



Preparing Data from CAHPS® Surveys for Analysis

Contents

Introduction.....	1
Creating Your Analysis Dataset.....	1
Task 1: Code and Enter the Data.....	1
Task 2: Conduct an Audit	3
Task 3: Identify and Exclude Ineligible Cases.....	3
Task 4: Clean the Data	4
Task 5: Identify Complete and Partial Completed Surveys for Inclusion.....	5
Task 6: Recoding Variables for Analysis.....	6
Examining Your Data With Descriptive Statistics.....	8
Identifying the Level of Analysis and Reporting.....	8
Calculating Frequencies.....	9
Calculating Top Box and Other Proportional Scores.....	9
Calculating Mean or Average Scores.....	12

Introduction

This document explains how to transform the raw data from the CAHPS surveys into data that can be analyzed using either the CAHPS Analysis Program for SAS (known as the CAHPS macro) or other statistical software. This document also provides an overview of how to compute frequencies, top box scores, and other proportional scores (e.g., % Always, % Usually, and % Never and Sometimes).

Comprehensive information about using the CAHPS macro once the data are prepared for analysis is provided in [Instructions for Analyzing Data from CAHPS Surveys](#) (PDF).

Creating Your Analysis Dataset

Before conducting analyses with your data, you need to carry out several tasks to prepare the data:

- Task 1: Code and enter the data.
- Task 2: Conduct an audit.
- Task 3: Identify and exclude ineligible cases.
- Task 4: Clean the data.
- Task 5: Identify and include only complete and partial completed surveys.
- Task 6: Recode variables for analysis.

Task 1: Code and Enter the Data

The first task is to enter the raw data into an electronic data file where each row or case in the data file represents a returned survey and each column represents a question. The responses to each question must use the code numbers, or precodes, contained in the questionnaires. Table 1 below shows examples of recommended coding when entering the data into an electronic data file.

Table 1. Sample coding of responses

Response	Code
Actual response	Use the corresponding survey precode indicated beside the response option. Precodes typically begin with 1 and number consecutively for each response option.
Item was appropriately skipped	Code as special missing for statistical software packages (e.g., “.S” for SAS). If using spreadsheet software, you can code as missing or 7 (or 77 if precodes exceed 7) if you want to track this type of response.
Item showed more than one response option completed when only one is appropriate (i.e., multiple marks)	Code as special missing for statistical software packages (e.g., “.H” for SAS). If using spreadsheet software, you can code as missing or 8 (or 88 if precodes exceed 8) if you want to track this type of response.
Item was left blank	Code as missing (. or blank).

The method you use to enter the data will depend on the method used to collect the data (e.g., whether questionnaires require data entry, can be scanned by a computer using optical scan forms, or are coded automatically from a Web survey).

Mailed questionnaires. If you use optical scan forms for your mailed questionnaires, the scanning equipment automatically enters the data into an electronic data file. The programs for reading in the optical scan forms should be programmed to capture the coding as outlined in Table 1.

If you do not use optical scan forms, the mail questionnaires are designed for direct data entry. The coding specialist or data entry personnel should first review the questionnaires to see whether the responses are legible and if there were mistakes in completing the questionnaire (e.g., multiple selections chosen). If it is unclear which answer the respondent selected (e.g., the respondent's pencil mark does not neatly fit within a single answer category, or two responses are marked), then the coding specialist will have to decide which response the respondent intended. If the intent is not readily apparent, the coding specialist should code the answer as missing or multiple marks, as indicated in Table 1. To ensure quality, two data entry specialists should enter the answers from paper-and-pencil questionnaires. Their results should then be compared to identify and correct data entry errors.

Telephone survey. If you use a CATI system for a telephone survey, an interviewer enters data directly into a data file that has already been programmed to refuse invalid responses. If interviewers use paper questionnaires to record answers given in a telephone interview instead, the process for coding and data entry is the same as for the standard paper version of the mailed questionnaires.

Web survey. Finally, if you have programmed the CAHPS instrument into a Web survey, issues with skip pattern errors and multiple marks will not be possible. The programmer should code the responses as outlined in Table 1.

