THE IRS TEST CALL PROGRAM

Mary Batcher and Fritz Scheuren, Internal Revenue Service

Since 1965, the Internal Revenue Service (IRS) has offered free telephone assistance to taxpayers on income tax matters. Currently, this service is offered at 31 toll-free telephone sites located throughout the United States [1]. IRS telephone assistors handle about 50 million U.S. calls per year. During the 1989 filing season, January through mid-April, approximately 20 million questions were answered. The single largest category of inquiries received during the filing season are individual tax law questions (Figure 1); that is, questions about individual income tax returns which can be answered using IRS Publication 17 or the 1040 instructions [2-3].

Figure 1.--Percent of Taxpayer Inquiries by Type, 1988 and 1989 Filing Periods

Type of Inquiry	1988	1989
Individual Tax Law..............	41.2	39.1
Non-Individual Tax Law...........	6.8	10.7
Account..........................	20.6	25.2
Procedural.......................	31.3	25.0

The rest are:

- account-related inquiries which include questions about refunds and payments due;
- procedural inquiries about the mechanics of filing returns: where to file, which form to use, where to get forms, etc.; or
- non-individual tax law questions, including questions on employment taxes, excise taxes, and so forth.

Taken together, account and procedural inquiries predominate year-round but are especially prevalent during the May to December period each year.

With the large volume of inquiries handled by the telephone assistance service, the accuracy of the information given has a potentially large impact and is, understandably, of interest both within and outside of IRS. The telephone assistance service is often the point of first contact between the taxpayer and IRS. Hence, accurate information, at that point, can go a long way toward improving the quality of the returns filed, with corresponding cost savings for the taxpayer and the agency. Inaccurate information, on the other hand, could lead to the filing of an incorrect tax return, resulting in a potential overstatement of the taxes due or, conversely, an underpayment and a loss of revenue to the government. Additional costs are incurred in the error resolution process.

This paper describes a major IRS effort to assess the accuracy of the IRS telephone assistance service. First, we begin by providing some background; next, we describe the test call survey in some detail. Then, we talk about the statistical estimation procedures used to provide both the national accuracy level and the detailed weekly feedback. Finally, some overall conclusions are drawn and some ideas for future study spelled out.

BACKGROUND

Public interest in the accuracy of information given by the toll-free assistance service is high. Some of this interest is generated by Congressional hearings. Every year, both IRS and the General Accounting Office (GAO) are called to testify before Congress on the accuracy of the information given by the toll-free service [4-5]. The Congressional testimony, in turn, generates a barrage of newspaper articles; indeed, IRS has been under fire about its performance in this area. ("The Taxman Flubbeth" [6] was among the gentler references in the press during 1989.)

IRS has initiated efforts designed to improve the quality of its toll-free assistance. For example, each of the 7 IRS regions recently set up Diagnostic Centers to study the telephone assistance service in the call sites in their region, identify problem areas, and make recommendations for improvement. The effectiveness of this new step, and of still other improvement efforts, is of major concern within the agency.

There is more than one potential option available for measuring the accuracy of the toll-free assistance service. Direct measurement of the taxpayer/assistor interaction, through monitoring or recording [7], posed problems. Monitoring for quality review had been carried on for several years at the call site level. The accuracy results from this process were generally very high and at odds with GAO assessments. There was a concern that the monitoring carried out at a local level was not impartial. Additional difficulties exist in standardizing a decentralized monitoring system and the technology was not available until recently to allow centralized monitoring [8].

For the last two years, IRS has operated a program of test calls to assess the accuracy of the technical tax law advice given by the telephone assistance service--specifically, questions about individual income tax returns which can be answered using IRS Publication 17 or the Form 1040 instructions. This test call program--the Integrated Test Call Survey System or ITCSS--is based on the earlier GAO test call program and developed and operated in cooperation with GAO [9-10]. 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 call site volume of calls answered. 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.

These two purposes, one to meet the accountability needs of the broader society, represented by Congress, and the other to serve an assessment function for evaluating improvement efforts, ultimately to better serve the public, shaped ITCSS.

