Examining Service Quality and Data Quality in Telephone Interviews

Julia Coombs and Kelly Govern[1]
U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract
While the benefits and limitations of measuring service quality in telephone surveys have been studied, not much attention has been given to measuring data quality. The Coverage Followup (CFU) operation, a telephone survey that resolved coverage errors in the 2010 decennial census data, included two independent quality measures: one to evaluate service quality and one to evaluate data quality. Thus, CFU provided a valuable opportunity to assess the usefulness of monitoring both aspects. The Service Quality Assurance (SQA) program scored recorded interviews on interviewers' telephone courtesy and accuracy of data capture for question groups, while the Data Quality (DQ) program scored recorded interviews on interviewers' script fidelity and accuracy of data capture for certain selected survey questions. The subtle differences between SQA and DQ provide an interesting look at the close relationship between the effectiveness of an interviewer and the quality of the data collected in a telephone survey. This paper looks at the development and design of the two quality measures, compares the information collected from both, and details how real-time SQA and DQ results catalyzed operational changes.

Key Words: data quality, service quality, telephone interview, census

1. Background

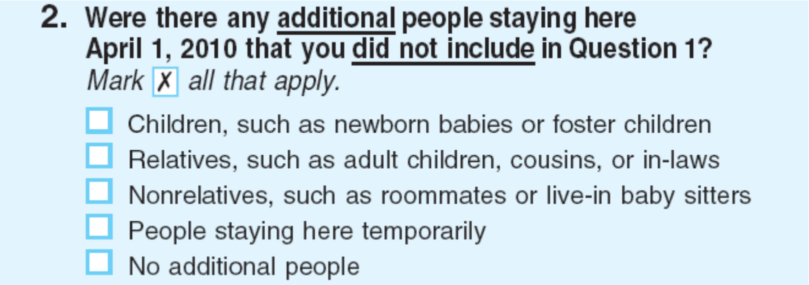
Ensuring that every person in the United States is counted once, only once, and in the right place is a vital goal of the decennial census. For many decades, the U.S Census Bureau has evaluated coverage in each census and documented that people are typically missed in the census. These people are referred to as census omissions. The Census Bureau has also documented that people are counted in the wrong place and found evidence that people are counted more than once during the census. Both of these errors are referred to as erroneous enumerations.

During the Coverage Followup (CFU) operation, computer-assisted telephone interviews were conducted with respondents to determine if changes should be made to their household roster as reported on their initial census return. The questions asked during the interview probed to identify if people were missed or counted in error and collected missing demographic data for all persons in the household. An interview was separated into modules, which were groupings of questions with similar purposes. Not all interviews entered every module, and not all questions within a module were necessarily asked if a module was entered. The modules were as follows:

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

- Modules A and P began the interview by verifying the household and identifying an eligible respondent. New interviews began in Module A, and partially completed interviews began in Module P.
- Module B was only entered if the respondent said that the incorrect household was reached in Module A. It attempted to collect information about the CFU household, and the interview could continue only if the respondent said that the CFU household had actually been reached.
- Module C verified the address of the household and collected missing tenure information. If the respondent reported that the address reached differed from the CFU address, the interview could continue only if the household had lived at the CFU address on Census Day or if the CFU address was a place the respondent sometimes lived or stayed.
- Module D removed duplicated or unknown roster members and probed for additional roster members.
- Module E asked if any household members moved out before Census Day.
- Module F probed for other places where household members sometimes lived or stayed.
- Module G collected missing demographic information.
- Module Q contained experimental questions.
- Module H ended the interview.

