Customer Service Satisfaction Survey: Cognitive and Prototype Test

Kevin Cecco and Anthony J. Young, Internal Revenue Service

The Internal Revenue Service (IRS) is committed to becoming a more modern, customer-oriented agency. This requires developing performance measures that balance taxpayers' needs with the IRS's internal operational needs. One prong of our balanced performance measures is a Customer Satisfaction index. This index is being developed, in part, from surveys collected from taxpayers who had direct telephone contact with the IRS.

The Customer Service organization within the IRS currently has a manual customer satisfaction survey in place to gauge taxpayer opinions and perceptions. This survey is offered to a sample of taxpayers regarding taxpayer assistance or issue resolution on several IRS toll-free telephone numbers. In an attempt to interact more efficiently with taxpayers, the Service has decided to automate the process of conducting telephone customer satisfaction surveys. The Customer Service Satisfaction Survey (CSSS) application will replace the current manual survey. The automated telephone survey should be cost-effective and just as accurate if we can encourage taxpayers to use the system and not hang up prior to completing the survey.

Moving from the manual telephone survey to an automated survey, the IRS obtained the services of Andersen Consulting (AC) to complete a series of cognitive tests. The objective was to develop the most efficient automated survey that taxpayers would be willing to complete.

As part of the study, several areas within the IRS worked with AC to complete the following activities:

Expert Review—This expert review of the CSSS application used best practices in order to suggest revisions to improve usability of the scripts and identify problem areas for cognitive testing. Exploration was done to find published documentation regarding automated survey research techniques and practices.

Cognitive Testing—This portion of the study consisted of cognitive testing of the CSSS scripts using concurrent think-aloud procedures. Rather than using a simulated environment for the testing, actual callers to the Atlanta Call Site were asked to participate in cognitive testing after they completed their calls.

Rapid Prototype Study—The final portion of the study used a Voice Response Unit (VRU) which played different scripts (or scenarios) for a caller. The purpose was to gather data for different length scripts, different scales, and call types. Participants in the prototype tests were solicited by a group of customer service representatives (CSR's) who asked each taxpayer to participate in the survey. If taxpayers agreed, they were transferred to the prototype VRU application.

■ Results from the Expert Review

The automated script was revised more than ten times, based on listening to the script after recordings were made and on recommendations from past experience with automated survey scripts. The result was a very organized script, which was easy to use for callers. The script was then tested qualitatively and quantitatively with the Cognitive and Prototype tests.

■ Methodology and Results from Cognitive Testing

Cognitive testing was completed during the week of December 14-18, 1998, using telecommunication monitoring equipment installed at the Internal Revenue Service's New Carrollton Federal Building. The test included 25 taxpayers who phoned the IRS Atlanta Call Center for assistance. The IRS decided that the best possible test process would include real callers. The 25 participants were divided into two groups:

- Phase 1.--15 taxpayers were asked to think

aloud as the survey script was read to them. They completed the required survey actions using the keypad of a telephone. Once they completed the first phase, major issues were identified, and changes were made to the script.

- Phase 2.--10 taxpayers were asked to complete the survey, but their *think-aloud* responses were restricted to areas in which they had difficulties or confusion.

Two members of the AC staff completed the cognitive interviews. The first person simulated the VRU by reading the question and playing back the confirmation response to the caller. The second AC team member probed the caller and documented responses, opinions, and perceptions. Following the call, a post-survey interview was conducted to gather additional information. The process worked extremely well and was easily set up with minimal cost and effort.

■ Key Findings from Cognitive Testing

Table 1 summarizes the key findings resulting from cognitive testing. The four main points highlight differences that were significant between phases 1 and 2, as well as aspects of the automated survey that were changed from phase 1 through to phase 2. The findings, coupled with the corresponding results, allowed the IRS to understand the behavior of taxpayers and make changes that improve the efficiency of the survey.

Table 2 provides a summary of responses to a survey conducted following the cognitive interview for each taxpayer. The table shows different responses to several questions between phase 1 and phase 2 of the cognitive interviews. The data indicate a general trend of improvement in ease, willingness, and information to answer questions between the first and second phases of cognitive testing.

Note: These data, from each of the two groups of taxpayers, show the amount and percent difference between them. Each row of data is ranked from the largest difference to the smallest. The three areas with the greatest differences are shaded gray.