SYSTEM OVERVIEW

The Integrated Test Call Survey System consists of a series of approximately 1,500 calls per week placed to the 29 toll-free call sites in the continental United States. Calls are made by test callers in Washington, DC. Results are entered on PCs during the test call. Callers merely record the presence or absence of certain key information; scoring is accomplished later, using a computer routine which determines whether particular combinations of response elements are present and whether the taxpayer service "assistor" elicited sufficient background information to be able to provide a correct answer that was not a "lucky guess." The test call system is comprised of several components: the categories of tax law tested, the test questions, and operational components. Several mechanisms are employed to monitor ITCSS quality, as described below.

Categories

For the 1988 ITCSS, test questions were classified into 22 categories derived from the individual income tax forms. A large sample of incoming taxpayer inquiries on individual tax law issues (about 70 to 80 thousand per week) was also classified according to these same categories and the results were used for poststratification weighting of the test call results. When the volume of taxpayer calls in each category was collected, it was clear that the distribution of incoming calls was not spread evenly across the categories; instead, it was concentrated in 5 or 6 categories, with others receiving relatively few calls (Figure 2).

As part of a redesign effort following the 1988 ITCSS, a new classification scheme was developed. This new categorization was data driven, using results of an existing "live call" quality management system for the toll-free assistance service. The Quality Management Information System, (QMIS), monitors a random sample of up to 200 taxpayer and assistor conversations per month at each of the 31 call sites [11]. The monitoring is done by local IRS quality assurance personnel. For the 1988 filing season there were some 14,200 taxpayer inquiries on individual tax law issues monitored in QMIS. Under that study, each monitored conversation was classified according to 141 categories of inquiry, 64 of which were individual tax law issues. (See [12] for category details.)

Figure 2.--Percent of Individual Income Tax Inquiries by Major Category, 1988 Filing Period

Category	Percent
Filing Information....................	20.6
Exemptions............................	8.0
Income................................	22.0
Adjustments to Income.................	10.7
Tax Computation.......................	7.3
Credits...............................	3.1
Other Taxes...........................	3.4
Payments..............................	6.5
Schedule A............................	15.0
Sale of Residence.....................	3.4

To develop the 1989 ITCSS categorization system, the QMIS results were used as a starting point, to construct categories that would be more equal in size than the 1988 categories, with the hope that the sample could become self-weighting. The QMIS categories related to individual tax law were sorted by size and aggregated to form new categories in which the volume of incoming calls was spread "relatively evenly" throughout (Figure 3). Initially, ten major categories were developed for reporting purposes (each having from six to eleven percent of the volume). Contained within those ten major categories were 42 subcategories. Subsequent analysis has reduced these numbers to 7 major and 34 minor categories. Although many of the category titles in 1988 and 1989 are deceptively similar, they are not equivalent. (For specific definitions of elements included under similarly named codes in both years, see [9] and [10].)

Figure 3.--Percent of Individual Income Tax Inquiries by Major Category, 1989 Filing Period

Category	Percent
Filing Information..................	15.5
Exemptions..........................	11.0
Income..............................	10.4
Capital Gains.......................	11.3
Pensions............................	19.3
Adjustments to Income...............	15.5
Tax Computation.....................	17.0

During the 1989 ITCSS operation, several mechanisms were introduced to monitor implementation. One of these was of the categorization of incoming taxpayer calls into the ITCSS categories. This process was the basis for the poststratification weighting of the test call results. To monitor the consistency of classification, a monthly sample of questions was developed and sent to the call sites for coding and then examined for consistency. The agreement was generally around 80 to 90 percent. This was not as high as hoped and, consequently, in 1990, although we will continue to seek greater agreement in the classification of calls, we will collapse subcategories until they are nearly equal in size (and weight) to minimize the effect of misclassification.

