# A Quality Measures Plan Within the IRS: A Case Study

*Glenn D. White, Internal Revenue Service*

The Quality Management Information System (QMIS) measures the quality of work performed by Taxpayer Service Division, of the IRS. It produces a national accuracy rate for specific types of work or sources of information (SOI) -- namely: Directly Prepared Returns, Correspondence, Written Technical Referrals, and Total Accounts work. These data are also used to identify trends and training needs.

In October 1992, a new entry system was implemented nationwide. This system replaced an existing system that produced data that were neither timely nor reliable. The new system now produces a weekly weighted accuracy report. This system eliminated the concern that quality review data were not received in time to take corrective actions.

During fiscal year 1993, the Mid-Atlantic Region of the IRS established a Centralized Quality Review System. This system was developed to eliminate the influences of local conditions and managerial biases that were perceived to affect quality review accuracy rates. A single district was selected as the centralized site for all written work. As expected, the initial results showed a substantial drop in accuracy after implementation. Now, with full implementation for all written work, the accuracy rate is beginning to increase.

Mid-Atlantic's analysis of the data produced by centralization caused concerns about the consistency of the national accuracy data. Mid-Atlantic suggested that the National Office staff in Washington, DC develop a test to determine if quality review was consistent nationally and volunteered to provide unidentifiable cases for the test.

Using the cases provided by Mid-Atlantic, a Consistency Study was conducted nationwide. The IRS maintains 63 district offices across seven regions in the nation. Taxpayer Service Division provides toll-free telephone assistance at 32 of these sites. A minimum of one toll-free and one non toll-free site participated from each region. Each site was asked to review the same cases and return them to Washington, DC upon completion.

Because the Consistency Study was not able to test consistency in the review of on-line or telephone accounts, we designed a second study aimed at reviewing consistency of our on-line quality review. We asked North Atlantic Region to conduct this test, because they had access to a "QUEST" box, which can be attached to a district office's phone system, allowing the monitoring of live calls from a remote location.

This test was conducted in three phases. The QUEST box was moved from Brooklyn to Boston and to Buffalo, the three toll-free call sites in North Atlantic Region. To be consistent, one quality reviewer and a backup were selected in each site for all three phases of the test. During the third phase of the test, the quality reviewers from each site met in Washington, DC to test both the consistency of the monitors and the review codes used to describe the quality review of the work products. The backup reviewers were used to continue the test in the district offices.

This paper describes both of these studies and their findings.

## ■ Consistency Study

The Consistency Study was conducted to determine if inconsistencies were present in the quality review process. Because of the differences in accuracy rates from site to site, the sites questioned whether the errors charged by quality reviewers were consistent. In response to this concern, we developed an approach to verify the consistency of the application of the review procedures. The purpose of the study was to determine whether the errors charged by quality reviewers were consistent throughout all IRS sites.

The Mid-Atlantic Region provided five cases for each SOI, with the exception of on-line accounts, which were excluded since there is no paper trail available for review. The cases were 'sanitized' to conceal confidential information. Sites were given sixty days to complete and return the reviewed cases. Twenty-seven sites volunteered, 16 toll-free and 11 non toll-free sites. All sites reviewed the same cases.

### *General Results*

Because of time limitations, not every case was reviewed, as planned, by each site. The actual number of cases worked, by type of work and SOI Code, were:

| Type of Work -- | SOI Code | Actual Cases Worked |
|---|---|---|
| Correspondence | 30 | 98 |
| Directly Prepared Returns | 40 | 118 |
| On-Line Adjustment | 43 | 79 |
| Written Adjustment | 44 | 110 |
| Written Technical Referral | 62 | 118 |
| Written Account | 72 | 122 |
| Total | | 645 |

The concern that the QMIS accuracy rates were inconsistent was correct. Low match percentages across all SOIs were found. This indicated that the sites are not evaluating the cases consistently in accordance with predefined standards.

### *Specific Results*

#### Error Code Consistency

Mid-Atlantic Region developed Master Codes to classify each reviewed case. These were then compared to the results from each site.

