Statistics of Income

2007 SOI Paper Series

Measuring the Quality of Service to Taxpayers in Volunteer Sites

by Kevin Cecco, Ronald Walsh, and Rachael Hooker

Measuring the Quality of Service to Taxpayers in Volunteer Sites

Kevin Cecco, Ronald Walsh, and Rachael Hooker, Internal Revenue Service

In 2000, the Internal Revenue Service (IRS) established an office, called SPEC (Stakeholder Partnerships, Education, and Communication), that aims to assist underserved segments of the taxpaying public in satisfying their tax responsibilities. These segments include elderly, disabled, low-income, multilingual, military, and other taxpayers who are otherwise unable to receive tax assistance. To achieve its mission, SPEC establishes and maintains partnerships with key stakeholders in local communities. With ongoing support and guidance from SPEC, stakeholder partners coordinate and manage site locations where taxpayers can receive support in tax preparation and answers to basic tax law questions from unpaid volunteers.

In order to effectively oversee partner relationships, SPEC must be able to measure the accuracy of the returns filed within SPEC sites. From the 2000 through 2005 filing seasons, SPEC relied on an unempirical approach to evaluating the quality of returns. Each filing season, a team of reviewers was sent to a select group of sites to pose as taxpayers and record the quality of service received. The results of these "shopping" reviews were used as qualitative indicators of the actual accuracy of returns prepared in SPEC sites.

The Statistical Support Section (SSS) of the Statistics of Income (SOI) Division of the IRS provides general statistical consulting services on request for various areas of the IRS, as well as for other branches of the Federal Government.

In late 2005, SPEC requested SSS's assistance in developing a new sampling methodology that could potentially result in more statistically defensible estimates of the accuracy of returns prepared in SPEC sites. This new methodology was tested during both the 2006 and 2007 filing seasons. While the test did establish the feasibility of the sample design, unexpected sources of nonsampling error arose during the test period. This paper details the proposed methodology and discusses the issues that may prevent this methodology from becoming a long-term solution to SPEC's quality measurement needs.

▶ SPEC Site Overview

There are over 12,000 SPEC sites. They are grouped into three basic partner types: Volunteer Income Tax Assistance (VITA), Tax Counseling for the Elderly, (TCE), and Military. Each partner type is geared toward serving a different segment of the taxpaying public. In addition, the country is split geographically into four distinct areas, which are themselves subdivided into 46-separate territories.

There are approximately 4,540 VITA and 7,822 TCE sites nationwide. They are physically located in public institutions within local communities. VITA sites, which tailor to low and moderate income taxpayers, are found in locations such as libraries, schools, and universities, while TCE sites, which accommodate elderly taxpayers, are found in establishments such as banks, senior centers, and churches. There are also roughly 200 military sites set up on various military bases within and outside the United States, as well as on military ships at sea.

During the 2006 filing season (January through April), there were over 2 million returns prepared in and filed from SPEC sites. Distributing returns by partner type shows that VITA, TCE, and military sites prepared 713,703, 1,059,288, and 324,197 returns, respectively.

▶ Return Review Pretest Phase

The new methodology for the SPEC Return Review was tested over the course of the 2006 and 2007 filing seasons. SPEC went through several steps in preparation for testing the methodology in the field. The following were conducted prior to the 2006 filing season:

- A data collection instrument (DCI) for the new Return Review was designed and tested.
- An online database to house review data and to generate reports was developed and tested.

- Internal clearances granting permission to capture review data at the site level were obtained. Capturing data at this level was necessary in order to calculate weighted point estimates and confidence intervals.

- An assessment was conducted of the capabilities and limitations of the resources allocated to implementing the Return Review.

▶ Estimates

The Return Review focused on the accuracy of tax-preparation services provided by volunteers in SPEC sites. The new DCI for the Return Review was broken out into eight major indicators of quality. For each sampled return, it was determined whether or not the volunteer successfully completed each indicator while helping the taxpayers file their returns. These indicators assessed the appropriateness and accuracy of different aspects of the return being filed, such as the filing status of the taxpayer, the number of dependents claimed on the return, the deductions and credits claimed by the taxpayer, and the total tax owed or due.

After all indicators were evaluated for a single sampled return, the overall return accuracy for that return was determined by combining the results for all indicators using a "pass/fail" methodology.

The primary goal of the Return Review was to obtain statistically valid estimates of the overall return accuracy for each partner type (VITA, TCE, military), separately, as well as all partner types combined over the course of the filing season. Each of these estimates was needed within 5-percent precision. 90-percent confidence intervals were calculated for each estimate.