Test Questions

When IRS began its test call program in 1988, the GAO test call system was used as a starting point. This led to one of the two design requirements imposed in the construction of test questions. That is, that as many of the existing 1987 (and earlier) GAO questions as possible be used to allow the 1988 IRS results to be benchmarked to earlier years.

The second design requirement imposed during 1988 question development was that each subcategory have at least two questions, with more in higher volume categories.

During the development of test questions and test call procedures, we were aware of the artificiality of test calls in assessing the accuracy of the information IRS assistors provide to the public. However, the measurement control possible in test calls combined with operational difficulties in the implementation of a nationwide monitoring system led us to a test call system in which we attempted to develop test questions that were similar to questions taxpayers ask. To this end, a goal in the development of both 1988 and 1989 test questions was that they reflect, to the greatest extent possible, the nature and difficulty of actual taxpayer inquiries. We could not do this constructively in 1988 because, while there were lots of experts in IRS and GAO, we had no actual data to draw on.

For the 1989 ITCSS test questions, we were able to use a sample of transcribed taxpayer questions as starting points for the development of new test questions. During the 1988 filing season the call sites were asked to have someone transcribe the opening question of taxpayer inquiries for two 3-hour blocks of time--one in the morning and the other in the afternoon of a different day--during the week of March 7. These transcriptions were studied as a means to improve test call procedures and became the starting points for the 1989 test questions. About 1,000 of the transcribed calls were individual tax law questions. Starting points were sampled from these, using systematic probability proportionate to size sampling. The measure of size was the number of questions falling into each of the categories described above. About 200 question starts were sampled, each of which served as the basis for a test question. Once the test questions were written, tested, and reviewed for technical merit, they were subsampled and combined with repeat questions from 1988, to obtain a core of about 85 to 90 test questions for use in the ITCSS measurement. This allowed an average of two test questions per minor category, with a few extras to use in categories with higher than average volumes of incoming taxpayer calls. The final set of new test questions was reviewed by IRS Office of Chief Counsel and other internal functions and then sent to GAO for concurrence. Samples of complete conversations were collected during the 1989 filing period; however, the extensive question review process precluded updating the 1989 test questions based on that information. This information was used in the 1990 question development.

Although the initial intent was to have at least two test questions in each subcategory, with additional questions in subcategories receiving high volumes of taxpayer inquiries, the joint IRS/GAO question review process involved a lot of give and take on the questions and several were deleted by one side or the other. There were 62 questions in the final 1989 set, spread somewhat unevenly across subcategories of tax law, with several subcategories containing more than three questions and a few with none. Even so, the 1989 question set we finally ended up with was richer and more representative than the set used by IRS in 1988 (which had 34 questions) or the GAO set of 21 in 1987.

Sample Design

As stated earlier, there were two purposes to be addressed by the ITCSS results: an overall accuracy level estimate for Congressional testimony; and a detailed week-to-week trend estimate, to provide the internal feedback needed for improvement efforts.

If we focus on each of these separately, we are led to different allocations of questions and test calls:

- To produce a good national accuracy level estimate, we would allocate questions to cover the subcategories of tax law, with more questions in larger categories with greater variance (i.e., with accuracy rates closer to .5); and we would place more calls to larger call sites and to those sites with lower overall accuracy;
- To produce a good estimate of week-to-week change at the call site level, the best approach is to use the same questions every week at every site. It would not be as important to cover all categories of tax law, except where category estimates are desired.

Our response to these conflicting goals was a compromise. We tried to place calls and test questions so that we would be able to meet both estimation needs: level and trend. A subset of 29 of the test questions was asked every week in each of the 29 call sites. These questions were allocated so that the seven major tax law categories each contained at least four questions, with the largest containing five. This subset of questions was used to provide weekly trend estimates. The remaining 33 questions were allocated to call sites over the 11 weeks of testing to augment the trend sample so that, by the end of the testing period, questions were balanced over both subcategories and call sites to produce an improved national accuracy level estimate.

