

Forensic Monitoring System

Technical Report

September 2007

Number W0702

Chuah Siang Chee

American Institute of CPAs



American Institute of Certified Public Accountants

Forensic Monitoring System

Chuah Siang Chee

The Forensic Monitoring project was charted with developing a system to identify test items that may have been compromised, and thereby endanger the integrity of the Uniform CPA Examination. This is to be done primarily by monitoring the proportion of correct responses for each item across testing windows. This will allow us to monitor the performance of items across time and to identify unusual fluctuations in item performance. If the proportion of correct responses is above and beyond the expected response rate, it would trigger further investigation into the integrity of the items. This would allow us to make a decision on retiring the item from the item pool.

Forensic analysis is conducted on the multiple-choice items and simulation items for all sections of the Uniform CPA Examination. The Uniform CPA Examination consists of four sections: audit and attestation (AUD), financial accounting and reporting (FAR), regulation (REG), and business environment and concepts (BEC). Three of the sections (AUD, FAR, and REG) consist of an adaptive multi-stage test (MST) for the multiple choice items, two simulation items, and constructed response items. The BEC currently has multiple choice items that are not administered using an adaptive algorithm. There are no simulation items associated with the BEC presently.

Description of System

This system was built on a Microsoft Access platform. It combines historical data across all previously administered CBT administrations of the Uniform CPA Examination. Initial data processing is done using SAS to generate text files which are then loaded on the Microsoft Access database. Currently, new data for each window is appended to the database at the end of each test window.

The Forensic Monitoring system tracks the proportion of examinees that correctly answer a particular item, or p-value. However, natural fluctuations in the abilities of examinees across test windows as well as restriction of range for multiple choice items because of the adaptive administration confound direct assessment of the proportion correct scores. In order to resolve these limitations, the observed proportion correct scores is contrasted with the expected proportion correct scores in order to determine if scores are within tolerances. The expected proportion correct score is calculated based on the IRT parameters and conditional upon the ability level of the examinee administered the specific test item. Difference scores from the expected minus the observed proportion correct scores allow us to control for natural fluctuations in examinee ability and the range restrictions from multi-stage testing. If ability-adjusted difference scores are not used, the tolerance levels would require a wide range, or increase risk of false positives would occur. This would in turn reduce the sensitivity of the forensic monitoring system to detect security breaches.

Data in the Forensic Monitoring System is partitioned according to months. Each test window consists of two months of testing before the item pool is replaced. The rationale for this decision was to detect if examinees might be harvesting items at the beginning of a test window

and distributing the information to conspirators later in the test window. By partitioning the first and second month of testing, the system would be able to detect significant compromises in the security of the test if such a situation were to occur.

Identification of Suspect Items

Identification of suspect items in the Forensic Monitoring software currently requires human interpretation. There are no strict rules as to what dictates a suspect item, we have several constraints proposed for identifying suspect items.

1. Real p-values are significantly higher than expected p-values.
2. Sustained increase of real p-values over expected p-values three or more testing windows.
3. Significant increase in real p-values in the second month of test window.

Several charts are provided to track the performance of a particular item across windows. These charts include the expected and observed p-value. Also included is the standard error of the p-value. This allows the number of examinees to be taken into consideration when interpretation results and serves as a guide in interpreting if there are significant differences between the expected and observed p-value.

The process of identifying suspect items is time consuming because each item must be individually inspected. In order to improve efficiency the Forensic Monitoring software provides several statistics to restrict the list of potential items. These include:

1. The absolute size of the fluctuations in the observed p-value. If there is excessive fluctuation, it may indicate that something has occurred to compromise or change the item from previous administrations.

2. The absolute size of the fluctuations in the difference between the expected and observed p-value. This is akin to fluctuations of the observed p-value, but takes into account the fluctuations of ability of examinees.
3. The difference between the expected and the observed p-value for the last window the item was administered in. This is provided in order to limit searches to the current or latest test window.
4. The difference between the first window difference score (which is the difference between expected and observed p-value) and the last window difference score, in which the item was administered.