Chart 1 shows the agreement rates of the QMIS personnel by SOI Code. The agreement rates are relatively constant over all the SOIs. Low percentages at this point illustrate confusion at the earliest steps of the review process. The actual percentage of site codes that agreed with the master code was 63.4 percent, shown as a horizontal line.

Chart 1: Error Agreement Rates with Master Codes by SOI Code

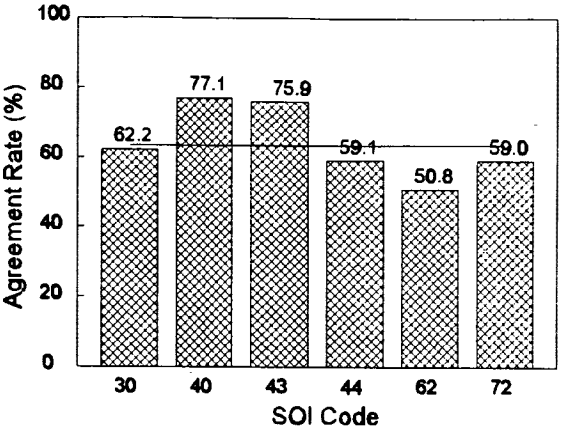
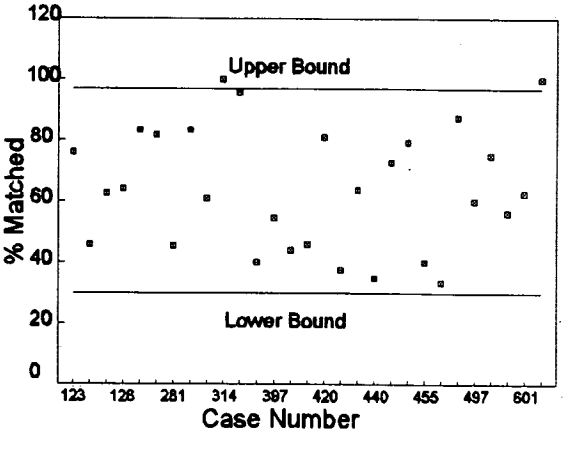



Chart 2 illustrates the 90 percent confidence interval with the percent matched for each of the case numbers. This chart indicates that the agreement rates are low over most of the case numbers, suggesting a lack of consistency in the review process. The rates ranged from 33.3 percent for case number 460 to 100 percent for case numbers 314 and 671. The standard error of this estimate, within case number, was 19.7 percent and the 90 percent confidence interval was (29.9, 96.9).

Chart 2: Percentages of Error Code Matches (90% CI)

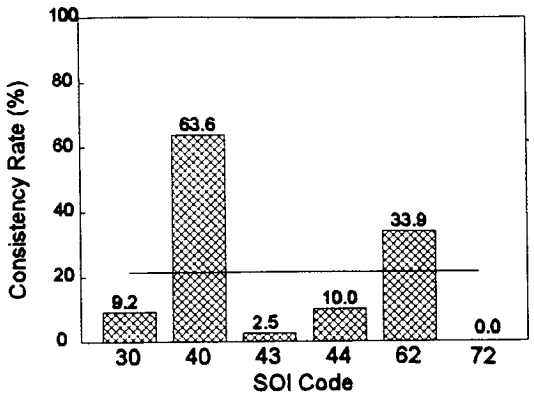



#### QIC Consistency Rates

Quality Issue Codes (QICs) were also developed that provided information about the subject matter of the call. An analysis was performed to identify

which QIC codes QMIS personnel were having problems identifying. A case was considered to be a match if all four of the site QICs matched the master QICs. Chart 3 shows the QIC consistency rates by SOI Code. These consistency rates are low and inconsistent over all the SOIs.