Before moving on to Task 2, preserve the original data file created when the raw survey responses were entered. Many interim files will be created along the way. Any changes and corrections during the cleaning and data preparation phase should be made on duplicate files. There are three reasons to work with a duplicate file:

- The original data file is an important component of the complete record of the project.
- Having an original file will allow you to correct any data errors made during the cleaning process.
- The existence of an original file is critical if someone wants to go back later and conduct other analyses or tests, such as extent of error tests or tests of skip patterns.

Entering Data for Multiple Survey Versions

If you are entering data for multiple versions of a survey, construct a separate data file for each version of the survey. A single data file should not contain data from different survey instruments. For example, do not include responses to the Adult and Child versions of the CAHPS Clinician & Group Survey in the same data file.

It should be noted that the CAHPS Analysis Program can analyze data from two survey versions if needed. However, unless your research questions necessitate combining the responses to two survey versions into one file, it is advisable to keep them separate. If you decide to include more than one version of the survey in a data file, be sure to indicate the survey version in each row.

Entering Data for Multiple Entities

Any analysis of CAHPS data is intended to assess, compare, and report on some type of entity or reporting unit. Examples of entities include health plans, hospitals, medical groups, clinics, sites of care, and individual physicians. If you are entering survey data for more than one reporting unit, then you also need to include two variables to identify the entity:

- An identification number (e.g., 123) for each reporting unit (e.g., a variable called “PlanID”)
- A text descriptor (e.g., Plan A) for each reporting unit (e.g., a variable called “Plan”)

If you want to report data for different types of entities, you will need to create variables for each type (e.g., one for clinic, one for medical group).

Submitting Survey Data to the CAHPS Database

If you are interested in submitting data to the CAHPS Database, please refer to the data file specifications for the CAHPS Health Plan Survey or the CAHPS Clinician & Group Survey:

- [CAHPS Health Plan Data Submission System](#)
- [CAHPS Clinician & Group Data Submission System](#)

Task 2: Conduct an Audit

If survey data are collected using paper questionnaires (whether mailed or administered by telephone), the second task is to audit a small random sample of the entered data by comparing hard-copy forms with the results of data entry. This step enables you to catch any systematic errors that show up consistently across all questionnaires. For example, if the optical scanning program was incorrect, Question 5 may be entered in the file where Question 6 was supposed to be. An audit will also identify possibly coding errors. If you find errors in the first subset of data, you may wish to expand the sample of entered data until no issues are found. Conducting the audit immediately after data entry allows you to catch errors before proceeding with data cleaning, recoding, and analyses.

For Web surveys, an audit should be conducted *before* launching the Web survey by entering test data and verifying that the Web survey is capturing the data correctly. Since there is no way to verify accuracy at the end by comparing entries to hard copy questionnaires, pre-testing is imperative to make sure the data output from the Web survey is correct.

Task 3: Identify and Exclude Ineligible Cases

Several situations render a case ineligible for analysis. One common situation that should be coded as ineligible is when the respondent reports that he or she has not visited the sampled entity (e.g., a physician or medical group). This would be indicated by a “no” response to Question 1 (e.g., “Our records show

that you got care from the provider named below in the last 6 months. Is that right?"). Other situations that should be coded as ineligible for analysis include the following, with exceptions noted:

- **Deceased.** If the individual has died between the visit(s) and receipt of the questionnaire, the case should be coded as ineligible.
- **Language barrier.** If the potential respondent has a language barrier, the instrument is not available in the respondent's preferred language, and no one was available to translate the questions for the respondent, the case should be coded as ineligible.
- **Proxy.** In some settings of care it may be acceptable for someone to answer the questions on behalf of the target respondent (proxy) and be included in the analysis. In these situations, a case-mix variable indicating whether the data were self-reported, reported with assistance, or reported on behalf of the target respondent may be included when analyzing the data. In most situations, however, proxy data should not be included in the analysis or in the numerator when calculating the survey response rate.
- **Child.** Pediatric patients (under 18 years old) are ineligible from the adult version of a survey unless the child is emancipated.