Below are samples of the charts available to interpretation. Figure 1 shows the difference between the expected and observed p-value, and standard error of an item across several test windows. Each data point represents a month of testing. Therefore, 05Q21 represents the first month of testing for the second quarter for the year 2004. As we can see from the chart, the difference between the expected and observed p-values is generally within a standard error. There does not appear to be a significant difference between the expected and observed p-value in the example given.

Figure 1.
Difference Between Expected and Observed P-value and Standard Error of a Test Item by Window

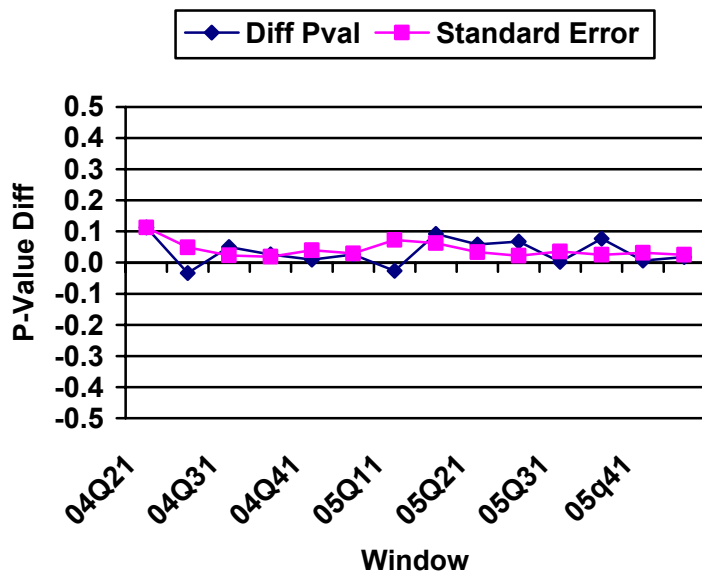
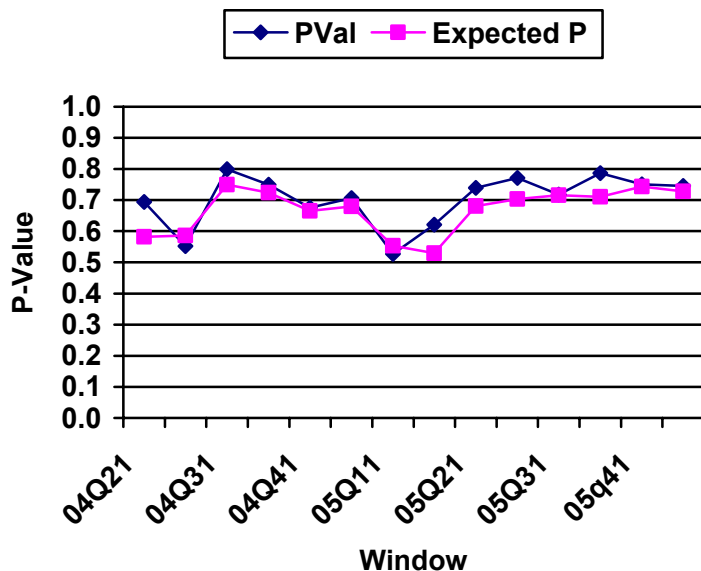