■ Methodology and Results from Prototype Tests

The purpose of the Prototype testing was to determine how response rates would vary, given the number

Table 1: Key Findings from Cognitive Testing

Finding #	Issue	Method	Result
1	Cognitive interviews allowed for a general improvement in specific questions found on the automated survey	Through the cognitive process, callers verbalized difficulty and confusion regarding the wording of several questions on the survey	Following Phase 1, certain questions were rephrased, while clearer instructions were prefaced before the questions.
2	Scaling responses to questions-Comparing the 1-4 Scale (i.e. very dissatisfied – very satisfied) to the 1-7 Scale (larger number identifies greater satisfaction)	Participants in Phase 1 were given both scales in answering questions in a randomized fashion. After completing the survey, the participants were asked which scale they preferred.	Post interview results revealed that ten of fourteen users (71.4%) preferred the 1-4 Scale.
3	Repeated instructions regarding the "type ahead" feature increased the usage of this feature in the second phase.	Participants in the second phase were given multiple instructions stressing the awareness of this feature. The "type ahead" instructions were only provided once during phase one.	Phase 1: 9 of 15 participants (60%) used "type ahead." Phase 2: 8 of 10 participants (80%) used "type ahead."
4	Use of "STAR" key (repeat question feature) diminished in Group 2.	Participants in both phases were given option of pressing the "STAR" key to repeat the prior question.	Phase 1: 7 of 15 participants (46.7%) used the "STAR" key to repeat one or more questions. Phase 2: 2 of 10 participants (20%) used the "STAR" key. Slight wording changes to questions, removal of vague language, and other minor system revisions probably led to this decrease in the usage of the "STAR" feature.

Table 2: Summary of Responses from Post-Cognitive Interview Survey

Interview Question	Score*		Improvement	
	Phase 1	Phase 2	Amount*	Percent
1. Overall Ease or Difficulty of This Survey	1.9	2.3	0.4	19
2. Willingness to Use This Automated Survey	2.3	2.6	0.3	14
4. Sufficient Information to Answer Questions	2.2	2.5	0.3	12
6. Ease of Understanding the Survey Instructions	2.9	3.0	0.1	2
7. Appropriateness of Survey for Participants' Knowledge and Experience	2.9	3.0	0.1	2
3. Ability to Do the Survey Correctly	2.9	2.9	0.0	0
8. Awareness of "Type Ahead" and Ability to Use It	N/A	2.9	N/A	N/A
Average Improvements (for questions with scores)	2.5	2.7	0.2	8

*A 3.0 scale where 3.0 is the highest score.

and type of questions on the automated telephone survey. To our knowledge, there is inconclusive documentation in the field relating to the optimal number of questions that should be included on an automated survey while still maintaining a respectable response rate. One belief is that an automated survey should not exceed about ten questions, because a caller may become impatient with the survey and simply terminate the call. Our study set out to determine how many questions could be included while still maintaining credible response rates.

For the non-tax season prototype test (conducted in December 1998), it was agreed to run scripts of various lengths from 8 to 30 questions in order to see what effect the length of survey had on user hang-up rates. Based on the objectives for the non-tax season prototype test, different scenarios were developed. For each call type, four different scripts were developed of different lengths. Each script was tested, first with 50 callers using the 1-4 scale, and then with 50 callers using the 1-7 scale. A scenario was defined as a test with a script of a certain length, using a certain scale, and consisting of a particular call type. Each scenario was tested with 50 callers. The prototype VRU application took care of switching from scenario to scenario as soon as 50 callers had been surveyed. Following the non-tax season prototype test, improvements were made to the script with the intent of collecting additional data during tax season.

The objective of the tax-season prototype test was to investigate two scenarios with similar attributes to those planned for the future pilot test in the summer of 1999. The first scenario used 20 questions for Account Call System (ACS) callers and 16 questions for toll-free callers. The second scenario had 14 questions for ACS callers and 12 questions for toll-free callers. Each scenario had 300 callers. However, there was no control of the blend of ACS and toll-free callers.

Based on the results of the cognitive interviews and the first phase of the prototype tests, it was decided to use a 1-4 response scale for the tax season test. The 1-4 scale was now somewhat different, however, in that it allowed one negative entry and three positive entries rather than two negative entries and two positive entries utilized during non-tax season testing. The wording of questions was done to determine the caller's satisfaction with the services provided.