Secondary goals of the review included:

- overall accuracy by geographic region (area)

- overall accuracy by area by partner type

- individual indicator for the nation

- individual indicator for the nation by partner type

- individual indicator by area by partner type

- individual indicator by area for all partner types combined.

90-percent confidence intervals were calculated for each of these estimates, as well. However, 5-percent precision was not required by SPEC.

▶ Sampling Frame

The list of SPEC sites is fluid. Each year, some sites that were previously operational close, while others open for the first time. Therefore, a new sampling population for the SPEC Return Review must be defined each year. In establishing the sampling population for the 2006 test year, SPEC chose to exclude sites that would be open for the first time during the 2006 filing season. In addition, sites closing after the 2005 filing season were removed.

Resources prevented SPEC from reviewing some sites in the population. Due to their physical locations, a minimal number of sites were deemed inaccessible. These sites included military ships, overseas military bases, sites located in Hawaii, and nine sites impacted by Hurricane Katrina. Inaccessible sites were removed from the sampling population. After removing new, closed, and inaccessible sites from the sampling population, 9,761 sites remained in the sample frame in 2006.

SPEC identified the 11-week period between January 30, 2006, and April 16, 2006, as the timeframe when a majority of returns would be prepared in SPEC sites during 2006. Any returns prepared in SPEC sites outside of this time period were excluded from the sample frame for the 2006 test year.

In summary, the final sample frame for the 2006 test year consisted of all paper and electronic tax returns prepared in the 9,761 SPEC sites that were open during the 2005 filing season and that were open and reviewer-accessible between January 30 and April 16, 2006.

▶ Basic Review Process

The only way to evaluate the accuracy of a return prepared in a SPEC site is to physically travel to the site

location. The need to review cases onsite puts distinct boundaries around SPEC sampling options and will ultimately drive the resource requirements for any quality review process.

There is a basic framework for any official review of the quality of returns prepared in SPEC sites. SPEC reviewers will have to travel to a select group of accessible SPEC sites over the course of a defined review period. The reviewers will need a sample plan that identifies sites to be visited and a specific timeframe for each visit. Once at a site, reviewers will use a predefined case-selection technique to sample a designated number of returns prepared within the site. The results for each selected return will be recorded on a predesigned DCI.

While there is a somewhat rigid structure associated with reviewing SPEC return accuracy, there are aspects of the process that can be modified. The number of site visits, the number of returns reviewed during site visits, the timeline for site visits, the process used to select sites to be visited, and the actual information gathered for each selected return all have some level of flexibility associated with them. Starting with the basic review process and making adjustments where possible, SSS worked with SPEC personnel to design a sample for the 2006 test year that met SPEC's needs without overburdening available resources.

▶ Sample Design

Based on the statistically reliable estimates required by SPEC, the sampling frame, and the basic procedures involved with reviewing returns, it was decided to employ a "two-stage stratified random sample of unequal-sized clusters selected with probability proportional to estimated size (PPeS)" sampling methodology for the 2006 test year.

The sampling frame was stratified first by the three partner types and then by each of the 11 weeks included in the sample period, for a total of thirty-three mutually exclusive strata. Stratifying by type of partner was an estimate-driven decision. Due to variability in the number of returns prepared between partner types, stabilizing sample sizes by type was necessary to facilitate estimates for individual partner types. Stratifying by week was a resource-driven decision. It allowed control of sample sizes by week, which was necessary to streamline reviewer travel time without overburdening allocated resources.

A two-stage sampling approach was utilized during the 2006 test year. The primary sampling units (PSUs) were defined as individual sites. Because sites within a given stratum had varying numbers of returns prepared, PSUs were treated as unequal-sized clusters. In the first stage, a unique random sample of PSUs was selected within each stratum using a PPeS methodology. Sampling of PSUs was done with replacement.

Master File data were used as the source for estimated Measures of Size (MOS) in the first stage of all PPeS selection procedures. Returns filed electronically from SPEC sites post to the IRS Master File database approximately 2 weeks after the date they were prepared. Paper returns take approximately 6 weeks to post. Master File provides weekly reports containing "date-posted" information for all returns filed from individual SPEC sites. Using Master File data from the prior year (2005), SSS obtained MOS for individual PSUs by applying necessary adjustments to account for the discrepancy between the date posted and the date prepared.