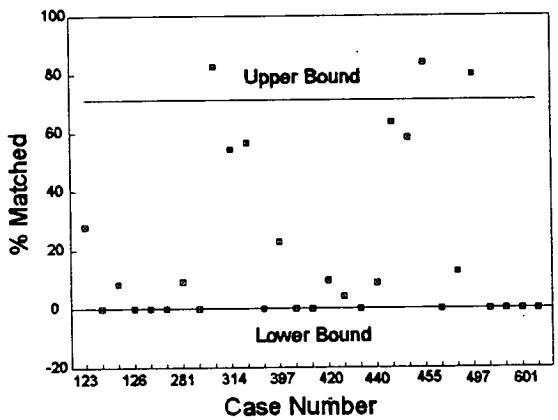
Chart 3 - QIC Consistency Rates by SOI Code




The actual percentage of site codes that completely agreed was 21.2 percent; ranging from 0.0 percent, for fourteen of the cases, to 84.0 percent, for case number 455. The standard error and 90 percent confidence interval were 29.4 percent and (0 to 71.2) percent, respectively.

Chart 4 shows the upper and lower bounds of the confidence interval and the percentage of matches by case number. The agreement rate between the sites and the master codes is low over most of the case numbers. This indicates a general lack of consistency in the review process.

Chart 4: Percentages of QIC Complete Matches (90% CI)




**Other Results**

Several other consistency and agreement rates were examined. One such analysis showed how close the site QICs were to a complete match to the master QICs. Information from this table was used to identify case numbers which might have an incorrect master QIC or where site training in this area was not adequate.

The Correspondence and Written Technical Referral cases would show the highest increase in the consistency rate if the rule for a match was changed from matching all four to matching at least three. The consistency rate for Correspondence would climb from 9.2 to 56.1 percent and for Written Technical Referrals from 33.9 to 89.0 percent.

Further analysis identified the QICs with minimal agreement among the sites. These codes should be looked at in greater detail to determine whether they were the correct QICs for these cases and, if so, whether proper training was given to the sites.

The percent agreement between the master QIC error codes and the sites for the most part was very low. In fact, the percentage of sites that matched the master QIC error codes was only 19.0 percent.

Another set of results showed which error codes within QICs were reported by the sites most frequently. These results could be used to further identify whether sites were considering other types of errors. Since there are a large number of different error codes for several of the QICs, it is apparent that there was a considerable amount of disagreement about the error codes to use.

### *Consistency Study Conclusions*

The results from this study indicate that the accuracy rates are inconsistent. The low match percentages indicate that the sites are not evaluating the cases consistently in accordance with predefined

standards. The following are several recommendations to improve the consistency of QMIS accuracy rates:

- Improve the training. Training appears to be an area in which substantial improvements need to be made. The reasons for the low agreement rates over most of the case numbers is a lack of consistency with the QMIS review process. This could be due to insufficient or incomplete training to all sites involved.

- Improve training materials. The Taxpayer Service Division has revised and rewritten all material used by the quality reviewer with a focus on consistency. Formal classroom training has begun with representatives from each district office.

- Review the cases to determine whether proper QIC and error codes were assigned to these cases. It would be worthwhile to review the individual cases, to determine if the proper QIC and error codes were assigned. There appears to be confusion among sites on evaluating these cases. Combining several of the error codes or procedural inquiries may alleviate some of the confusion.

- Conduct a retest at approximately the same time of year under similar constraints and conditions. A comparison could be done to determine whether any improvement to the QMIS rating consistency occurred.

The study verified the concern about the inconsistency of the quality review process. However, because the review cases were sanitized, with key information removed, the percentage of inconsistent cases found should be used as an indicator only. The sanitization required reviewers to make some assumptions, since insufficient data were included with the case. To accurately measure the percentage of inconsistency in the review process, future test cases must simulate actual cases.

## Remote QUEST Study

On-line accounts is an area of work (SOI) where inquiries are resolved while taxpayers are on the phone. There is no paper trail to be quality reviewed, so the work is reviewed by monitoring the call as it occurs. With available technology, district offices only have the ability to review their own on-line calls. Regions and Washington, DC cannot monitor districts without going to the call site to be monitored. There are the same concerns about the consistency of on-line accuracy rates as with written work.

