QUESTION DEVELOPMENT PROCESS FOR THE IRS TEST CALL SYSTEM

Jonathan Daniels and Mary Batcher, Internal Revenue Service

KEY WORDS: Quality improvement questionnaire design, integrated test call, survey system

The Internal Revenue Service (IRS) offers free telephone assistance throughout the United States to taxpayers to help them with questions about their personal income tax. This is a year-round service with seasonal peak volume during the period from January to mid-April. Since 1988, IRS has operated a program of test calls designed to assess the accuracy of the technical tax law advice given by the telephone assistance service--specifically, questions about individual income tax returns. This test call program -- the Integrated Test Call Survey System or ITCSS -- is based on an earlier General Accounting Office (GAO) test call program and has been developed and operated in cooperation with GAO [1-4].

This is the latest in a series of papers documenting that effort. This paper will focus on the procedures used to develop test questions for the 1992 ITCSS that would be similar to actual taxpayer questions in the nature of the inquiry and the phrasing of the question. After providing some background, we will describe the question development process and present some results. We will conclude with some future plans for this study.

BACKGROUND

The "integrated" portion of the title refers to the integration of data used to arrive at the ITCSS estimates, which are based on test calls, volume of taxpayer inquiries by tax law topic, and the volume of calls answered at each telephone answering site. The ITCSS has two major purposes:

- to provide Congress and the public with an overall measure of accuracy during the filing season; and
- to provide the call sites with detailed weekly feedback that they can use to assess the effect of improvement efforts and to target such efforts.

The test call system, itself, is described in [5]. Briefly stated here, it is a system of test calls placed to each of the 31 toll-free telephone assistance sites in the continental United States. Calls are placed from a central location in Washington, D.C., by a permanent staff, which included 9 test callers in 1992. For 1990 and 1991, the callers used a predetermined, scripted set of 42 test questions developed by IRS staff, with the concurrence of GAO. These test questions were classified by tax law category. In addition, two assistors in each call site categorized all of the inquiries they received from taxpayers into those same categories. These category volumes were then used for poststratification weighting of the accuracy rates. The major design components were thus, the call sites, the test callers, the test questions, and the tax law categories. These components and the basic design are described in greater detail in [6]. Our focus in this paper is the test questions and their development process.

The decision to use a measurement system based on test questions was guided by the need to strike a balance between measurement control and making the ITCSS reflect the actual interaction between a taxpayer seeking information and an IRS employee responding to the inquiry. Of measurement systems considered, the one that best reflects the actual interaction between caller and assistor is provided by monitoring live calls and assessing their accuracy. However, this monitoring of live calls does not afford a high degree of measurement control. The monitors would be required to make an instantaneous determination of accuracy, there would be no control over the type of tax law issue tested, and the result would be more subjectivity in the scoring than was acceptable for the public testimony use of the data.

The greatest degree of measurement control is provided by making test calls using scripted test questions, controlled calling procedures, and predetermined scoring guidelines. However, this measurement approach is more artificial and less reflective of the actual interaction between taxpayer and assistor. There is some risk that we are actually measuring something other than what goes on between taxpayer and assistor. There are several ways to minimize this risk. They include protecting the test questions from identification, developing questions that are similar to those that taxpayers actually ask, and training callers to ask questions in a style similar to an actual taxpayer call; i.e., in a

"natural" manner. Changes in the nature of test questions over the years of ITCSS operation and overall test reliability are described in [6].

TEST QUESTIONS

The basic format of the test questions consists of a script of the question itself, some background information to be provided if the assistor requests it, a set of possible probes that the assistor may ask, and a set of responses. The test caller works from a standard computer assisted telephone interview (CATI) system, which provides the text of the scheduled call, background information, and a list of possible probes and responses. The caller poses as a taxpayer in making the call and simply records which of the possible probes and responses shown on the screen were addressed by the assistor. Scoring of the call as correct or incorrect is accomplished by a computer algorithm, which combines probes and responses marked as present, to see whether they correspond to any of several possible correct combinations. The scoring is not made available to the test caller. Callers are not allowed to deviate from the script except to adjust for sex and age differences by saying they are calling for someone else.