A site visit was conducted for each PSU selected in the first stage. Site visits occurred during the specific weeks associated with each PSU's stratum. The basic sampling unit within a PSU was defined as a single paper or electronic tax return filed. An equal size subsample of returns was selected during each site visit. Sampling units were selected on a "first-come, first-served" basis by reviewers. Sampling at the second stage was done without replacement.

▶ Sample Size Determination

Several pieces of information were taken into account when determining the sample size for each stage of the 2006 sample plan. Working with SPEC personnel, SSS established that confidence intervals for all primary goal estimates should be at the 90-percent level with a 5-percent margin of error. The results from

the "shopping" review conducted during the 2005 filing season were used as very conservative predictors of the overall accuracy expected in 2006. In addition, resources restricted the number of visits that could be conducted each week and over the course of the filing season, while sites' hours of operation and time constraints limited the number of reviews that could be physically performed during a single visit.

Prior to determining the actual sample size for the first stage, it was decided to make sample sizes consistent across strata. In other words, the same number of visits would be conducted for each partner type each week. Streamlining the logistics of reviewer travel planning in this way was necessary to help minimize travel costs and to design a viable sample plan. To preserve the EPSEM nature of the PPeS design, it was also decided to make sample sizes consistent in the second stage. In other words, the same number of returns would be sampled and reviewed during every site visit.

Given these constraints, along with a lack of auxiliary and historical information about the sample frame, SSS utilized an unscientific ad hoc simulation process to determine the first and second stage sample sizes for the 2006 test year. SSS recommended that SPEC conduct 25 visits to each partner type each week and sample 3 returns during each visit during the 2006 Return Review. With 3 partner types and 11 weeks included in the review period (33 total strata), this design resulted in a total of 825 planned visits to be conducted and 2,475 returns to be sampled during 2006.

▶ Estimate and Margin of Error Calculations

The combined ratio estimator was used to calculate all primary and secondary goal estimates (see Estimates section). The generalization of the Hansen-Hurwitz estimator appropriate for two-stage cluster sampling was used to estimate the total accurate and the total applicable returns (or individual indicators), separately, across all relevant strata. A ratio estimate was then calculated by dividing these two estimated totals.

For each ratio estimator, the formula for the estimated variance of the Hansen-Hurwitz estimator for two-stage PPS sampling was used to estimate the variance and covariance of its numerator and denominator across all relevant strata. The estimated variance formula for a combined ratio estimator was then used to estimate the variance of the ratio estimator. The Korn-Graubard adaptation of the Exact binomial interval was then used to calculate the upper and lower bound of the 90-percent Exact confidence interval for the estimate.

Most IRS quality measures include point estimates and margins of error in reports. IRS executives and personnel have experience dealing with and interpreting results of this nature. For this reason, SSS opted to express SPEC confidence intervals as point estimates and margins of error. The midpoint and half-width of the 90-percent confidence intervals were reported as point estimates and margins of error, respectively.

▶ 2006 Results

The table below summarizes the primary goal estimates and margins of error calculated by SSS using the formulas described in Estimate and Margin of Error Calculations section. These estimates were provided to SPEC. However, because 2006 was a test year, these results were only used internally and were not provided to SPEC partners or other external stakeholders.

National Results—2006 Test Year

	Point Estimates[+]	Margin of Error[+]*
VITA	89.96%	2.98%
TCE	89.94%	2.73%
Military	90.46%	2.58%
All Partners	90.14%	1.86%

[+] Estimates inflated due to nonsampling error (see Weaknesses and Limitations).
* Assuming 90-percent confidence.

Similar results were also calculated for each of the secondary goals outlined in the Estimates section. This included estimates for each of the four areas and for each of the eight individual indicators on the DCI. All results were provided to SPEC.

▶ 2007 Filing Season

A second test of the new sampling methodology was conducted during the 2007 filing season. Based on findings from the 2006 test, some modifications were made to the sample design.

The first-stage sample size of 25 site visits per stratum required for the 2006 test was based on a conservative estimate of the actual accuracy of returns prepared in SPEC sites. However, estimates from the 2006 review allowed SSS to update sample sizes for the 2007 review. It was determined that the first-stage sample size could be reduced to 15 site visits per stratum in 2007.

SPEC sites can vary in size considerably. The average daily volume of returns prepared in a given site can vary from less than 1 to nearly 200. During the 2006 test, SPEC reviewers had difficulty finding and sampling the required three returns during visits to smaller sites. This resulted in missing data and an inefficient use of reviewer time. To alleviate the issues with obtaining samples from smaller sites, it was decided to remove smaller sites from the sample frame used for the 2007 test. More specifically, the final 2007 sample frame included only those sites that prepared at least 50 returns during the 11-week period of the 2006 filing season. One consequence of this decision is that estimates from the 2007 review will not represent the quality of returns prepared in smaller SPEC sites.