The QUEST Study was developed to measure the difference in quality review accuracy rates between monitoring at the site and monitoring from a different location. North Atlantic Region was asked to participate in this study because they have a QUEST box which permits monitoring of calls in progress. However, the QUEST box is not used for general remote monitoring because of the cost of the unit and the excessive time involved in using the system to review cases. Representatives from North Atlantic Region and Washington , DC developed the procedures for the QUEST Study. Three phases were designed.

### *General Results*

Phase I of the test was conducted for five consecutive weeks. The accuracy rate reported by the reviewed site during that period was 91 percent, the accuracy rate produced from the remote monitoring site was 66 percent. Due to the large difference in accuracy rates, the test was expanded to the Boston and Buffalo Districts.

Phase II of the test was also conducted for five consecutive weeks. For this phase the Brooklyn District office was the remote monitoring site and Boston was the reviewed site. The QUEST box was moved from Brooklyn to Boston, prior to the start of Phase II. At the end of the test period, the reviewed site reported an accuracy of 91 percent, and the remote monitoring site reported 69 percent.

The QUEST box was next moved to Buffalo, the reviewed site, for Phase III. Both Boston and Brooklyn were remote monitoring sites for this phase. This phase of the test was scheduled for three weeks; with two remote monitoring sites the sample size could be doubled each week. At the end of the testing period, the reviewed sites reported an accuracy rate of

82 percent, and the remote monitoring sites reported a combined accuracy of 80 percent.

The first two phases of the tests validated the concern that on-line accounts' accuracy rates were not reliable because of inconsistencies in the review process. However, the third phase of the test showed very little difference in the remote on-line accuracy rates. If we assume that the remote monitors were more objective and consistent, the results of the three phases of the test indicate different review practices and/or standards in the three sites.

### *Specific Results*

Testing the difference between two sample proportions, where $p_1$ and $p_2$ are the accuracy estimates from the two respective populations, gives the following:

Phase I: Boston Monitoring Brooklyn

| Week Starting | Brooklyn | | Boston | |
|---|---|---|---|---|
| | Total | Ok | Total | Ok |
| 6/7 | 12 | 11 | 12 | 8 |
| 6/14 | 12 | 11 | 12 | 7 |
| 6/21 | 12 | 12 | 12 | 8 |
| 6/27 | 20 | 16 | 13 | 10 |
| 7/4 | 11 | 11 | 9 | 5 |
| Totals | 67 | 61 | 58 | 38 |

Accuracy $p_1 = 0.9104$ $p_2 = 0.6552$
Overall accuracy $p = 0.7920$
z score =3.51.

Phase II: Brooklyn Monitoring Boston

| Week Starting | Boston | | Brooklyn | |
|---|---|---|---|---|
| | Total | Ok | Total | Ok |
| 8/15 | 19 | 19 | 19 | 13 |
| 8/22 | 19 | 16 | 19 | 11 |
| 8/29 | 19 | 17 | 19 | 13 |
| 9/5 | 15 | 14 | 15 | 10 |
| 9/12 | 19 | 17 | 19 | 16 |
| Totals | 91 | 83 | 91 | 63 |

Accuracy $p_1 = 0.9121$ $p_2 = 0.6923$
Overall accuracy $p = 0.8022$
z score = 3.72.

Testing the following hypotheses:

$H_o$ = There is no difference between the two populations

versus

$H_a$ = There is a difference between the two populations

yields the same result from both phases. The null hypothesis, $H_o$, is rejected with near certainty. There is a significant difference between the accuracy estimates obtained from self-monitoring versus remote monitoring.

In other words, Phase I provides evidence that Brooklyn's accuracy, obtained from self-monitoring, was different from that obtained via monitoring by Boston, with 99.5 percent confidence. Similarly, Phase II provides that Boston's accuracy obtained from self-monitoring was different from that obtained via monitoring by Brooklyn, with even greater confidence.