Task 4: Clean the Data

In many cases, the data set you have created will have imperfections even after pre-testing programs and conducting an audit. Please take the following steps to check for and fix these imperfections before reporting any results. The cleaning of data is less imperative with Web surveys and optical scanning because most of these problems cannot happen with these data collection methods. Nevertheless, one should ensure against these problems in every mode.

Check for out-of-range values

Out-of-range responses occur when respondents provide inappropriate responses for a particular question. For example, if the valid response choices for a question are 0 or 1, a value of 2 would be out of range. To detect out-of-range values, run the data through a computer program that generates a report showing the item distributions or frequencies. If you find a value that is impossible (or unlikely) given the response options, review the questionnaire and revise the data:

- If the value was due to a data entry error, fix it.
- If the value truly was an out-of-range response, set the out-of-range value to "missing," which drops it from the data analysis for that particular question.

Carefully document the results from this review process, including any changes to the data set. Maintain an audit trail so it is possible to go back to the original data file.

Check for skip pattern problems

Response inconsistencies generally arise when a respondent misunderstands a question or does not successfully follow instructions to skip questions. An example of a response inconsistency would be if a respondent answered that he or she had no visits to the provider in the past 6 months, but then answered follow-up questions about visits in the past 6 months. If there are inconsistencies between the response to

the screener question and the following question's response, assume the screener response is correct. This forward-cleaning approach requires the following cleaning steps:

- If the respondent answered the screener question such that the subsequent dependent items should be missing, but the items are answered, the subsequent dependent items should be set to missing.
- If the screener question is left blank, and the subsequent dependent items are answered, the responses are retained and the screener question should remain as missing.

Check for duplicates

Duplicates can occur if the respondent responds via two data collection modes (e.g., Web and telephone, or Web and mailed questionnaire) and one mode arrives at the same time or soon after the other and the case slips through the receipt control system, or if there are errors in data entry (the same questionnaire is entered or scanned twice). If two questionnaires are received for the same respondent, keep the first questionnaire that is received and remove the duplicate unless the questionnaire received later has substantially more completed items than the first questionnaire.

Task 5: Identify Complete and Partial Completed Surveys for Inclusion

Once the data are cleaned, the next step is to code questionnaires as complete, partially complete, or incomplete and then remove the incomplete cases from the data file. Only complete and partially completed questionnaires should be retained for the analysis dataset.

- **Complete questionnaire:** A questionnaire is considered complete if responses are available for at least half of the key survey items and at least one reportable item.
- **Partially completed questionnaire:** A questionnaire is considered partially complete if responses are available for at least one reportable item, but less than half of the key items.
- **Incomplete questionnaire:** A questionnaire is incomplete if the individual did not answer at least one reportable item.

Key items are the survey questions that all respondents should answer, including:

- Questions confirming eligibility for the survey.
- The screeners for the questions included in the core composites measures.
- The primary rating question.
- Demographic and other background items.

Reportable items are the questions included in the composite and rating measures.

Key and Reportable Items in the CAHPS Clinician & Group Survey and the CAHPS Health Plan Survey

For a list of the key items, please refer to the guidance on fielding those surveys:

- *Fielding the CAHPS Clinician & Group Survey* (To download, go to [Get the Clinician & Group Survey and Instructions.](#))
- *Fielding the CAHPS Health Plan Survey* (To download, go to [Get the Health Plan Survey and Instructions.](#))

For a list of the reportable items, visit:

- [CAHPS Clinician & Group Survey Measures](#) (version 3.0 and 2.0)
- [CAHPS Health Plan Survey Measures](#) (versions 5.0 and 4.0)

Task 6: Recoding Variables for Analysis

This section outlines the types of variables found in the CAHPS surveys and discusses what variables may need to be recoded for analysis.