At the time this paper was written, all sample review and site volume data were not yet available. For this reason, point estimates and confidence intervals have not yet been calculated for the 2007 test year.

▶ Weaknesses and Limitations

The tests conducted during the 2006 and 2007 filing seasons established that SPEC is capable of successfully carrying out the new sampling methodology proposed by SSS. SPEC resources were able to complete the necessary site visits during the designated weeks and, with the exception of small sites, were able to consistently meet second-stage sampling requirements. However, the following weaknesses and limitations of the new design have proven to be unavoidable:

- Previsit procedures are a source of nonsampling error. SPEC's current relationship with partners requires that sites be notified about a return review visit 5 days in advance. Therefore, the level of service provided by volunteers on the day of a site visit may not be an accurate indicator of the level of service provided throughout the rest of the filing season. Influencing volunteer behavior by providing advanced notice of a site visit is a source of nonsampling error, which could positively skew estimates of quality under the new methodology.

- The makeup of the SPEC review team is a potential source of nonsampling error. Due to resource limitations, SPEC is unable to employ an independent team of SPEC reviewers to carry out site visits. Instead, visits are conducted by SPEC partner relationship managers located in each area of the country. Because these managers work with the SPEC partners on a regular basis, they may have difficulty reviewing sampled returns objectively. Manager bias cannot be verified, and its impact on the final results cannot be measured. Yet failing to employ an independent review team is a potential source of nonsampling error, which could positively skew estimates of quality under the new methodology.

- The process for evaluating returns may not capture all errors and is a source of nonsampling error. During a single case review, the reviewer does not witness the actual interaction between the taxpayer and the SPEC volunteer first-hand. Instead, to evaluate the accuracy of a prepared return, the reviewer compares the physical return prepared by the SPEC volunteer with all information provided by the taxpayer, including his or her answers to a tax-related questionnaire and all relevant tax documents. Reviewing returns after the fact could lead to reviewers missing certain errors on the return. For example, if volunteers improperly interview a taxpayer, they could either fail to obtain important information ini-

tially omitted by the taxpayer or overlook incorrect information provided by the taxpayer. Errors of this type will not be identified by a SPEC reviewer. Failing to detect all errors is a source of nonsampling error, which could positively skew estimates of quality under the new methodology.

- Estimates do not represent the entire population of SPEC sites. To alleviate the issue of obtaining adequate sample from smaller sites, it was decided to remove these sites from the sample frame. It has been determined that sampling small sites is not an efficient use of SPEC resources. Consequently, estimates of quality under the new methodology will not represent smaller sites.

- The timing of reports is not convenient. The new methodology requires weekly volumes to produce estimates. However, due to the discrepancy between the date posted and the date prepared for individual returns, all volumes are not available until 6 weeks after the end of filing season. Consequently, SPEC will not have a measure of their quality until well after the filing season is over. In addition, while the new methodology provides feedback for making adjustments for the following filing season, it does not provide information on a flow basis which can be used during the current review period.

Each of these weaknesses and limitations is inherent to the review process and sampling procedures associated with the proposed sampling methodology. Collectively, they may prevent the new design from being a viable long-term solution to SPEC's quality measurement needs.

▶ Future Plans

As shown in the table in the 2006 Results section, the results from the 2006 test year show estimates of quality near 90 percent. Preliminary results from 2007 appear to support these figures as well. However, these estimates are significantly higher than the qualitative results obtained from prior quality measurement efforts which utilized a "shopping" methodology. The gap between the expected and the actual results from the 2 test years may be attributed to some of the inherent problems discussed in theWeaknesses and Limitations. section

There is a strong indication that results from the new methodology are positively skewed. Because of this, the new methodology may not provide SPEC with a realistic assessment of their quality and may not allow them to accurately and consistently identify potential areas of improvement. While "shopping" is not considered statically reliable, results from SPEC's prior "shopping" reviews have proven useful in focusing improvement efforts.

Discussions between SPEC, IRS executives, and SSS are currently underway to weigh the pros and cons of both the old "shopping" technique and the new statistically valid sampling methodology tested during 2006 and 2007. The future direction of SPEC's quality measurement efforts will attempt to strike a balance between obtaining useful quality data and the efficient use of resources.