The questions used in the ITCSS testing undergo a lengthy review and negotiated agreement process between IRS and GAO technicians and attorneys. This process has made the cost of adding questions to the ITCSS question bank very high, both in terms of dollars and human energy. There is, therefore, a real incentive to keeping good questions in the system as long as possible. This must be balanced against the need to keep the test questions current with what taxpayers are asking and to protect them from being used so frequently that they are identified. Given that background, it may not come as a surprise that the test questions remained the same in 1990 and 1991 or that there was a need to develop many new questions for the 1992 measurement. The remainder of this paper describes the development of new questions, some characteristics of the question bank, and changes to the question development process, resulting from what we learned from the 1992 question development and approval process.

QUESTION DEVELOPMENT

A goal of the ITCSS was to estimate the accuracy of assistors on the IRS toll-free tax information line. Therefore, to best reflect the accuracy of the IRS assistors, it was necessary to construct test questions similar to actual taxpayer inquiries. Two criteria were identified:

- the question set had to be distributed over the various tax law categories to mirror the actual call volume distributions; and
- the test questions had to be similar in content and style to actual taxpayer questions.

The question development process centered on these two vital requirements.

The first phase of the question development process was to derive the distribution of test questions over the tax law categories. This was done by using volume records from the district offices. Two assistors in each district office recorded the specific tax law categories of all of the telephone inquiries they received. These assistors recorded the category of each taxpayer's call they received by placing a "tick" mark on a data collection sheet. By combining the data from the district offices, a national distribution of taxpayer inquiries by tax law category was derived. This was used to determine the percentage distribution of test questions. For example, for 50 total test questions, if the ticker volume for a given category was 20 percent of the total volumes ticked, then ten test questions would be included in that category. A percentage was assigned to each minor category by dividing the actual call volume of the category by the total call volume. Then, the number of questions desired was distributed according to the percentage of total volume in each minor category.

It was decided that 60 test questions were desired. Additional test questions were needed from the previous year because of changes in the system. It has always been a goal to have a large bank of test questions to move in and out of testing. While the increase to 60 test questions is not a large bank of questions, it is progress in the right direction. The 60 questions were allocated to minor categories, as described above. Due to rounding, the total number of questions to be developed dropped to 57. Extra test questions were developed in categories with higher variance. Because, in general, the accuracy of the telephone assistance is around 85%, the greatest variance is in those categories having the lowest accuracy historically, i.e., closer to 50 percent, where the variance of a binomial is highest. One extra question was added in each of the minor categories Dependents, Nontaxable Income, and Employee Business Expense, which had low accuracy rates in 1991. An extra question was also included in the Estimated

Tax category, whose proportion of total call volume showed substantial periodic fluctuation. At this stage, the distribution of 61 test questions over specific minor categories had been developed.

The second phase of the question development process was to create the specified number of test questions in each minor category. It was very important that each test question closely resemble a taxpayer inquiry for two principle reasons. First, as previously discussed, unbiased results can be obtained only if the subject matter of the controlled test is reflective of the subject matter encountered by the assistors during taxpayer inquiries. Second, if the test questions did not closely resemble taxpayer inquiries, the test questions could stand out to those being tested and possibly be identified and handled differently than an unidentified question would be handled.

Throughout the year, a sample of live conversations between taxpayer and assistor are transcribed for use in understanding the nature of the interaction. They are used to study conversational style, complexity of the inquiry, the give and take between taxpayer and assistor, and the types of questions taxpayers ask. In general, they are used to help us understand what is a typical question and a typical conversation. To develop test questions similar in style and content to taxpayer inquiries, these transcribed taxpayer conversations on the IRS toll-free telephone assistance line were used. Questions within the transcribed conversations were to be used as "question starts." The question development team would model a test question based on one of these "starts."

