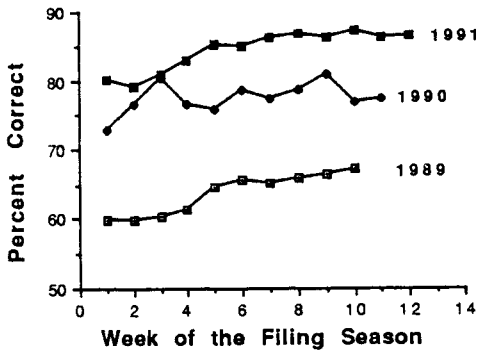
ASSESSING THE TEST USED IN THE MEASUREMENT OF IRS TELEPHONE ACCURACY

Robin Lee and Mary Batcher
Internal Revenue Service, SOI Division, R:S:P, P.O. Box 2608 Washington DC 20013-2608

Beginning in 1988, the Internal Revenue Service has operated a program of test calls – the Integrated Test Call Survey System (ITCSS) – designed to assess the accuracy of the information provided to the public by its telephone assistance program [1-4]. Although during the early years of measuring the accuracy, results were disappointing for the last two years, substantial improvements in measured accuracy have been realized (Figure 1).

Figure 1. National ITCSS Accuracy Rates



This improvement has been gratifying to see, but, unfortunately, it was not totally clear just how much of the change might have been influenced by the test itself or by changes in the operation. This paper describes some work directed at assessing the effect that the test itself has had on the measurement results.

Test Questions

The test used in the test call program consists of scripted questions about the filing of individual income tax returns. One of the early design considerations was how to strike the balance between measurement control and making the test call program more nearly reflect the actual interaction between the taxpayer seeking information and the IRS employee responding to the inquiry. Although some consideration was given to assessing accuracy by monitoring live calls, the complexity of scoring and the difficulties in ensuring comparability of measurement allowed only limited experimentation with monitoring of actual calls. The accuracy rate from monitoring did, however, correspond quite closely to the accuracy rate indicated by the test call program.

There were other options available for making the test questions more reflective of actual inquiries, without sacrificing measurement control. In the first year of test call operation, the initial test questions were written by subject matter experts, drawing upon the basic reference documents. Little was known about the difficulty level or complexity of actual requests for information.

For the second year, test questions were based on a sample of actual taxpayer inquiries. During the 1988 filing season (the period between January and mid-April when most people file their tax returns), the opening question asked by taxpayers was transcribed in a sample of incoming calls. The sample of transcribed calls, then, formed the basis for the development of new test questions. This was done in an attempt to make the test questions reflect actual inquiries, both in terms of content and difficulty. A note of caution is in order here. As part of the question development process, all test questions received IRS legal review and were then reviewed and agreed to by the General Accounting Office, in their role of monitoring the accuracy of the IRS telephone assistance at the request of Congress. During this review process, the initial transcribed questions were sometimes altered with the intent of making them more precise, or more difficult, or whatever the reviewers felt improved the question. The result was occasionally a question bearing little resemblence to the actual inquiry from a real person. These manipulations primarily affected the difficulty level of the question and the extent to which it was phrased like a real inquiry. The question topics were still closer to the real world we were trying to measure than was the original question set. Some of the initial questions were retained in the new question set, but only after comparing them to the transcriptions to ensure that an actual call had been received by the assistance service on that topic.

The test questions are scripted for the test callers, with background provided to allow them to respond to requests for further information from the telephone assistance staff. In general, background

information is required for the test questions, since the question itself does not present all pertinent information. The assistance staff must ascertain key background information to respond correctly to the question. If a correct response is provided and the necessary background information has not been elicited, it is considered a lucky guess and the call is scored as incorrect.

The test callers merely code the presence of certain requests for information, called probes, and responses provided by the telephone assistance. Scoring itself is accomplished by a computer program which identifies and tabulates the presence of combinations of probes and response points.

Composition of the Test

The test has changed throughout most of the operation of the test call program. Some of these changes were necessitated by changes in the tax law; most, however, were a result of attempts to improve the test call system during its early years. Because there were substantive changes in the test questions and basic operation of the test call program between 1988 and 1989, our interest was in examining the properties of the 1989, '90, and '91 tests.

In 1989, most of the test questions were based on transcribed conversations. However, about one-fourth of the questions were carried over from the 1988 test, yielding a mix of questions based on actual conversations and those based on rational consideration of the content area.

The goal for the 1989 question development process was to have at least two questions in each of 35 minor tax law categories, with more questions in high-volume or low-accuracy categories. This was not achieved; the question developers and reviewers had great difficulty reaching accord on the minimum of two questions per category and, indeed, were not able to develop and agree to any questions in three of the minor categories. Therefore, in 1990, the test design was based on the seven major categories, rather than minor categories. The 1991 test was only minimally changed from 1990, to revise a few questions that IRS judged were not working as intended.

In summary, then, for the three years that we considered, the test changed most between 1989 and 1990, with the 1991 items essentially unchanged from 1990. A core set of 17 items is common and unchanged throughout all three years.