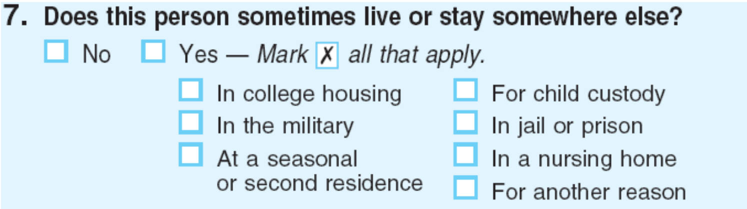
The cases selected for followup were household returns that had household sizes larger than the return could capture (a household size of six or more on the English Mailout/Mailback return, for example), returns with a reported household size that differed from the number of collected person records, returns with one of the coverage probes selected, returns that matched to administrative records data suggesting an omission, or returns with an indication of duplication based on computer matching. Some additional case types were sent to CFU as well for experimental purposes. The coverage probes were two questions on the initial census return; the undercount probe was a household-level question that asked for any missing persons, and the overcount probe was a person-level that asked if each household member sometimes lived or stayed in a group quarter. Figure 1shows the wording of the undercount coverage probe, and Figure 2 shows the wording of the overcount coverage probe.



2. Were there any additional people staying here April 1, 2010 that you did not include in Question 1?
Mark ☒ all that apply.

□ Children, such as newborn babies or foster children
□ Relatives, such as adult children, cousins, or in-laws
□ Nonrelatives, such as roommates or live-in baby sitters
□ People staying here temporarily
□ No additional people

Figure 1: Undercount Coverage Probe



7. Does this person sometimes live or stay somewhere else?

□ No □ Yes — *Mark* ☒ *all that apply.*

□ In college housing
□ In the military
□ At a seasonal or second residence

□ For child custody
□ In jail or prison
□ In a nursing home
□ For another reason

Figure 2: Overcount Coverage Probe

Preparing the 2010 CFU universe was an iterative process that took place over 11 waves while initial census returns were being processed. This practice minimized the time between the completion of the original return and the CFU interview, which in turn minimized recall bias. Cases selected for the CFU operation had to meet one of the coverage issues and had to be a non-group quarters return. Only one return per housing unit could be sent to CFU; thus, returns with more than one coverage issue were sent only once, and housing units with more than one return eligible for CFU (for example, housing units with both an initial and a replacement mailout census return) would have only one return sent to CFU.

The management of the 2010 CFU operation was contracted out as part of the Decennial Response Integration System contract (DRIS). Cases selected for CFU were sent to DRIS and loaded into a central dialer. Once an interviewer became available to conduct an interview, the dialer began dialing. If a respondent was available, then the dialer connected the respondent to the interviewer. However, if the dialer identified certain types of tones (busy signal, fax machine, no answer, etc), the dialer ended the call and dispositioned it appropriately. The dialer also had the capability to leave automated voice messages when an answering machine message was detected. Sequential dialing was implemented to allow for the calling of up to three phone numbers per case to increase the likelihood of reaching a respondent. Each phone number was dialed a minimum of three times before moving to the next phone number for the case, if an additional numbers existed. However, if a respondent requested a call back at an alternate phone number, that number was the only phone number dialed for the case. CFU was conducted from April 11, 2010 to August 14, 2010, in eleven call centers spread across the country. The interview was available in English, Spanish, Chinese, Korean, Vietnamese, Russian, and Telephone Typewriter (TTY)/Telecommunications Device for the Deaf (TDD).

To ensure that interviewers were collecting respondent data appropriately, two quality programs were included in the 2010 CFU operation. The Service Quality Assurance (SQA) program measured interviewer performance, and the Data Quality (DQ) program measured the accuracy of the data collected.

2. Limitations

One case in the SQA operation's data file had a score of zero for all scorecard items. This case was not included in any analysis contained in this paper.

Because the DQ operation's data file relied on a data source with some known error for some fields, it contained some error. Most notably, 84 cases had no call center identifier. These cases were included in overall totals in this paper but were omitted when the DQ universe was divided by call center.

Since not every question was asked in every interview, the scored frequency of some items differed. This could potentially limit some score comparisons. Most notably, one DQ item was scored only 76 times while another item was scored 35,675 times.