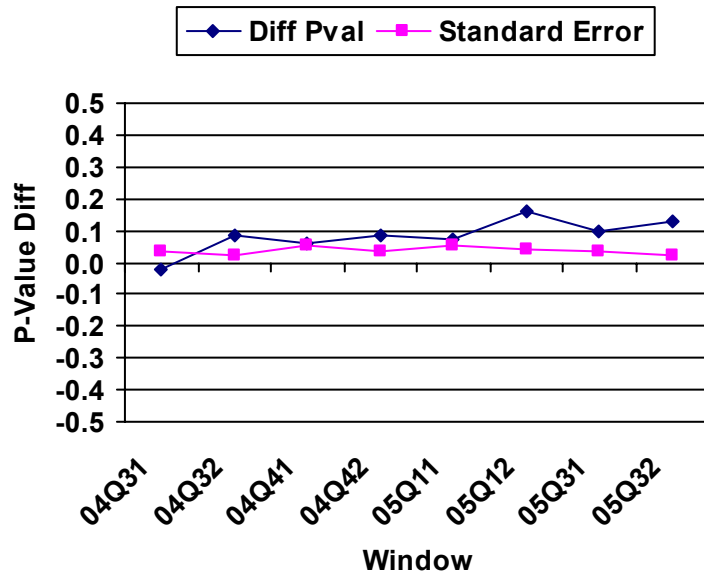
Figure 2 shows the expected and observed p-value for a test item across test windows. There does not appear to be a significant deviation between the expected and observed p-value. Interpretation of the observed without the inclusion of the expected p-value would be difficult because the p-value fluctuates considerably in this example. Therefore, including the expected p-value is an important instrument for interpreting results for the Forensic Monitoring system.

Figure 2.
The Expected P-value and observed P-value of a Test Item by Window



The Forensic Monitoring system is currently operational. Analysis of the results beginning 04Q2 to the last completed window (05Q4) is complete. Results suggest that there are no significant compromises in test security. A small number of items have been flagged as suspect items, 4 items out of the item pool. Figure 3 illustrates the performance of a particular item that has been identified as suspect. There appears to be a slight increase in the proportion of examinees correctly answering the item over time.

Figure 3.
 Difference Between Expected and Observed P-value and Standard Error of a Suspect
 Test Item



However, such small fluctuations do not necessarily mean that the item has been compromised. The p-value is expected to fluctuate to a small degree across windows. In addition, the small number of items that were flagged as suspect items suggests that there is currently no significant compromise to the security of the test.

Other Applications

In addition identifying items that may be compromised the Forensic Monitoring program has other potential applications. These include the ability to identify items where the current accounting rules pertaining to that item may have changed or to identify inconsistencies between item parameters and data. The Uniform CPA Examination is based on governmental rules that frequently change. Consequently, there is a need to actively perform obsolescence analysis of

the items in order to assure that the examination has kept up with the code. If there has been a change and the keyed response is no longer correct, the item statistics for that item might reflect this discrepancy. The system could also be used to study the lifecycle of the item. Indications that an item properties change over time might inform policy decisions about item use and retirement.

Schedule of Analysis

It is proposed that the Forensic Monitoring analysis be run at the completion of each testing window. This is the most opportune to run the analysis because all item parameters and statistics will be available. It is also the soonest that the analysis can be conducted. Conducting the analysis in the middle of the test window, before the completion of the window would exclude items that might not have their final item parameters.

Procedural Recommendations

For the Forensic Monitoring program to be effective, there need to be procedures in place to respond once items have been flagged as behaving suspiciously. Below are several recommendations for procedures.

- Having a second psychometrician to evaluate items already flagged as being suspicious. The identification of suspect items is a subjective process, and having concurrence should improve the process.
- Once an item is flagged as suspicious, a test developer should review the identified item in order to determine if there is a reason for the behavior of the item statistics. If no

acceptable reason can explain the item behavior, the status of the item should be escalated in order to determine an appropriate response.

- Once an item's status has been escalated, a process needs to be in place to retire an item, or in the extreme case, to not score the item. In order to implement such a process, a senior manager should be designated with the task of approving the decision to permanently/temporarily retire an item, or not to score the item. The procedure also requires that the personnel from content and production be informed if an item has been retired from the item pool. Accordingly, if an item is marked as a do not score, the appropriate personnel from production need to be informed to make the necessary changes to the scoring algorithm.

Future Enhancements

Future enhancements under development include the analysis by test center. This would enable the system to detect conspiracies that are localized to certain areas or districts. The rationale is that, if there is a local organization that has somehow compromised the security of the test, many of the students would be geographically close. Consequently, test scores for a local test center should be higher than surrounding areas.

One current shortcoming of the system is its inability to identify individual examinees that might be cheating. Research is currently under consideration in order to determine a manner in which individual cheaters might be identified.