THE IRS TEST CALL PROGRAM: DESIGN AND ESTIMATION ISSUES

Mary Batcher and Fritz Scheuren, Internal Revenue Service
Internal Revenue Service (R:S:P) P.O. Box 2608 Washington, D.C. 20013-2608

KEY WORDS: Quality Measurement, Sample Design, Estimation

The quality of service organizations is difficult to define and measure. Quality indicators that are extensively used in manufacturing settings do not carry over easily to service settings [1]. This complicates improvement efforts in service organizations and poses interesting statistical challenges, particularly in a government service organization like the Internal Revenue Service (IRS).

The present paper discusses some of the challenges we have faced in the design and implementation of a large-scale program of test calls developed to assess one dimension of service quality: accuracy and completeness of information provided to the public by the IRS telephone assistance operation. We begin by presenting some of the background leading to the development of the test call program and discuss some improvement efforts. We then describe the design and estimation procedures and present results. Finally, we conclude with some plans for the future.

BACKGROUND

Since 1965, the Internal Revenue Service 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 by IRS employees, called "telephone assisters" [2]. These assisters handle about 50 million U.S. calls per year. During the 1990 filing season, January through mid-April, approximately 17.5 million inquiries were handled. Questions range from issues as simple as the location of the nearest IRS district office to complex tax law issues. 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

Figure 1.--Percent of Taxpayer Inquiries by Type, 1988, 1989, and 1990 Filing Periods

Type of Inquiry	1988	1989	1990
Individual Tax Law	41.2	39.1	33.7
Non-Individual Tax Law	6.8	10.7	10.4
Account	20.6	25.2	28.2
Procedural	31.3	25.0	27.7

can be answered using IRS Publication 17 or the Form 1040 instructions [3-4]. 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 [5].

For the last three years, 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 which can be answered using IRS Publication 17 [3] or the Form 1040 instructions [4]. 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 developed and operated in cooperation with GAO [6-7-8]. 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.

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 the ITCSS.

IMPROVEMENT EFFORTS

During the first year of the ITCSS measurement (1988), the accuracy rate remained at a fairly steady level--at about 70 percent. This was a source of concern to IRS, since the agency had believed that it was doing much better than that.

Between the first and second years of the test call program, each of the seven regional offices set up diagnostic centers to assist the call sites in their region to identify areas where improvements were most

needed and to help identify successful improvement strategies. In spite of these and other improvement efforts, the 1989 accuracy not only failed to improve but actually declined to the low 60s. Although this was not the direction of change anticipated, the regional diagnostic centers had been in existence for only a few months and might not be expected to have an effect so quickly.

Throughout the 1989 filing season, the call sites and regional offices tried various strategies for improvement. One of the more promising of these was the introduction of a probe and response guide in Midwest Region. The probe and response guide is a set of instructions, by category of tax law, specifying the background information that the telephone assister should determine and the areas to be addressed in responding to the taxpayer. A preliminary version of the probe and response guide was introduced in one call site and the accuracy monitored for several weeks. The accuracy increased and remained significantly higher than it was prior to the use of the guide. An expanded version was then developed, in conjunction with the call site, and introduced in a second site. Again the accuracy increased. Following up on this promising effort, IRS' national office developed a probe and response guide and mandated that some version of a probe and response guide be put in place in all call sites beginning with the 1990 filing season.

The ITCSS is the designated indicator for the success of this and other improvement efforts for the quality of the toll-free assistance service. Its primary uses are then to measure the effectiveness of improvement efforts and to provide an overall gauge of the accuracy of the toll-free service. The sample and estimation procedures were designed to meet these needs within available resources.

DESIGN

The test call system itself is described in [9]. Briefly stated here, it is a system of test calls placed to each of the 29 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 of 8 test callers. For 1990, 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, four assisters 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 29 call sites, the 8 test callers, the 42 test questions, and the tax law categories. Each of these is described in greater detail in [9].