Also, the data are subject to error arising from a variety of sources.

3. Methods and Results[2]

Both operations drew interviews to review from a pool of calls randomly recorded by the eyeQ360 application. EyeQ360 was an audio-visual recording program that recorded both an interview's dialogue and the interviewer's screen randomly throughout the day. Thus, a SQA or DQ monitor could see interviewer navigation and response selection during an interview while listening to the conversation between the respondent and the interviewer. An interviewer could not detect if a particular interview was being recorded by eyeQ360 while on the call.

1.1 SQA Methodology

Each of the eleven call centers had its own SQA team that was dedicated to fulfilling the CFU requirement that each interviewer be monitored at least twice a day 95 percent of the time. The SQA team was comprised of SQA monitors who scored the calls, SQA supervisors who managed the SQA monitors, and an SQA manager who oversaw all SQA activities within that call center. Monitors would arbitrarily select recordings from a list of all non-scored recordings for an interviewer so that each interviewer had two scored calls per day. This decentralization ensured that monitors scored calls from interviewers in the same call center and allowed each call center freedom in achieving a high rate of monitoring two calls per interviewer per day. However, this decentralization also introduced a risk that SQA scores could vary by call center due to variance in how each call center understood the SQA rules. As a result, each call center's SQA manager participated in weekly calibration sessions with a central SQA team and Census Bureau representatives. During these calibration sessions, the scoring of a preselected call would be reviewed to ensure that the call center's implementation of an SQA rule was in line with the prescribed implementation. The calibration sessions also served as a forum for resolving a call center's SQA questions.

The interviewer's performance was evaluated using three criteria: Critical, Universal, and Code of Conduct. Seven of the eight Critical Criteria corresponded to the first seven modules in the CFU application; these capture the interviewer's ability to collect accurate and complete data. Each module-based Critical Criteria was scored only if the interview entered the module. Two additional modules existed in the CFU interview – Module Q, which contained experimental questions, and Module H, which ended the interview – but they were not specifically included in SQA's module-specific Critical Criteria.

The eighth Critical Criterion evaluated the interviewer's adherence to scripting and was scored in every call. The Critical Criteria were scored only *Pass* or *Fail,* and failing even one Critical Criterion meant that the call failed. A failed call resulted in immediate feedback and supervisory coaching for the interviewer.

[2] All data in this paper is taken from a larger forthcoming Census report, cited in the References section.

The Universal Criteria evaluated an interviewer's customer service soft skills, call handling efficiency, and behaviors that drove interviews to completion. There were seven Universal Criteria. Unlike Critical Criteria, the Universal Criteria were scored on a gradient scale of *Meets Standard*, *Needs Improvement*, or *Needs Significant Improvement*, as opposed to *Pass* or *Fail*. Scoring a *Needs Significant Improvement* on any of the Universal Criteria—even on every one of them— did not fail the call. Coaching may have still occurred, but the call was not considered to have failed. All Universal Criteria were scored in every call.

Code of Conduct violations included behaviors such as the use of profanity, disconnecting the caller, avoiding or manipulating the call, or any other behaviors as identified by the call center's Code of Conduct policy. Similar to the Critical Criteria, Code of Conduct was either scored as a *Pass* or a *Fail*. Any Code of Conduct violation resulted in a score of zero for the overall evaluation score, and the call was considered to have failed. An interviewer that failed a call for a Code of Conduct violation was given immediate coaching or possible disciplinary action.

The SQA score of a call was calculated using the following formula:

$$\textit{SQA Score} = \frac{\textit{Number of Earned Points} * 100}{\textit{Number of Possible Points}}$$

The number of earned points was the sum of points given for each item based on the SQA scoring standards. A call failed if the SQA score was below 94 percent.

While SQA scores were reported to interviewers to directly improve performance, scores were also monitored daily across the operation. Doing so allowed the CFU team to track interviewing difficulties and to release refresher training and job aids when necessary to improve scores.