Test Call Operation

During the first 5 weeks of the 1988 ITCSS, 870 test calls a week were placed from a central location in Washington to each of the 29 call sites; this was increased to 1,296 call attempts per week during later survey weeks. Question attempts were made every week in every site, with test calls placed from early January through the middle of April. Calls were scheduled to cover as many of the normal operating hours in the call sites as possible, with time zone differences incorporated into the schedule. Each test caller entered the call results directly into a microcomputer database, coding whether the required background information had been solicited and whether the assistor's response corresponded to the predetermined correct answer for that question.

All responses were reviewed for technical accuracy by GAO and by IRS's Office of Chief Counsel.

The 1988 ITCSS was put in place in a very short time, without a permanent staff of test callers. There were frequent substitutions of callers, who were on loan from other functions. Consequently, callers had minimal or no training in test call procedures, many of which were developed by the callers on the job and formalized only gradually. Test callers were assigned question/site combinations according to the scheduling in the sample. When they finished their calls, they were allowed to return to their regular assignments. During the 1988 filing season, there were 31 test callers. Caller effects were a major concern, and individual caller accuracy rates were closely monitored, but efforts to reduce caller effects met with only limited success. The weekly range of caller accuracy rates varied from a maximum of 90.4 to a minimum of 64.1, with a median of 78.4.

The operation of the test call program during the 1989 filing season consisted of a permanent staff of eight test callers, based in Washington, DC, who placed approximately 1,500 test calls per week to the 29 call sites in the continental United States [13]. Test calls were scheduled to cover as much of the normal operating hours in the call sites as possible, within time zone limitations. They were scheduled Monday through Friday, with extended hour and weekend operations generally not covered. Test calls for the 1989 ITCSS began in January and were continued throughout the year, leading directly, without a break, into the 1990 ITCSS. This was a change from the 1988 test call program, which ended after the filing season. The intention is to continue to place calls year round, with increased numbers of calls and test questions during the filing season. Another change introduced in 1989 was a shift from recording results of call attempts to counting completed calls, with calls rescheduled until completed. (Calls not completed within the week, however, were not carried over to the next week.)

The test callers placed calls throughout the workday, with staggered breaks and lunches. They were given the questions and times to call the various call sites, with test calls scheduled every 10 minutes and a single question asked in each call. The 10-minute scheduling was determined as optimal during preliminary 1989 testing (and from the earlier 1988 experience); it provided the maximum number of calls that could be routinely completed.

Test callers in 1989 were trained in test call procedures, using all of the test questions, to ensure consistency in the way they asked the questions, responded to probes for background information from the assistors, and coded the responses. This training included role playing, with callers playing the part of assistor and test caller; probing skills; tax law training, so they could distinguish the various forms that responses can take; and training in the mechanics of the system. The 1989 test callers were IRS staff who, for the most part, had previously spent time as telephone assistors answering taxpayer inquiries. Test callers who had not been assistors were temporarily detailed to a field office so they could have the experience of having been telephone assistors.

As already noted, the 1989 sample of test calls was developed to include a set of 29 questions that were called to every site, every week, with the remaining calls allocated so that the largest call sites received more calls. The sample was balanced over weeks of the filing season, callers, and days of the week. Although the callers called each site and called with all questions, the same caller/question/site combinations occurred throughout the filing season. This allowed us to monitor change in accuracy without confounding site and caller effects.

We continued our internal monitoring of caller effects in 1989. (GAO also did some live monitoring that proved very helpful [14].) At the beginning of the 1989 ITCSS operation, the test callers seemed to fall into two clusters, based on their average accuracy level. Throughout the 1989 ITCSS operation, test call procedures and problems were discussed daily in group meetings of the test callers. After a few weeks, the accuracy differences were reduced but not eliminated; nonetheless, we judged the differences to be small enough so that the caller effect could be said to be under control. (The weekly difference in percent accurate between the two caller groups ranged from a high of 12.5%, in the second week of testing, to a low, at the end of the filing season, of 0.6.% The median difference was 5.9%.)