In order to meet the needs of the data users, it was necessary to design a system that could provide a fairly precise national accuracy rate for Congressional testimony and public release two or three times during the filing season and that could also provide regular feedback to the call sites and regions. The number of calls we could schedule was limited by the number of test callers, which was constrained, in turn, by the size of the test call facility. Allowing time for lunch, breaks, and daily caller meetings, left us with 6.5 hours of calling time per day per caller.

Although callers followed a schedule, if a call could not be completed at its scheduled time (most often because of busy signals), it was rescheduled later in the week. However, rescheduled calls that were not completed by Friday afternoon were lost. During the 1989 filing season a number of calls were lost because they had not been completed by Friday afternoon. To minimize lost calls, we only scheduled 56 calls for Friday (7 per caller), allowing the rest of the day to be used for making up calls that were not completed when scheduled. We were left with a sample of 1,320 calls per week to distribute over the categories, questions, call sites and callers.

If we distributed the calls equally over the call sites, we would have had 48 to 49 calls per site per week. However, some sites were considerably larger than others; one in particular received about three times as many individual tax law inquiries as the smaller sites. Additionally, the evaluation of an expert system that was being tested in one site in 1990 required a larger sample. We, therefore, allocated 90 calls per week to the expert system site and to the site with the very large volume of tax law inquiries, 60 per week to 7 other large sites, and 36 per week to the remaining smaller sites.

The next step was to determine the number of times each question would be asked in the call sites. We wished to represent the 7 major categories tested in proportion to their size. To improve the week-to-week change measures, we also wished to retain the same weekly sample of questions for each site throughout the filing season. The 42 test questions were not evenly distributed across the categories, nor were they distributed proportionate to category size. We were, therefore, in a position of selecting subsets of questions for some categories, especially in the smaller sites.

We selected the question set for each site in the following manner. We began by identifying the number of questions we wished to include in each category, based on category volumes. When this was less than the full number of available questions, we manually developed subsets of questions, having similar expected accuracy, based on preliminary testing with the questions. If the questions within a category were very different in their expected accuracy rates, all questions in the category were included. These subsets of questions were then randomly sampled to develop the question set for each site. This process resulted in greater representation in large categories and more variable categories, not unlike Neyman allocation [10].

It was important to us that no site be penalized by the set of questions assigned to it. We, therefore, balanced the sample on accuracy by calculating an

expected accuracy rate for each site and nationally, based on the preliminary testing with the mix of questions assigned. For those sites whose expected accuracy rates differed from the national accuracy rate by more than ±1%, another sample was substituted. The expected accuracies were then recalculated and extreme samples substituted for, until all sites were within range.

The sample was then distributed so as to spread callers as evenly as possible across sites and questions. Test calls were scheduled, by caller, to meet time zone restrictions and to fit the caller's work schedule.

ESTIMATION

The data needs of users of the test call system required the production of weekly call site estimates by category. Although, at the national level, the sample was fairly large, it was very small for the production of call site by category estimates. In prior years, the small sample size was compensated for in two ways: by the use of moving averages (two-week in 1989 and four-week in 1988) and by using a James-Stein type of procedure for the call site by category estimates [11-12]. However, the use of moving averages made it difficult for the data users to track improvement efforts. So, to accommodate their wishes, we began using one week estimates in 1990. This led, predictably, to greater fluctuations in the weekly call site, regional, and national numbers, but the estimates were still judged to be usable as is; however, the call site by category cell counts were very small indeed--4 to 5 calls per cell in the small sites. Even with the help of the James-Stein type procedures we had used in prior years, this was an extremely small number upon which to base an estimate. Our solution to this problem was to produce model estimates only at the call site by category level.

We fit a contingency table model to the data, where the marginals were fixed by the observed data for that week and the starting values for the interior of the table were the cumulative data, beginning with week 6 of the filing season, when we officially began the 1990 test call program [13]. We then iterated until the marginals were recovered. By starting with the cumulative, we were able to have a solution that reflected call site by category interactions.