Phase III: Boston and Brooklyn Monitoring Buffalo

The third phase of the QUEST study was having Boston and Brooklyn monitor Buffalo for three weeks. Testing first for any evidence of a difference between Boston and Brooklyn's monitoring showed:

| Time Period | Boston | | Brooklyn | |
|---|---|---|---|---|
| | Total | Ok | Total | Ok |
| 10/4-22 | 33 | 28 | 21 | 15 |

Accuracy $p_1 = 0.8485$ $p_2 = 0.7143$

Overall accuracy $p = 0.7963$
z score = 1.19.

Testing a similar hypothesis as above with a z score of 1.19 shows no evidence to suggest a difference between Boston and Brooklyn's accuracy rate for Buffalo.

Testing if there is a difference between the two sample proportions or, in other words, a difference between the accuracy rate Buffalo obtained by self-monitoring versus the combined rate that Boston and Brooklyn obtained by remote monitoring gives the following:

| Week Starting | Buffalo | | Boston/Brooklyn | |
|---|---|---|---|---|
| | Total | Ok | Total | Ok |
| 10/4 | 12 | 8 | weekly | |
| 10/11 | 10 | 10 | data not | |
| 10/18 | 12 | 10 | available | |
| Totals | 34 | 28 | 54 | 43 |

Accuracy $p_1 = 0.8235$ $p_2 = .7963$
Overall accuracy $p = 0.8068$
z score = .32.

Testing the following hypotheses

$H_o$ = There is no difference between the two populations

versus

$H_a$ = There is a difference between the two populations

we find the sample statistic, z, does not fall in the critical region. In other words, there is no evidence to suggest that Buffalo's accuracy rate is different from the accuracy rate that Boston and Brooklyn obtained while monitoring on the QUEST box.

### *Consistency of Monitors and Review Codes*

The quality reviewers from each of the three toll-free sites involved with the QUEST Study met in Washington, DC to measure how consistently they monitored on-line calls and applied the review codes to those calls.

On-line calls were monitored using the QUEST box connected to the phone lines in the Buffalo District office. For this study, we reviewed all calls coming into the site, not just account calls, due to the limited amount of monitoring time.

The three quality reviewers monitored 18 on-line calls. There was no discussion among the monitors while the calls were being monitored and evaluated. There was total agreement for all eighteen calls on the major category codes selection (QIC). Disagreement was found, however, for the more specific subcategory codes.

There was total agreement as to the case being correct or incorrect in 15 of the 18 calls monitored. Hence, there was an 83.3 percent agreement among the three monitors. The standard error on this estimate is about .09. In the three disagreed calls, two quality reviewers out of the three were in agreement. In two of these cases, agreement was total after group discussion. In the remaining disagreed call, the problem was technical expertise. This would not have occurred in a real review situation, because quality reviewers only work in their area of expertise.

The study showed that the percent of agreement is high when the monitoring is performed off-site, away from local influences. Although a sample of size 18 is not large, the 83.3 percent agreement would appear to be a high level of agreement. The three monitors appeared to be using the same criteria in evaluating the calls; hence, differences in the accuracy rates cannot readily be attributed to the monitors.

The quality review codes need to be reviewed to determine if definitions are clear and if there are too many or too few codes to select under a majority category.

## ■ Comparison of Consistency and QUEST Studies

Both the Consistency and QUEST Studies validate the concern that the accuracy rates produced by the quality review system are not consistent

among district offices. The Consistency Study focused on written work and the QUEST Study on online work.

The Consistency Study results showed lower percentages of consistency, because the review was performed in the district office. Centralization of quality review should help eliminate local bias and promote consistency of reviews.

The QUEST Study was conducted off-site, eliminating local influences and provides us with a baseline to measure future improvements in our online work.

Quality review will need to direct its efforts toward eliminating local influences from the system, where possible; providing consistent guidelines for reviewers; and evaluating the codes used to identify the various aspect of the review.

## ■ Acknowledgments

The author would like to thank Audrey Moore of Taxpayer Service Division Quality Review Team for providing background information, reviewing and interpreting the data. Special thanks are also made to Eric Falk of Statistics of Income for analysis of the Consistency Study. ■