The final overall score for the calls scored in the 2010 CFU operation was 99.0 percent, which was more than the required threshold value of 97.0. Interviewers were monitored twice every day 99.7 percent of the time. Table 1 breaks down the SQA score by call center. The range of average SQA scores is within one percentage point; both Denver and Stockton had the highest average SQA score of 99.3 percent, and Kennesaw had the lowest average SQA score of 98.6 percent. The small range of scores could be attributed to the weekly calibration sessions that the SQA managers of each call center had with a central SQA team and Census Bureau representatives to ensure consistent scoring.

Table 1: Average SQA Score by Call Center

Call Center	*Average SQA Score*	*Number of Calls Scored in CFU*
Denver, Colorado	99.3%	27,940
Kennesaw, Georgia	98.6%	93,734
Lawrence, Kansas	99.0%	21,563
London, Kentucky	98.8%	77,968
Monticello, Kentucky	99.2%	40,974
Murray, Utah	99.1%	82,904
Ogden, Utah	99.2%	36,340
Phoenix, Arizona	98.9%	26,175
Sandy, UT (ACS) [3]	98.7%	15,406
Sandy, UT (Vangent)	99.0%	90,528
Stockton, California	99.3%	46,107
Overall	**99.0%**	**559,639**

[3] Two call centers were located in Sandy, Utah, and they are distinguished by including the name of the telephony subcontractor that ran the call center. ACS stands for Affiliated Computer Services, one of the telephony subcontractors.

Table 2 shows the SQA scores of the eight Critical Criteria for the scored calls. The seven module-specific criteria had similar scores, but the "Read scripts verbatim" criterion had a lower score than the others. This corresponds to observations made during the operation that noted some interviewers' tendency to amend or abridge the given script.

Table 2: Average SQA Score by Critical Criteria

Critical Criteria	*Average SQA Score (in Percent)*	*Percent of Monitored Cases With a Failure*
Module A	99.0	1.0
Module B	99.9	0.1
Module C	99.8	0.2
Module D	99.6	0.4
Module E	99.9	0.1
Module F	99.2	0.8
Module G	99.7	0.3
Read scripts verbatim	95.4	4.6

Table 3 shows the SQA scores of each Universal Criterion as well as the percent of scored cases that received a "Needs Improvement" or a "Needs Significant Improvement" score. All of the Universal Criteria had a score of over 96.9 percent. A slightly higher percentage of cases were scored as "Needs Significant Improvement" in the Universal Criterion "Effectively and efficiently navigate systems" than in other criteria.

Table 3: Average SQA Score by Universal Criteria

Universal Criteria	*Average SQA Score (in Percent)*	*Percent of Monitored Cases with a Needs Improvement Score*	*Percent of Monitored Cases with a Needs Significant Improvement Score*
Display courtesy and professionalism	97.5	2.8	0.8
Display enthusiasm and confidence	98.3	2.1	0.5
Provide accurate and complete information	98.1	2.1	0.7
Effectively control the call	97.1	3.4	0.9
Effectively use active listening and probing questions	97.3	2.9	0.9
Effectively and efficiently navigate systems	96.9	2.2	1.8
Appropriately document and disposition the call	97.9	2.3	0.8

Each call was also given a Code of Conduct score. Out of the 559,639 calls scored in SQA, 156 calls had a Code of Conduct failure.

Overall, 1.6 percent of the monitored calls failed. So, not only were scores high, but there were not many failed calls. SQA scores, while generally high, also increased over time, as seen in the figure in the Appendix. Some early dips can be attributed to the staggered call center opening schedule, but the scores trend upwards over time.

1.2 DQ Methodology

The DQ operation for CFU measured the accuracy of the CFU data collected through the telephone interviews.[4] Instead of being organized at the call center level, all scoring occurred at the operational level. A small group of monitors supervised by the DQ manager scored a small daily sample of English and Spanish recordings where the interview was completed within the call. The scores were used for monitoring trends and data quality at the operational level, and interviewers were never notified of scores.