RESULTS

The 1990 results are presented in Figure 2, along with the 1989 results for comparative purposes. It is apparent that there was a substantial increase in the national accuracy rate between 1989 and 1990. This follows upon 2 years of a relatively flat trend line and an actual decline between 1988 and 1989 (Figure 3).

One of the statistical questions we examined was whether the increase in the measured accuracy rate between 1989 and 1990 was due to real improvement or might have been a measurement artifact. While we cannnot, of course, be certain, the evidence for real improvement is compelling. For instance, we examined the relationship between 1989 and 1990 accuracy using only those questions common to both years and, again, there is an increase, from 71.4% correct in 1989 to 81.6% correct in 1990.

Estimates of the accuracy rate would not be

FIGURE 2
NATIONAL ITCSS ACCURACY RATES, 1989 & 1990

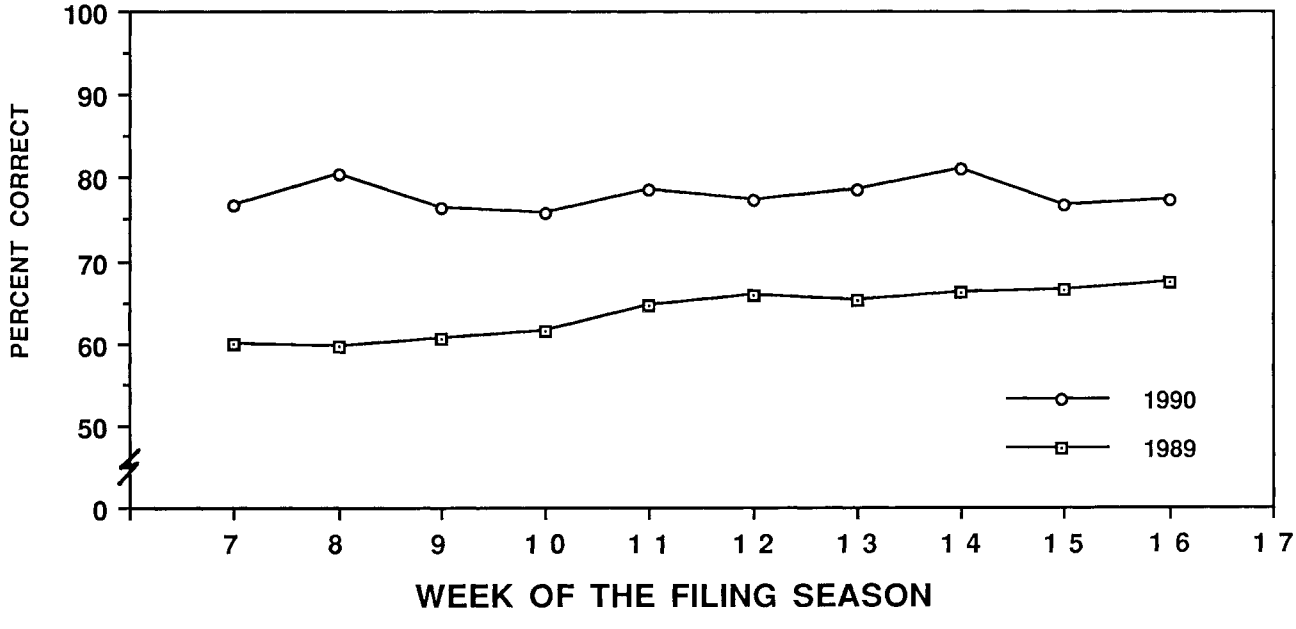
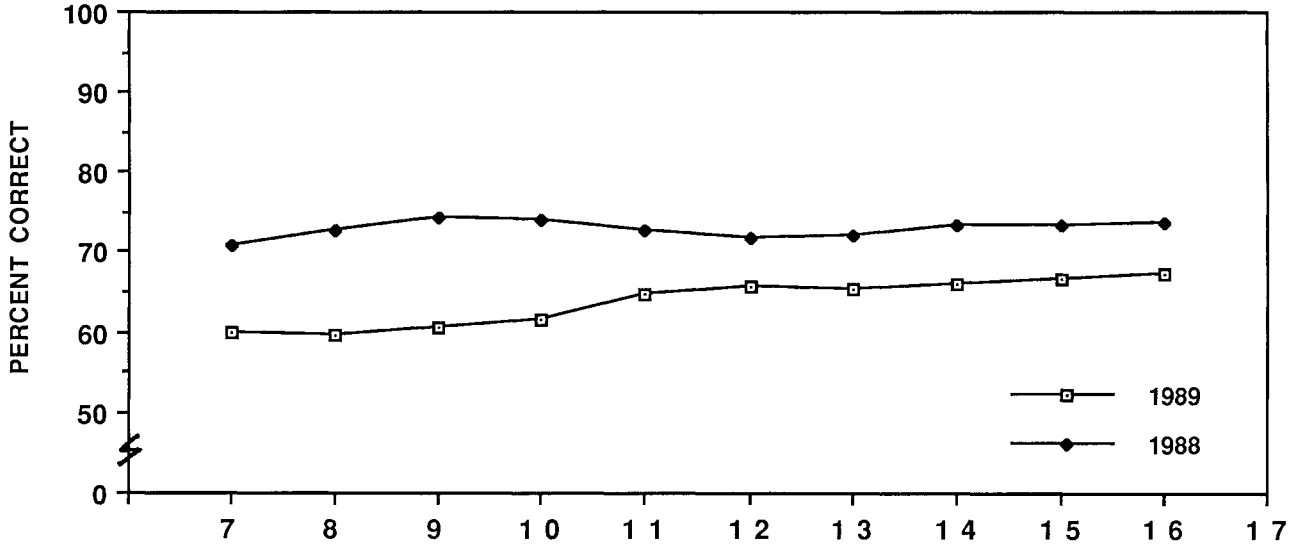