ESTIMATION

The estimation problem was complicated by the need of the call sites for more detailed weekly data. The number of test calls could not be increased; the system was at its capacity. At the national level, the trend sample consisted of well over 100 calls per week in each major category. At the call site level, there were only four or five calls per week in the trend sample in each major category. These small site-level sample sizes were addressed in two ways. The first was through the use of two-week moving averages, where, although estimates were produced every week, they were for the most recent two weeks combined.

The second way we addressed the problem of small sample size was in the form of the estimator employed. We used a James-Stein type of procedure, where the actual call site data were combined with data developed using a contingency table model, to produce improved call site by category estimates.

As is well known, James-Stein estimators reduce the mean squared error of unbiased sample estimates by shrinking the sample estimate toward a group centroid. The amount of shrinkage is determined by the ratio of the centroid variance to the unadjusted estimate variance [15].

Our problem was to determine the accuracy rates for the cells of a three-dimensional contingency table, where the dimensions were accuracy, call site, and tax law category. At its most general, we had a 2 x 29 x 34 table--very sparse, indeed; even with 29 x 29 x 2 = 1682 calls every 2 weeks, there would be only about one call per cell. The approach taken was as follows:

- The table was collapsed to look just at major categories, i.e., to 2 x 29 x 7.
- The accuracy rates, $\hat{p}_{ij}$, i=1, ..., 29; j=1, ..., 7, were then weighted by major category and call site volume and aggregated to the national level.
- A contingency table model of conditional independence was fit next, where the marginal frequencies of call site by accuracy and major tax law category by accuracy were fixed.
- The model estimated cells were to be combined at this point with the actual cell data, $\hat{p}_{ij}$, in a James-Stein type of procedure;
- If we denote the model-based estimate as $\tilde{p}_{ij}$, our estimate, $\hat{\hat{p}}_{ij}$, is then, $\hat{\hat{p}}_{ij} = \alpha\tilde{p}_{ij} + (1 - \alpha)\hat{p}_{ij}$.
- The value of the mixture parameter, α, was determined by the fit of the conditional independence model relative to the fit of a model of complete independence. More precisely, it is the ratio of information statistics of the two models, conditional independence divided by complete independence [16]. This ratio of information statistics operates like the ratio of variances in the James-Stein estimator.
- For the weekly trend estimates, α was evaluated using preliminary data and set at a fixed .5 throughout the 1989 filing season.

To obtain the level estimates the test call results were time-weighted to take account of the higher volume periods of the filing season. The level estimates were based on all data collected during the test period.

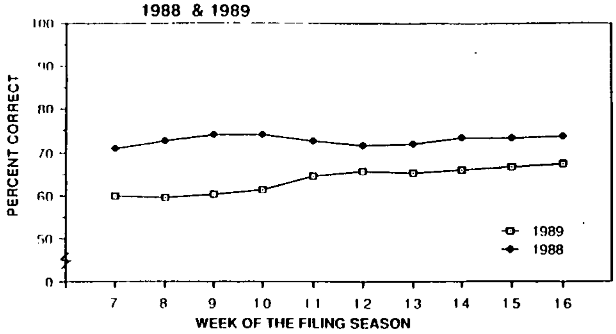
The final 1989 national accuracy level estimate for percent correct was 62.8, with an estimated standard error of approximately 1 percent. The 1989 trend estimates are shown in Figure 4. There was very little change over the course of the filing season--an increase of 7.5 percent from the two-week moving average for weeks 6 and 7 to the final two-week average for weeks 15 and 16. The 1988 trend results are also presented in Figure 4. Again, there was very little change, with an overall increase of 2.9 percent from the week 7 estimates to the week 16 results. There was a decline of 6.5 percent in the accuracy measurement between 1988 and 1989. Figure 5 displays the trend lines for 1988 and 1989. The shapes are fairly flat with a generally stable difference after the eleventh week. As noted earlier, changes in the accuracy between the two years are confounded with

Figure 4.--Individual Income Tax Law Inquiries: Percent Correct by Category and Week,Two-Week Moving Averages, Weeks 7 Through 16