Types of Variables in the CAHPS Surveys

The following are the variable types for most of the core items in CAHPS surveys:

Table 2. Sample Types of Variables in CAHPS Surveys

Variable type	Sample Scale	Min – Max Response Values
Dichotomous	1= Yes 2= No	0 – 1
Global rating	0 = Worst to 10 = Best	0 – 10
3-point response scale	1 = Yes, definitely 2 = Yes, somewhat 3 = No	1 – 3
4-point response scale	1= Never 2 = Sometimes 3 = Usually 4 = Always	1 – 4

To facilitate analyses in most software packages, the variables used to compute CAHPS reporting measures, the case-mix adjustment variables, and any variables used to identify the survey version should be numeric.

Recoding Variables

Although not required, it is best to recode certain variables for ease in analysis, reporting, and interpretation. When recoding variables, always save the original version of that variable. The types of precodes (e.g., see Tables 1 and 2) that require recoding are presented below.

Dichotomous Variables. For easier interpretation of the results, the “positive” response should have the highest value. Raw data for this type of variable will need to be recoded as the precodes in CAHPS surveys typically set the values of the responses to 1=Yes and 2=No rather than 0 and 1 where 1=Yes or the highest category. Below is sample SAS code for recoding; the logic can be applied in any software package.

Sample SAS recode where Q05 is a Yes/No item with logic to recode from 1 = Yes and 2 = No to 0 = No:

- `Q05_Orig = Q05; If q05 = 2 then q05 =0;`

Variables with Three-Point Response Scales. In some cases, reverse coding may be necessary to ensure that the most positive response has the highest value—for example, where “Yes, definitely” is the most positive response. The precodes have the most positive category as the lowest value. Below is sample SAS code for recoding; the logic can be applied in any software package.

Sample SAS recode where Q17 uses a 3-point response scale where 1 = Yes, definitely, and 3 = No with logic to recode to 3 = Yes, definitely, and 1 = No.

- `Q17_Orig = Q17; If q17 in (1,2,3) then q17 = 4 - q17; else q17 = .;`

Negatively Worded Items. Reverse coding is also required for some of the supplemental items where the question is negatively worded – that is, responding negatively to the question (e.g., Never) is the positive response. For example, in the Health Literacy (HL) Item Set for Clinicians & Groups, one question asks: In the last 6 months, how often did this provider use medical words you did not understand? The responses to that question would need to be reverse coded so that the most positive response (Never) would be 4. Please review any supplemental items you use to determine which ones might need reverse coding for ease in interpretation.

Sample SAS recode where HL2 is a 4-point response item where 1 = Never and 4 = Always with logic to recode 4 = Never and 1 = Always.

- `HL02_Orig = HL02; If HL2 in (1,2,3,4) then HL2 = 5 – HL2; else HL2 = .;`

Setting Skip Pattern and Multiple Marks to Missing

If you coded the multiple marks and skip patterns with numeric values rather than special missing as per Table 1, you will need to consider setting them to missing for analyses. That is, if you try to calculate mean scores and include a 7 or 8 for the skipped or multiple marks responses, these will be counted in your mean even though they should not be. If your software package is not able to exclude these responses, set them to missing.

Examining Your Data With Descriptive Statistics

This section outlines several ways you can examine your CAHPS survey data. It provides guidance on calculating frequencies, ad means, and top box and other proportional scores, all of which can be done with a variety of software programs, such as Excel, SPSS, R, or SAS. To support survey users in [analyzing CAHPS survey data](#), AHRQ offers access to the CAHPS Analysis Program for SAS – often referred to as the CAHPS macro – which is designed to generate statistically valid comparisons of performance (learn more in the box below).

Overview of the CAHPS Analysis Program

The CAHPS Analysis Program was developed by AHRQ’s CAHPS Consortium to generate output that can be used to calculate the performance of various entities—such as medical groups, individual physicians, health plans, hospitals, counties, States, regions, or any other appropriate units of measure—to show how one entity’s performance compares to the overall performance of all other entities. The program generates both text output and SAS data sets that provide information on the scores for global ratings, composite measures, and individual items.