Monitors evaluated eyeQ360 recordings of interviews from a sample of cases and focused on 15 critical questions to ensure script adherence and accurate data capture. The 15 critical questions were as follows:

- Is there anyone I've mentioned that you don't know?
- Who is the person(s) you don't know?
- Is your name correct?
- I'd like to make sure we are not missing anyone who lived or stayed here {fill address} on April 1, 2010. Other than the people we've already mentioned, were there: Any newborns or babies?
- Any other relatives who lived or stayed here?
- Anyone else who stayed here often?
- In Spring of 2010, was anyone attending college?
- Who was attending college?
- Where did {Name from who was attending college} stay while attending college?
- In April or May, did {fill "you" if person count=1, else "anyone"} stay somewhere else for an extended time or live part of the time at another residence?
- Who was staying elsewhere for an extended time during April or May?
- In April or May, where did you live or stay most of the time?
- Were you staying at {fill address} or at the other place on April 1, 2010?
- {Were you/Was Full Name} staying in any of those places on April 1, 2010?
- What was {fill Full Name's} age on April 1, 2010?

Each question was scored as *Accurate*, *Inaccurate*, *Uncertain*, or *Not Scored*, and a corresponding reason code to describe why a question was scored as such was always selected. As in the SQA operation, not every question was presented in every monitored interview. Table 4 shows the definitions of the reason codes within each score.

[4] The DQ methodology of the 2010 CFU operation was largely modeled on the methodology for measuring data quality in telephone interviews presented by Ryan King in (King, 2008).

Table 4: DQ Scorecard Options

Score	*Reason Code*	*Definition*
Accurate	Accurate	Interviewer read the critical question verbatim and the response matched the output
Inaccurate	Question not Read	Interviewer did not read the critical question
Inaccurate	No Match	Respondent's response did not match what the interviewer selected in the tool
Inaccurate	Data Error	Respondent's response matches what the interviewer selected but not the captured output in the data file
Uncertain	Not Read Verbatim	Critical question was not read verbatim
Uncertain	No Clear Answer	Critical question had a complex exchange between the interviewer and respondent where no clear answer was given by the respondent
Uncertain	Inaudible	Audio from either the interviewer or respondent was inaudible
Not Scored	Question Absent	Critical question was not included in the monitored recording

Each call scored in DQ was given a Quality Improvement Index (QII) score. The QII was calculated using the following formula:

$$\textit{QII Score} = \frac{\textit{Number of Accurate Critical Questions}}{\textit{(Number of Accurate Critical Questions + Number of Inaccurate Critical Questions)}}$$

Since the QII tracked changes over time and the SQA score was reported to the interviewers as a performance measure, the QII was not reported as a percent, but the SQA score was. The overall QII of scored calls was 0.994, and 11,583 calls were evaluated in the DQ operation. Since the scores were not reported to interviewers, there is no "fail" score. The high overall score implies that for most of the time in CFU, the interviewer, respondent, and instrument communicated successfully. Table 4 shows the QII scores by call center and over time. In this table, the time periods are based on each call center's start date, not on the operational start date. Scores were very high overall, and nearly every call center shows improvement over time. Monticello is the only call center with scores that show any downward movement, but the change is not large. The call centers have similar overall QII scores, which was expected because the same group of monitors scored all of the recordings across all call centers.