Major Category	Filing Period Week									
	7	8	9	10	11	12	13	14	15	16
	Part I - 1989									
Overall Total..................	59.8	59.5	60.4	61.4	64.6	65.7	65.3	66.0	66.5	67.3
Filing Information...............	66.0	69.7	67.8	66.2	69.3	68.9	72.6	70.5	65.9	69.0
Exemptions........................	76.0	81.9	80.7	78.8	82.4	80.8	78.6	81.7	80.0	74.4
Income............................	44.1	46.9	48.5	54.9	53.2	55.7	62.6	65.1	60.5	57.7
Capital Gains.....................	33.8	33.5	30.9	29.3	34.6	34.3	31.3	35.1	40.7	39.2
Pensions..........................	67.0	61.8	57.6	57.0	67.8	72.3	69.5	68.6	73.0	76.7
Adjust. to Income................	56.4	54.5	61.2	61.8	59.4	62.9	61.8	60.5	64.6	65.5
Tax Computation..................	65.5	63.4	71.7	77.0	77.8	77.5	76.9	77.4	74.3	75.1
	Part II - 1988									
Overall Total..................	70.9	72.6	74.2	74.1	72.8	71.7	72.0	73.4	73.5	73.8
Filing Information...............	81.4	88.2	90.0	91.1	91.3	92.1	93.4	93.6	92.8	92.3
Exemptions........................	66.9	62.6	61.8	64.1	58.4	47.0	41.7	45.5	44.8	40.2
Income............................	64.6	62.7	61.0	59.4	58.0	48.8	51.8	60.1	57.3	54.5
Adjust. to Income................	77.7	77.6	82.1	80.2	79.9	86.6	77.9	67.8	71.3	71.9
Tax Computation..................	75.2	71.6	71.6	67.9	65.4	78.8	76.5	72.7	74.6	75.1
Credits...........................	49.5	59.3	59.1	52.7	52.0	50.8	57.6	64.2	55.9	53.1
Other Taxes.......................	55.0	64.3	82.9	75.0	73.5	91.5	89.4	86.0	79.8	77.3
Payments..........................	76.6	80.0	79.6	82.8	84.1	86.5	90.9	92.6	92.7	93.7
Schedule A........................	68.1	71.7	79.5	81.3	79.1	74.6	78.2	74.5	70.9	76.8
Sale of Residence................	71.8	76.7	68.4	64.6	62.2	56.2	57.8	68.3	57.1	46.5

Note: As explained in the text, the 1988 and 1989 categories were defined differently even when labelled identically.

FIGURE 5.--NATIONAL ITCSS ACCURACY RATES, 1988 & 1989



questions effects. However, a real decline in accuracy appears to have occurred; obviously this is of grave concern to us.

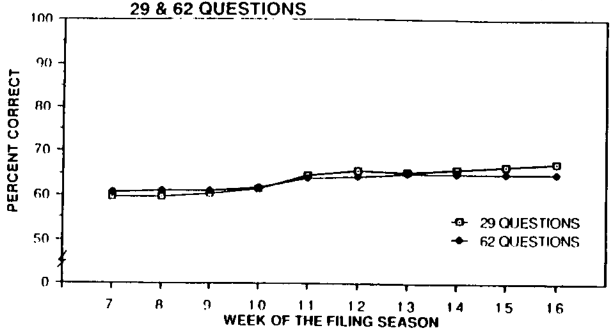
CONCLUSIONS AND AREAS FOR FUTURE STUDY

In earlier sections, we mentioned areas of concern and suggested some of the ways we hope to be able to deal with these. Some of the issues mentioned include the categories and the caller effect. We are also interested in ways to improve the test questions. Although the transcribed conversations are being used in question development, we are far from fully exploiting the full potential of this rich source of information. The use of techniques from the area of cognitive psychology in improving survey instruments and procedures [17] is one that we feel could benefit both the test questions and the test caller procedures. We have arranged with the Bureau of Labor Statistics to use their cognitive lab to study ITCSS test questions and caller procedures.