The CAHPS Analysis Program automatically adjusts the survey data for respondent age, education, and general and mental health status. Case-mix adjustment makes it more likely that reported differences are due to real differences in performance, rather than differences in the characteristics of enrollees or patients.

Consider using the CAHPS Analysis Program if:

- You are including more than 2-3 entities in your analysis.
- You would like to make statistically significant comparisons between entities.
- You want to case-mix adjust your scores.
- You want to weight the items differentially.

To determine whether using the CAHPS Analysis Program would benefit you, review the “Pre-Analysis Decisions” in [Instructions for Analyzing Data from CAHPS Surveys](#) (PDF).

Identifying the Level of Analysis and Reporting

CAHPS survey data are often collected at three levels: respondents, entities (e.g., medical groups, hospitals, health plans), and groups (e.g., health systems, sponsors, state organization). Therefore, the results can also be examined at each of these levels.

- **Respondent level:** A respondent is defined as an individual who has completed or partially completed a CAHPS survey. Respondent-level survey results are calculated across all respondents in the database. If the dataset includes multiple entities, respondent-level results would ignore the association of the respondent with a given entity.
- **Entity level:** Entity level results are calculated across respondents within each given entity. When results are calculated this way, and there is more than one entity, the results can be case-mix adjusted to account for different patient characteristics within each entity that might affect scores.

- Group level:** Group-level results are calculated across respondents within a given group. When there is more than one group, results can be case-mix adjusted to account for different patient characteristics within each group that might affect scores.

Calculating Frequencies

Frequencies indicate the number and percentage of respondents answering each response option for the items on the survey. The frequencies can be calculated in two ways: excluding missing values from the calculation of percentages or including them. See Tables 3 and 4 for sample output showing these two options.

Table 3. Sample Frequency Display With Missing Values Excluded

Item	Statistic	Never	Sometimes	Usually	Always (Top Box)	Total
Got appointment for urgent care as soon as needed	N	5,617	12,673	36,113	114,123	168,526
	Percent	3%	8%	21%	68%	100%

Table 4. Sample Frequency Display With Missing Values Included

Item	Statistic	Never	Sometimes	Usually	Always (Top Box)	Missing	Total
Got appointment for urgent care as soon as needed	N	5,617	12,673	36,113	114,123	250,375	418,901
	Percent	1%	3%	9%	27%	60%	100%

The decision to include or exclude the missing response depends on your research question or personal reporting preferences. For questions based on skip patterns, it may be more useful to exclude the missing values, so the frequency results show only those respondents who should have answered the question. As a data quality check, it also may be helpful to visualize how many respondents were excluded from the question.

Calculating Top Box and Other Proportional Scores

As noted, CAHPS surveys use several different response scales. Table 5 displays most of the response scales used in CAHPS surveys and shows how the options can be categorized into three proportions. Collapsing the response options in this way allows for easier reporting and interpretation of scores.

Top box scores are generally calculated as the percentage of respondents who chose the most positive option(s) on a given response scale (e.g., “Yes, definitely” or “Always”). The main advantage of using top box scores in reporting is that it is easier to explain and interpret one score for providers, staff, and consumers. A disadvantage is that top box scores only provide information for one end of the response scale; users have no idea of the distribution of responses across the remaining options (e.g., were the remaining scores all very negative or in the middle?).

Table 5. Crosswalk of Proportional Scores and Response Scales

Response Scale Type	Lower Proportion (“Bottom Box”)	Middle Proportion	Top Box
Dichotomous	No	-	Yes
Global rating	0-6	7-8	9-10
3-point response scale	No	Yes, somewhat	Yes, definitely
4-point response scale	Never, Sometimes or Definitely no, Somewhat no	Usually or Somewhat yes	Always or Definitely yes

Calculating proportional scores for an individual survey item: A proportional score is calculated by aggregating results across respondents for the reporting unit or entity. For example, in a sample of 10 respondents, if 4 out of 10 respondents answered “Always” to a particular item, the top box score for that item would be 40 percent [i.e., $(4 \div 10) * 100 = 40\%$]. (Note: If you are reporting out across all respondents, ignoring the reporting unit or entity, the top box scores cannot be case-mix adjusted for patient characteristics since the reporting unit or entity is ignored.)