Data from the first phase of the prototype test provided conflicting results. On the negative side, the initial transferring of taxpayers from Customer Service Representatives to Quality Reviewers revealed a rather low participation rate for the automated survey. Of the nearly 3,000 phone calls to CSR's, only about one-third of taxpayers agreed to be transferred from a CSR. This lower-than-expected participation rate was partially due to the CSR's not understanding or following the instructions properly when transferring taxpayers to the Qual-

ity Reviewer. Other telecommunication and data collection problems also hindered participation among taxpayers. Table 3 provides a quick overview of the limited success the IRS had during Phase 1 in transferring callers from CSR's to the automated survey.

Results from the Phase 1 Prototype Test summarized in Table 4 clearly show how hang-up rates gradually increase as the number of questions increase on the automated survey. The prototype test shows that most callers will complete the survey, but as the length of the survey increases, they tend to hang up at a higher rate. It would appear that the percentage of completed surveys remained credible through the 20-24 question range.

Table 5 summarizes the participation rate from the tax-season phase of the prototype test. The participation rate effectively doubled from Phase 1 to Phase 2 of the study. Participation rates during Phase 2 were more in line with what we expected compared to Phase 1. Additional field training and awareness of the survey could further improve the participation rate of the IRS automated customer satisfaction survey.

Table 6 summarizes hang-up rates for Phase 2 of the prototype test. In contrast to intuition, the hang-up rates for ACS calls decreased as the number of survey questions increased, while hang-up rates for toll-free calls, during Phase 2, increased as the number of survey questions increased. The nature of the call could be a possible explanation for the difference in rates between the two types of calls. ACS callers must identify themselves during the call, leading to a situation where taxpayers feel they should participate in the automated survey. On the other hand, toll-free callers do not always identify themselves during a call. Consequently, the toll-

Table 3: Phase 1 – Customer Service Representative Transfer to Automated Survey Analysis

Total Calls Gated	Calls Successfully Transferred	Participation Rate
2,953	880	31.9%

Table 4: Phase 1 of Prototype Test (Non-tax Season) – Hang-up Rates by Scenario

Scenario	Number of Questions	Call Type	Surveys		Hang-up Rate
			Transferred	Completed	
1	8	Toll-Free	100	90	10.0%
	9	ACS	98	85	13.3%
2	12	Toll-Free	47	32	31.9% *
	14	ACS	100	87	13.0%
3	20	Toll-Free	100	82	18.0%
	24	ACS	100	77	23.0%
4	26	Toll-Free	100	63	37.0%
	30	ACS	14	11	21.4% *

* Situations where computer malfunction or human error occurred

Table 5: Phase 2 – Participation Rates

Total Calls Gated	Calls Successfully Transferred	Participation Rate
1,174	762	64.9%

Table 6: Phase 2 of Prototype Test (Tax Season) – Hang-up Rates by Scenario

Scenario	Number of Questions	Call Type	Surveys Transferred	Surveys Completed	Hang-up Rate %
1	12	Toll-Free	226	183	19.0
	14	ACS	70	59	15.7
2	16	Toll-Free	227	159	30.0
	20	ACS	76	70	8.0

free caller might not be as persuaded to complete an automated survey. In any case, results from Phase 2 of the prototype test reveal an inconclusive picture. Additional data should be collected before making any clear statements about participation rates for the automated surveys.

■ General Recommendations and Conclusions

Based on the results of the entire CSSS Usability Research Study, it is recommended that a pilot test version of the CSSS application should:

- ❒ Be similar enough to the manual survey in order to correlate manual and automated survey data.
- ❒ Be configurable to allow elimination of questions so as to shorten the survey time and increase participation rates if needed.
- ❒ Use the 1-4 scale.
- ❒ Provide clear instructions regarding the ability to use "type-ahead".
- ❒ Provide prompts on the use of the "*" key until the user has made use the first time.
- ❒ Provide adequate length of time in the timeout values so that callers can use a telephone with touch-tone keys in the handset.
- ❒ Collect data on the use of the "9" response to support research into issues that cause this response to be used.
- ❒ Limit ability to add questions by providing placeholder questions that can be turned on after prompts are recorded.

The CSSS should also make use of the scenario that asks the largest number of questions and still maintains a credible response rate. From Phase 1, the scenario that best achieves this goal is Scenario 3, which asks 20 questions for non-ACS callers and 24 questions for ACS callers, while maintaining completion rates of 82 percent and 77 percent, respectively. From Phase 2, the preferred scenario is Scenario 1, which asks 12 questions for non-ACS callers and 14 questions for ACS callers, while maintaining completion rates of 81 percent and 84 percent, respectively. The plan for a summer 1999 pilot test is to use an automated survey similar to Scenario 2 of the second phase of the prototype report.