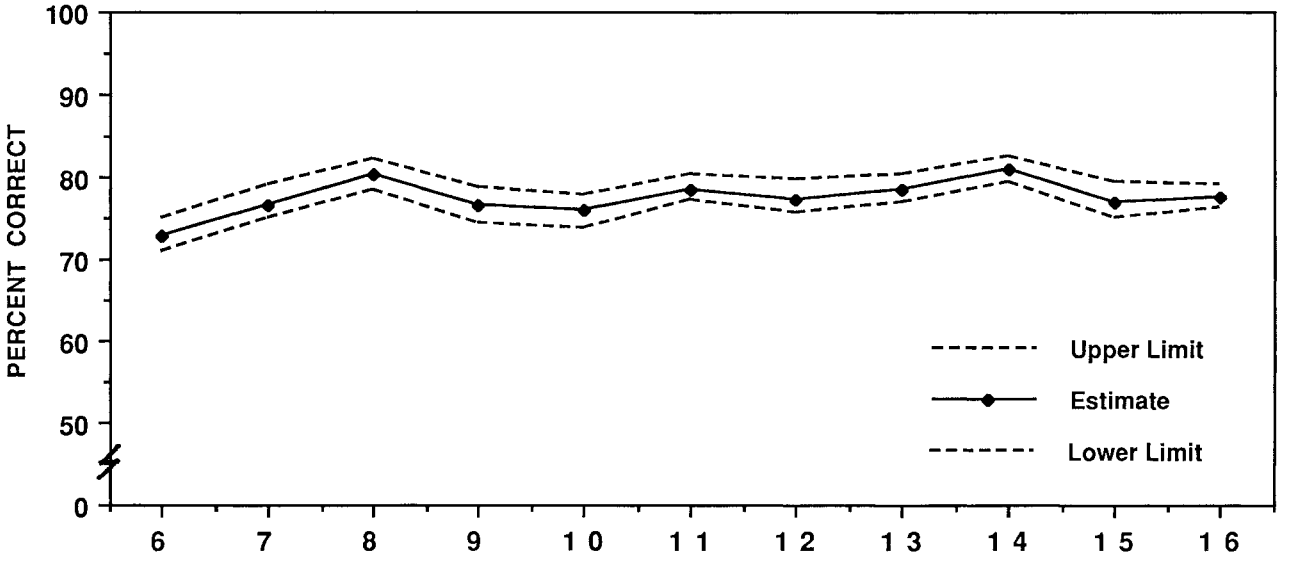
FIGURE 3
NATIONAL ITCSS ACCURACY RATES, 1988 & 1989



