American Indian or Alaska Native
⁴ ☐ Asian Indian
⁵ ☐ Chinese
⁶ ☐ Filipino

⁷ ☐ Japanese
⁸ ☐ Korean
⁹ ☐ Vietnamese
¹⁰ ☐ Other Asian
¹¹ ☐ Native Hawaiian
¹² ☐ Guamanian or Chamorro
¹³ ☐ Samoan
¹⁴ ☐ Other Pacific Islander

87. Are you eligible to get health services from an Indian Health Service, tribal, or urban Indian health program? (RC/OMB30)

¹ ☐ Yes
² ☐ No → **If No, go to #89**
³ ☐ Don't Know → **If Don't Know, go to #89**

88. Did you ever get health services from an Indian Health Service, tribal, or urban Indian health program? (RC/OMB30)

¹ ☐ Yes
² ☐ No

89. What is your preferred language? **(RC,CuC/T,C,OMB60/ CG2-AS-CU22)**

¹ ☐ English → **If English, go to #91**
² ☐ Spanish
³ ☐ Chinese
⁴ ☐ Other
Please specify: _____

90. How well do you speak English? **(RC,CuC/T,C,OMB60/OMH-4302-4)**

¹ ☐ Very well
² ☐ Well
³ ☐ Not well
⁴ ☐ Not at all

91. Did you have health insurance in the United States at any time between January 1st and December 31st, 2013? **(RC/T,C)**

AIR
AMERICAN INSTITUTES FOR RESEARCH®

¹☐ Yes
²☐ No

92. How confident are you that you understand health insurance terms? **(RC/OMB30)**

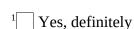¹☐ Not at all confident
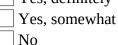²☐ Slightly confident
³☐ Moderately confident
⁴☐ Very confident

93. Do you feel comfortable using the internet through a computer, tablet, or smart phone? **(RC/C)**

¹☐ Yes, definitely
²☐ Yes, somewhat
³☐ No

94. Did someone help you complete this survey? (**RC/HP5-AM-38**)

¹☐ Yes
²☐ No →**Thank you. Please return the completed survey in the postage-paid envelope.**

95. How did that person help you? *Mark one or more*. **(RC/HP5-AM-39)**

¹☐ Read the questions to me
²☐ Wrote down the answers I gave
³☐ Answered the questions for me
⁴☐ Translated the questions into my language
⁵☐ Helped in some other way

*Please print:*_____

_____

Thank you.
Please return the completed survey in the postage-paid envelope.

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

About American Institutes for Research

Established in 1946, with headquarters in
Washington, D.C., American Institutes for Research
(AIR) is an independent, nonpartisan, not-for-profit
organization that conducts behavioral and social
science research and delivers technical assistance both
domestically and internationally. As one of the largest
behavioral and social science research organizations
in the world, AIR is committed to empowering
communities and institutions with innovative
solutions to the most critical challenges in education,
health, workforce, and international development.