Calculating top box and other proportional scores for a composite measure: The scores for a composite measure are equal to the average or mean of the proportion of responses (excluding missing data) in each response category across the items in the composite. The following steps show how proportions for composite measures are calculated:

- Step 1 – Calculate the proportion of cases in each proportional score category for each question in a composite.
- Step 2 – Calculate the average proportion responding to each category across the questions in the composite.

As shown in Table 6, top box scores for composite measures are calculated by averaging the top box scores for the items within the composite. This methodology can be applied to any of the CAHPS composites. For example, the “Helpful, Courteous, and Respectful Office Staff” composite has two items. If the top box score for the first item is 85 percent and the second item is 95 percent, the composite score would be 90 percent (i.e., $[85 \text{ percent} + 95 \text{ percent}] \div 2 = 90 \text{ percent}$). The same method is used for the other proportional scores.

Table 6. Sample Calculation of Top Box and Proportional Scores

Survey: CAHPS Clinician and Group Survey 3.0

Composite Measure: Getting Timely Appointments, Care, and Information

Item in Composite	Response Scale	Lower Proportion Score (Never, Sometimes)	Middle Proportion Score (Usually)	Top Box Score (Always)
Got appointment for urgent care as soon as needed	Never, Sometimes, Usually, Always	12%	20%	68%
Got appointment for check-up or routine care as soon as needed		7%	21%	72%
Got answer to phone question during regular office hours on same day		16%	23%	61%
Composite Proportional Score		12% = (12% + 7% + 16%) / 3)	21% = (20% + 21% + 23%) / 3)	67% = (68% + 72% + 61%) / 3)

In the Child HCAHPS Survey, the “Involving Teens in Their Care” and “Preventing Mistakes and Helping You Report Concerns” composite measures have items on two different response scales (Always to Never and Yes, definitely to No). Table 7 illustrates how scores for these items can be combined to create proportional scores in this scenario.

Table 7. Sample Calculation of Top Box and Proportional Scores when Scales Differ

Survey: CAHPS Hospital Child Survey
Composite measure: Involving Teens in Their Care

Item in Composite	Response Scale	Lower Proportion (Never, Sometimes or No)	Middle Proportion (Usually or Yes, somewhat)	Top Box Score (Always or Yes, definitely)
Providers involved teen in discussions about care	Never, Sometimes, Usually, Always	10%	25%	65%
Provider asked teen about readiness to leave the hospital	Yes, definitely, Yes, somewhat, No	5%	40%	55%
Provider talked with teen about care after leaving the hospital	Yes, definitely, Yes, somewhat, No	15%	10%	75%
Composite Proportional Score	--	10% = (10% + 5% + 15%) / 3)	25% = (25% + 40% + 10%) / 3)	65% = (65% + 55% + 75%) / 3)

Missing Data. If an entity is missing an item from a composite measure for all respondents, the proportional score is still calculated on the remaining items, dividing by the number of items for which there are responses.

Each item in a composite is generally equally weighted so comparisons from year to year are easier to interpret. Another option is to weight the item based on the number of responses to each item. (The CAHPS Analysis Program can do this for you.)

Calculating Mean or Average Scores

The mean or average score for an item is calculated by taking the mean across all response categories. For example, in a sample of 10 respondents, if 4 out of 10 respondents answered “Always” or “4” on the response scale, 2 answered “Usually” or a “3”, 2 answered “Sometimes” or “2”, and “2” answered “Never” or “1,” the mean would be 2.8 out of 4 (4+4+4+4+3+3+2+2+1+1 / 10). The top box score for this measure would be 40%.

Mean scores offer the advantage of reflecting the full range of the response options. Two physician practices with the same top box score (40%) look identical. If one physician practice had the responses shown above while the other had only 4’s and 3’s, the practice with the 4’s and 3’s would have a higher mean score. The disadvantage of using the mean score is that it can be harder for providers, staff, and consumers to interpret.