The transcriptions of telephone conversations with taxpayers were categorized by ITCSS minor category, according to the type of question. Each conversation was read to determine the number of distinct questions in the call. An individual test call is designed to test the assistor's accuracy in answering one specific question. Therefore, those transcribed questions which contained more than one principle question were not used. At this point the set of transcribed questions, categorized by minor category, included only single issue conversations. Seventy-five percent of the transcribed conversations were single issue questions.

The next step in formulating "question starts" was to identify the principal question in each call. Within each category, questions were separated into groupings based on similar questions. Then, the question groupings with the largest number of questions were identified to be used as "question starts" to develop a test question. Through this method, the most commonly asked question types within a minor category would serve as a model for the development of a test question. This method of development ensured that the test question would be representative of the taxpayer inquiries which were to be tested, and would be similar in style, so as to not stand out as a test question during testing. For example, if four new test questions were needed within the major category Pensions, then the four transcribed question groupings within this category which contained the greatest number of questions were identified and used to develop one test question from each grouping. The Taxpayer Service Division of the Internal Revenue Service used the subject matter and the nature of the test questions to aid them in their wording of the test questions. Once they formulated the test questions, the test questions went through extensive telephone testing, with the aid of the Bureau of Labor Statistics' Collection Procedures Research Laboratory. Furthermore, the questionnaire design group of the General Accounting Office reviewed language aspects of the test questions. Then, the test questions were submitted to Chief Counsel of the Internal Revenue Service and the General Accounting Office for legal review.

RESULTS

When a test question has completed the writing and testing process described above, it is still far from being an acceptable question for use in the ITCSS testing. It must undergo a process of negotiation between IRS and GAO attorneys, operational, and technical staff. Questions are often changed substantially during this process.

When the final review for question accuracy occurs, wording changes are made to address, not only technical accuracy, but to make the questions less specific, i.e., less memorable, to make the questions harder or easier, or for various other reasons that emerge during the negotiation process. This is the nature of negotiations. However, wording changes at this late stage can cause other problems and the questions must be retested for clarity and for unanticipated responses and probes. Over the lifetime of the ITCSS, we have attempted to reduce the number of iterations a question must go through. We have provided the question starts, results of testing, and technical research to the negotiators in advance. However, the process remains a barrier to developing the kind of flexible question bank needed to address the credibility concerns arising from the small number of test questions and the number of times each is asked in a filing season and in its lifetime in the system. Both sides in the negotiation are anxious to improve this process but have been unable to solve the problem.

As Figure 1 shows, in the target test question distribution (columns 1 and 2), test questions are distributed

Figure 1. --Test Question Distribution by Minor Category

Minor Category	Target		Actual	
	Percent	Number of questions	Percent	Number of questions
Filing Requirement	10.66	6	19.23	6
Estimated Tax	3.80	3	4.62	3
Dependents	6.12	5	7.69	5
Personal Exemptions	1.10	1	1.54	1
Filing Status-Head of Household	2.43	1	1.54	1
Filing Status-Other	1.93	1	1.54	1
Income	3.92	2	4.62	3
Interest/Dividend Income	3.30	3	4.62	3
Other Income	2.99	2	3.08	2
Non-taxable Income	2.10	1	1.54	1
Capital Gains/Losses on Sch. D	4.51	3	4.62	3
Sale/Exchange of Residence	3.85	2	3.08	2
Other Gains/Losses	3.18	2	3.08	2
Pensions and Annuity Income	3.92	2	3.08	2
All IRA Inquiries	7.12	4	7.69	5
Other Retirement Plans	2.68	2	0	0
Taxation of Soc. Sec. Benefits	2.40	1	1.54	1
Lump Sum Distribution	2.11	1	1.54	1
Employee Business Expense	3.75	3	4.62	3
Other Adjustments to Income	0.52	0	0	0
Medical and Dental Deductions	1.66	1	1.54	1
Tax Deduction	2.46	1	1.54	
Interest Deduction	3.20	2	3.08	2
Miscellaneous Deductions Sch. A.	2.07	2	3.08	2
Other Allowable Sch. A. Deduc.	1.83	1	1.54	1
Unallowable Sch. A. Deduc.	----	----	1.54	1
Standard Deduction	2.04	1	3.08	2
Itemized vs Standard Deduction	1.32	1	3.08	2
Child and Dependent Care Credit	2.19	1	1.54	1
Self Employment Tax	3.77	2	3.08	2
Earned Income Credit	4.57	3	4.62	3
Other Credits/Taxes/Payments	2.43	1	3.08	2
TOTAL	100	61	100	65