We are now in the planning stages for 1990. During the 1989 ITCSS, the presence of both level and trend accuracy rates caused difficulties in interpretation for the data users; therefore, the decision was made to have only one set of numbers. We examined the relationship between the accuracy calculated using the 29 trend questions and those using the full set of 62 questions. The results are displayed in Figure 6. There was virtually no difference between the pure trend results and those combining both level and trend calls.

For 1990, we plan to produce weekly estimates using all questions. Questions well be allocated approximately proportionate to the category volume of calls, a minimum number of calls per question/site combination placed, regardless of category volume. For the beginning of the 1990 survey, 1989 volumes will be used. The sample proportions will be adjusted, if needed, during the 1990 filing period. We do not anticipate other major changes to the estimation procedures themselves for 1990.

FIGURE 6.--NATIONAL ITCSS ACCURACY RATES, 29 & 62 QUESTIONS



In the estimation of accuracy, ITCSS is only one of several indicators we can look at. Although ITCSS was the primary indicator in 1989, we also used both centralized monitoring of live calls in two sites and a sample of transcriptions of entire conversations between taxpayers and telephone assistors (with all identifying information removed). In the area of tax law, comparisons of the accuracy measurement from these indicators to the test call data showed a high degree of consistency. We intend to continue to use these other indicators in 1990 to benchmark the ITCSS measurement.

ACKNOWLEDGMENTS

The authors wish to thank Nancy Collins and Mary Keriotis for their assistance in compiling material included in the paper; John Czajka and H. Lock Oh for their continuing role in designing the test call system and developing and producing estimates; and William Kalsbeek, who discussed this paper at the Washington meetings. We have attempted to address his comments in the body of the paper. Thanks are also due to Wendy Alvey and Bettye Jamerson.

FOOTNOTES AND REFERENCES

[1] Free walk-in assistance is available even more widely throughout the United States. There are also toll-free telephone sites for international locations and Puerto Rico.

[2] Internal Revenue Service (1988). Your Federal Income Tax, Publication 17.

[3] Internal Revenue Service (1988). Instructions for Form 1040.

[4] LeBaube, Robert (1989). Statement of the Director Taxpayer Service Division, Internal Revenue Service, before the Subcommittee on Commerce, Consumer, and Monetary Affairs, Committee on Government Operations, House of Representatives, March 15, 1989.

[5] Stathis, Jennie S. (1989). Statement before the Subcommittee on Commerce, Consumer, and Monetary Affairs, Committee on Government Operations, House of Representatives, March 15, 1989.

[6] "J Street," The Washington Post Magazine, April 9, 1989.

[7] Recording the calls, while legal, was viewed as an invasion of taxpayer and assistor privacy.

[8] Internal Revenue Service (1989). "QUEST--Quality Evaluation System for Taxpayers."

[9] Batcher, Mary and Collins, Nancy (Eds.) (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.

[10] Batcher, Mary and Collins, Nancy (Eds.) (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.

[11] Internal Revenue Service (1989). "Taxpayer Service Quality Review Handbook," Manual Number 6541.

[12] Internal Revenue Service (1989). "Taxpayer Service Quality Reviewers' Guidelines," Manual Number 6542.

[13] Mainland calls to Alaska and Hawaii could be identified by the telephone sound quality. Therefore, the ITCSS only tested continental U.S. sites to reduce the risk of test question disclosure.

[14] United States General Accounting Office (1989). "Tax Administration: Monitoring the Accuracy and Administration of IRS' 1989 Test Call Survey."

[15] James, W. and Stein, C. (1961). "Estimation with Quadratic Loss," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Berkeley: University of California Press, 361-79.

[16] Scheuren, Fredrick J. (1972). "Topics in Multivariate Finite Population Sampling and Data Analysis," George Washington University, Ph.D. dissertation.

[17] Fienberg, S. and Tanur, J. (1989). "Combining Cognitive and Statistical Approaches to Survey Design," Science, Vol. 243, No. 4894, 1017-22.