Data

The data analyzed for all three years are cumulative accuracy rates for questions by sites: the total number of correct responses given by each site for each question divided by the total number of times the question was asked during the entire filing period.

In designing the ITCSS sample, sites offering toll-free telephone assistance were grouped into three categories: large, medium, and small, according to the volume of incoming calls. The weekly sample size, based upon actual call volume data, thus varied by site. Small sites, where sample size was smaller than the total number of questions, had missing data by design because the question by site sample was fixed for each caller every week in order to obtain more stable weekly trend estimates. Therefore, sites differed in the total number of questions asked and some questions were never asked in some sites. Consequently, the proportion of correct responses – instead of raw scores – had to be used and the missing data had to be imputed in 1990 and 1991.

For imputations, the mean accuracy rate for each question, adjusted for overall performance level of that site relative to the average of all sites, was used. For example, if site A had no data on question 3, the imputed value was the mean accuracy rate for question 3 multiplied by the ratio of site A's cumulative accuracy rate based on that of all items to the national cumulative accuracy. Comparing each site's cumulative percent correct, with and without imputations, revealed that most sites' overall accuracy rates changed very little as a result of imputation. (No imputation was necessary for 1989, where the sample was designed without the constraints stated above.)

Test Reliability

When we construct a test to measure a trait, the questions included in the test represent only a small sample of items drawn from a universe of many possible items [5].

A test is reliable if the observed scores on the test actually given reflect the scores that would have been obtained had this hypothetical universe of questions been given. Each observation has an observed score, x, which has two components: a "true score" component (t) and an "error score" component (e) [6], such that $x = t + e$.

The true score can be conceived as the mean of a large number of administrations of the test to the same person. With the following assumptions on e in the model:

(1) $e \sim N(0, \sigma_e^2)$,

(2) $COV(e_i, e_j) = 0$,

(3) $COV(t, e) = 0$,

we can show that

$E(x) = t$ and

$\sigma_x^2 = \sigma_t^2 + \sigma_e^2$.

The reliability coefficient is defined as the proportion of the "true variance" to the total observed variance:

$$\rho_{xx'} = \frac{\sigma_t^2}{\sigma_x^2}.$$

This formula, however, is not useful for estimating reliability, because neither σ_t^2 nor σ_e^2 is directly observable.

There are many different ways to estimate reliability but Cronbach's Alpha is one of the most commonly used coefficients as a measure of the internal consistency of the test [7]. Cronbach's Alpha can be interpreted as the correlation between this test and all other possible tests containing the same number of items, which could be constructed from a hypothetical universe of items that measure the same characteristic.

Internal Consistency of the Test

The internal consistency of the test is, in some sense, an indication of test efficiency. Cronbach's Alpha can be thought of as the average inter-item correlation and reflects the extent to which the individual test questions are homogeneous in measuring a trait of interest. Alpha is defined as follows:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum \sigma_I^2}{\sigma_x^2}\right],$$

where n is the number of questions, σ_I^2 is the item score variance, and σ_x^2 is the total score variance. When we calculated Alpha for the overall filing season measurement, the results were mixed. Table 1 presents Alphas for three years, 1989 through 1991.

Table 1. Coefficient Alpha, 1989-1991

Year	Alpha	n
1989	0.85	62
1990	0.87	43
1991	0.67	43

In general, the more questions there are, the more reliable the test is. Thus, it is interesting to note that the tests were comparable in reliability between 1989 and 1990, despite the considerable change in the test length and the questions themselves; whereas, between 1990 and 1991, with virtually no change in the test, Alpha dropped by 0.2.

It is unclear what contributed to this drop in the reliability estimates; however, it indicates that sites' performance measured on the accuracy dimension was not consistent from one question to another.

Item Analysis

We examined some traditional item analytic statistics. However, these should be interpreted with caution. Traditional item analysis methods were developed for achievement tests designed to provide maximum spread in scores. The deviation of the score from the overall group mean was the most important measure for an individual. Our case is actually one of measuring the extent to which the call sites have achieved an accuracy goal. In the ideal situation, all sites would measure 100 percent accuracy.

Item Difficulty. – The difficulty of a test item is the proportion of subjects taking the test who respond correctly to the item. When the goal of testing is to achieve maximum separation among those tested, the ideal item difficulty is .5. This allows for maximum variation among subjects. However, in our case, we were measuring the extent to which the desired accuracy goal had been achieved. We tended to use the item difficulties to identify test questions that

were deviant in their accuracy from the rest of the test, those questions that were either extremely difficult or extremely easy. The means and the ranges of item difficulties for the 1989-1991 tests are presented in Table 2, for the entire set of questions--the set that remained the same (set 1), and the set that changed (set 2).