complete without some indication of the sampling error. The design complexity, with caller and question effects, did not allow a simple direct estimate of the standard errors. We used a bootstrap estimate, where, for each call site, we produced 101 bootstrap samples of size equal to the site sample [14]. The results, nationally for the filing season as a whole, are presented in Figure 4. The average length of the 90 percent confidence intervals for each week ranged from about 4 at the national level to approximately 20 at the site level, with regional levels being intermediate.

A basic problem in the ITCSS is the need to provide call site by category estimates. The data are very thin at this level and we will continue to explore options to maximally use the available information in the

FIGURE 4
ITCSS WITH 90% CONFIDENCE LIMITS



production of these estimates. Simulation studies using the 1990 data are being planned.

A second area of the ITCSS where we want to focus improvement efforts is the test questions and question process, itself. We have enlisted the support of the Collection Procedures Research Laboratory at the Bureau of Labor Statistics to analyze the test questions, themselves, and the processes involved in the interaction between test caller and assister and between assister and taxpayer [15]. We hope to use this information eventually to improve the way we place test calls.

Finally, in addition to improvements in the test call system, itself, we are working with the call sites to help them introduce improvement strategies in a more structured fashion. Although we have seen improvement in the accuracy overall, and particularly in certain sites, it has been very difficult to isolate potential explanatory factors. We will be attempting to influence the way improvement strategies are introduced and monitored, so that we can better isolate their effects.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of Wendy Alvey in the preparation of handouts and visuals and in providing editorial review of the paper and of Clementine Brittain in preparation of the paper for the Proceedings.

FOOTNOTES AND REFERENCES

[1] King, Carol (1987), "A Framework for a Service Quality Assurance System," Quality Progress.

[2] 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.

[3] Internal Revenue Service (1989), Your Federal Income Tax, Publication 17.

[4] Internal Revenue Service (1989), Instructions for Form 1040.

[5] There has been a decline in the individual tax law percentages over the three years shown, with a fairly sizeable drop in 1990. This may be partially attributable to a corresponding growth in the IRS Tele-Tax system, where taxpayers can dial a toll-free telephone number to hear a recorded message on one of several individual income tax topics.

[6] 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.

[7] 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.

[8] 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.

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

[10] Cochran, William C. (1977), Sampling Techniques, Third Edition, New York: John Wiley and Sons.

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

[12] Brandwein, A. and Strawderman, W. (1990), "Stein Estimation: The Spherically Symmetric Case," Statistical Science, 5, 356-69.

[13] Bishop, Y., Feinberg, S., and Holland, P. (1975), Discrete Multivariate Analysis, Cambridge: The MIT Press.

[14] Efron, B. and Tibshirani, R. (1986), "The Bootstrap Method for Assessing Statistical Accuracy," Statistical Science, 1, 54-77.

[15] van Melis-Wright, M., Batcher, M., Stone, D. and Scheuren, F. (forthcoming). "Cognitive Psychological Approaches in the Evaluation of Information Exchange Processes," 1990 Proceedings of the American Statistical Association, Section on Survey Research Methods.