across the minor categories to best reflect the volume distribution of calls to the IRS telephone assistance line. The actual test question distribution (columns 3 and 4) that was developed is very close to the target distribution, both in terms of total questions, as well as the distribution of questions across minor categories. Because of the challenges posed by the negotiation and approval process, the actual distribution does vary slightly from the target distribution.

The accuracy of the 1992 test questions during the February through April filing season is illustrated in Figure 2. All differences cited are significant at the 95% level. Questions in 1991 that were reused in 1992 were categorized as:

Revised slight changes were made in certain words or figures, but the basic question remained the same

Comparable the question remained exactly the same in 1991 and 1992.

The questions that were newly developed for 1992 showed statistically significant differences in accuracy compared to the revised and comparable questions used in 1992. There are several possible reasons for these differences in accuracy. The difference between new questions and previously used questions in 1992 may illustrate the need for a large test question bank to protect the test questions from identification. Additionally, the difference in accuracy rate for the comparable questions in 1992, as opposed to 1991 may show a real improvement on the part of the assistors. The difference in accuracy rate for the revised questions in 1992 may illustrate an improvement in the test questions themselves, to better reflect the accuracy of the assistors. These are all possible interpretations of the results to consider.

Figure 2.--Filing Season Test Question Accuracy, by Year and Question Type (February - late April)

Question Type	1992	1991
New	86.75%	----
Revised	92.53%	81.28%
Comparable	91.80%	87.42%

NEXT STEPS

A goal for the ITCSS is to change to keep pace with operational changes, such as an increasing use of specialization by telephone assistors, the need for a large and flexible question bank, changes to the negotiation process and the GAO role. Basic question development is expected to become a year-round process, with the goal of establishing and maintaining a very large question bank. Along with the notion of a flexible system with a large number of test questions, comes the need to develop equivalencies among the questions and to learn how to design a statistical system that effectively uses the greater flexibility. This is a major design concern for subsequent years.

NOTES AND REFERENCES

[1] Collins, Nancy (Ed.) (1988), "1988 Integrated Test Call Survey System--Volume I: Working Papers" and "1988 Integrated Test Call Survey System--Volume II: Statistical Documentation," Internal Revenue Service.

[2] Collins, Nancy (Ed.) (1989), "1989 Integrated Test Call Survey System--Volume I: Design and Development" and "1989 Integrated Test Call Survey System--Volume II: Implementation," Internal Revenue Service.

[3] Collins, Nancy (Ed.) (1990), "1990 Integrated Test Call Survey System--Volume I: Design and Implementation Issues" and "1990 Integrated Test Call Survey System--Volume II: Results and Improvement Initiatives," Internal Revenue Service.

[4] Daniels, Jonathan (Ed.) (1991), "1991 Integrated Test Call Survey System--Volume I: Design and Operation" and "1991 Integrated Test Call Survey System--Volume II: Results," Internal Revenue Service.

[5] Batcher, Mary and Scheuren, Fritz. (1989), "The IRS Test Call Program," **1989 Proceedings of the American Statistical Association, Section on Business and Economic Statistics**.

[6] Lee, Robin and Batcher, Mary. (1991) "Assessing the Test Used in the IRS Test Call Program," **1991 Proceedings of the American Statistical Association, Section on Survey Research Methods.**