Table 5: Average QII by Call Center

Call Center	*Average QII*	*Number of Calls Monitored*
Denver, Colorado		
First Week	0.984	30
First Month	0.993	89
End of Operation	0.996	318
Kennesaw, Georgia		
First Week	0.975	281
First Month	0.988	746
End of Operation	0.994	2,147
Lawrence, Kansas		
First Week	0.995	589
First Month	0.997	946
End of Operation	0.997	1,136
London, Kentucky		
First Week	0.989	191
First Month	0.995	568
End of Operation	0.996	1,329
Monticello, Kentucky		
First Week	0.975	408
One Month	0.982	863
End of Operation	0.988	1,287
Murray, Utah		
First Week	0.996	72
First Month	0.993	302
End of Operation	0.997	1,427
Ogden, Utah		

First Week	0.993	14
First Month	0.993	103
End of Operation	0.998	427
Phoenix, Arizona		
First Week	0.988	550
First Month	0.990	929
End of Operation	0.991	1,161
Sandy, UT (Vangent)		
First Week	0.993	342
First Month	0.995	588
End of Operation	0.996	1,366
Sandy, UT (ACS)		
First Week	0.982	111
First Month	0.986	161
End of Operation	0.988	258
Stockton, California		
First Week	0.986	52
First Month	0.993	148
End of Operation	0.998	643
Overall		
First Week	0.987	2,640
First Month	0.991	5,360
End of Operation*	0.994	11,583

* Due to an error in the call detail record, the call center for 84 DQ cases could not be determined. These 84 cases are included in the overall number but are not included in the call center numbers. Therefore, the number of cases scored by call center may not sum to the number of cases scored overall.

Table 5 looks at the QII by critical question. Not all questions were asked in every interview, and the QII for a question includes only the cases where that question was scored as “Accurate” or “Inaccurate.” Most question scores were over 0.96, but the QII of the question asking for the name of the unrecognized roster member was below 0.90. This is likely due to interviews where the interviewer had to loop through Module D to delete duplicated or unknown persons. Interviewers sometimes had difficulty dropping duplicated persons, and they would sometimes not reread required scripted text, which would affect QII scores.

Table 6: Average QII by Critical Question

Critical Question	*Average QII*	*Frequency Scored*
Any unrecognized roster members	0.996	10,011
Name of unrecognized roster member	0.874	323
Respondent’s name is correct	0.980	1,588
Missing babies	0.999	11,501
Missing relatives	0.998	11,452
Missing people who stayed often	0.998	11,447
Anyone in college	0.998	5,865
Name of college student	0.981	2,476
College address	0.994	2,962
Anyone stay at another address	0.998	11,511
Name of person staying at other address	0.963	408
Lived at which address most of the time	0.991	4,274
Staying at which address on Census Day	1.000	76
Anyone stay in a group quarters	0.993	35,675
Age of added person on Census Day	0.980	965

4. Conclusion

While SQA and DQ both strove to evaluate quality, the two programs went about it in different ways. One reason for this was that the programs had different purposes. SQA focused on interviewers and directly impacted interviewers, while DQ evaluated the data and helped identify new areas of training. Also, the scale of the two projects differed. SQA strove to evaluate two interviewers per day, while DQ looked at a small percentage of daily, random, completed English or Spanish interviews. In addition, the SQA scoring method spanned the whole interview, but the DQ scoring method focused on fifteen questions in two modules.

Having both operations was useful at an operational level, though. Both allowed for real-time quality tracking that gave a great opportunity to improve the responses gathered in CFU. SQA’s Universal Criteria allowed for some tracking of interviewer performance, and DQ’s reason codes showed how the interviewer behaviors may have affected the data. Since all of DQ’s critical questions were in the two modules that addressed erroneous enumerations, any low SQA score trends in those two modules prompted a closer examination of DQ scoring for any issues with specific questions. Together, both operations lead to specialized training, which in turn ensured that CFU was delivering the best results it could.

Acknowledgements

The authors wish to acknowledge the whole CFU team, including those who worked at the Census Bureau and those who worked for the CFU contractor.

References

Govern, Kelly, Julia Coombs, and Robert Glorioso 2011. 2010 Census Coverage Followup (CFU) Assessment Report. To be released.