Table 2. Item Difficulties,1989-1991

Item Set	1989		1990		1991	
	Mean (%)	Range	Mean (%)	Range	Mean (%)	Range
Set 1 ...	71.8	30-97	81.8	51-99	85.9	70-99
Set 2 ...	62.0	03-96	74.4	41-98	82.2	53-99
Total	64.7	03-97	77.3	41-99	83.9	70-99

Because the questions in set 1 did not change, the gain observed in this set is not confounded with the test. Thus, if the test questions were comparable in difficulty, set 2 should be expected to show a gain of approximately the same magnitude. The trend showing differential improvement between two sets of questions would imply that one set (i.e., set 2, consisting of new questions) contributed to the accuracy improvement more than the other; hence, some of the gain may have resulted from the reduced test difficulty. The gain in accuracy rates between years is summarized below in Table 3.

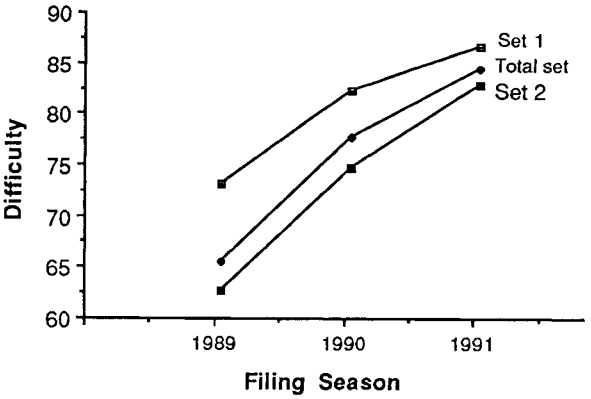
Table 3. Percent Changes in Cumulative Accuracy by Question Set.

Item Set	1989-1990	1990-1991	1989-1991
Set 1	10.0	4.1	14.1
Set 2	12.4	8.3	20.7
Total	12.6	6.6	19.2

There seems to be a differential gain, although it does not appear significant, especially in light of the reduced reliability of the 1991 data. The graphic presentation of the accuracy data (in Figure 2) shows that the lines are reasonably parallel, suggesting that the improvement was uniform across two sets of questions and can be regarded as real.

Item Discrimination.—Item discrimination is the extent to which test items separate the high scorers on the total test from the low scorers. This index can only achieve its maximum value for items at a difficulty level of .5. As the telephone assistance approaches its goal of 80 percent correct in 1990 and 85 percent correct in 1991, the achievable discrimination is reduced.

Figure 2. Difficulty Index by Year



The correlation of each question with the total test is a by-product of the calculation of Coefficient Alpha and is an indication of how well each question is performing when compared to the total test. Low correlation indicates that the item is not discriminating among the high and low call sites. Substantial negative correlation indicates that sites with low accuracy on the total test are performing better on the item than high accuracy sites. Means and ranges of item-total correlations for 1989-1991 are presented in Table 4.

Table 4. Means and Ranges of Item-Total Correlations.

Year	Correlation	
	Mean	Range
1989	0.27	-0.10 to 0.61
1990	0.36	-0.05 to 0.72
1991	0.19	-0.17 to 0.53

Again, a trend similar to that in reliability analysis is exhibited, in that the quality of measurement data on accuracy appeared to have dropped in 1991. This may be partly attributed to the reduced variance resulting from the high accuracy rate in 1991; however, the increased number of questions with negative correlations in 1991 indicates that the sites' performance on each item is not consistent with their performance on the test as a whole.

Conclusions and Future Plan

The tests were comparable in reliability between 1989 and 1990 and the improvement seems to reflect a real change in the quality of IRS service through the toll-free telephone assistance program. Also, the drop in Coefficient Alpha from 1990 to 1991 implies that the 1991 accuracy data were subject to more error than in 1990.

The sites' ranks on accuracy have not been stable, as shown in the rank order correlations between years in Table 5.

Table 5. Correlations between Call Site Ranks

Years	1990	1991
1989	0.59	0.61
1990	1.00	0.58

The shift in sites' ranking between years, however, shouldn't have affected the Coefficient Alpha, since it is a measure of internal consistency. It was speculated that the interaction among three components in the ITCSS measurement system – test questions, test callers and call sites – contributed to this drop in Alpha. A log-linear modelling approach will be applied to test this interaction effect explicitly.

The questions used to test the accuracy of the telephone assistance have remained virtually unchanged for two years, increasing the risk of question disclosure, as well as reflecting issues that people were asking about three years ago. The test will be revised, with many of the questions replaced and additional questions added. The item-total correlation will be used to guide the 1992 test development effort. Items from the 1991 test with negative correlation or correlation near zero will be revised or replaced, if necessary.

One area for additional research is exploration of other types of item-analytic indices – especially item discrimination indices – for the situation where the major purpose of testing is not to maximize the variability, but to measure the achievement level with a goal of 100 percent accuracy.

References

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

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

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

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

[5] Brennan, R. L. (1983). Elements of Generalizability Theory, Iowa City, Iowa: The American College Testing Program.

[6] Lord, F. M. and Novick, M. R. (1968). Statistical Theories of Mental Test Scores, Reading, Mass.: Addison-Wesley.

[7] Cronbach, L. J., Gleser, G. C., Nonda, H., & Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles, New York: Wiley.