King, Ryan 2008. Measuring Data Quality for Telephone Interviewers. In *ASA Proceedings,* Section on Survey Research Methods. New Orleans, LA: American Association for Public Opinion Research.

Appendix

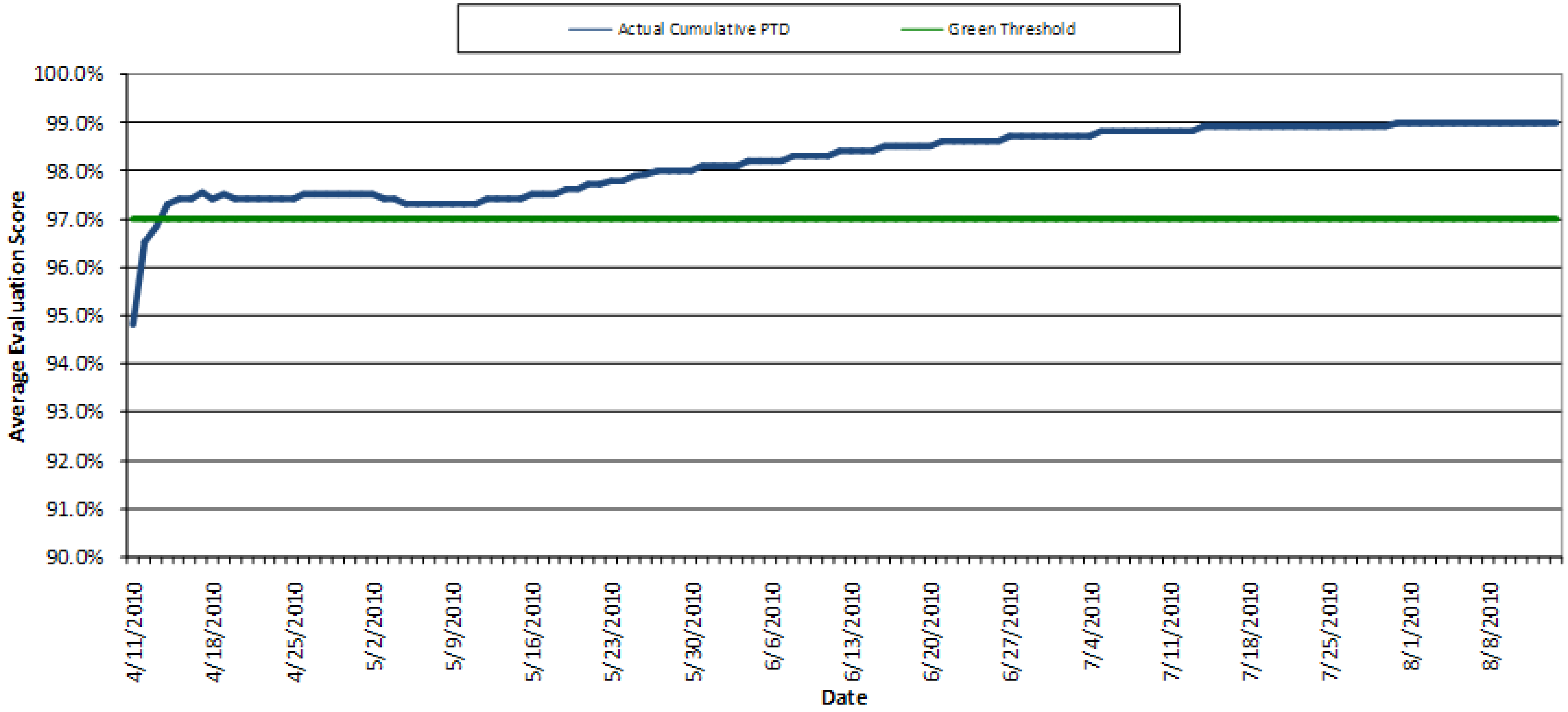


Figure: Actual cumulative SQA scores over the operational period. The green threshold is the target cumulative SQA score.