Report No. SR01-10-02

QA/QC Procedures Based on Program Data and Statistical Process Control Methods for I/M Programs

prepared for:

U.S. Environmental Protection Agency
Certification and Compliance Division

October 2001

prepared by:

Sierra Research, Inc.
1801 J Street
Sacramento, California 95814
(916) 444-6666

Report No. SR01-10-02

QA/QC Procedures Based on Program Data and Statistical Process Control Methods for I/M Programs

prepared for:

U.S. Environmental Protection Agency
Certification and Compliance Division

Under Contract No. 68-C7-0051
Work Assignment No. 3-03

October 30, 2001

prepared by:

Richard W. Joy
Garrett D. Torgerson
Michael St. Denis

Sierra Research, Inc.
1801 J Street
Sacramento, CA 95814
(916) 444-6666

QA/QC Procedures Based on Program Data And Statistical Process Control Methods for I/M Programs

Table of Contents

page

Table of Contents
(continued)

page

1. INTRODUCTION

Over the past 10 years, a number of "enhanced" vehicle emissions inspection and maintenance (I/M) programs have been implemented, including programs with both centralized and decentralized inspection networks. Key design elements incorporated into many decentralized enhanced programs include the establishment of a Vehicle Inspection Database (VID) and automated electronic transmission (ET) of all test data on a real-time basis to the VID. Over a dozen states, including some with basic inspection programs, have now incorporated ET and a VID into their decentralized I/M programs. While not usually considered as such, VIDs and electronic transmission of inspection data are also typical elements of centralized inspection programs (i.e., test data are routinely transferred to a central database). More than a dozen additional states* and the District of Columbia have either contractor- or government-operated centralized inspection programs, all of which are presumed to include electronic data transmission to a central database.

The widespread implementation of ET systems has greatly increased I/M programs' accessibility to the resulting vehicle inspection data and other information (e.g., equipment calibration results) recorded by the inspection systems and transmitted to the VID. This in turn has raised state expectations regarding the potential benefits of analyzing the resulting data for a variety of reasons, including quality control/quality assurance (QA/QC) purposes. One key area of interest, particularly for decentralized programs, is conducting statistical analyses on the data in order to identify potential problem test stations and inspectors. Analytical software, typically referred to as "triggers" software,** has therefore been developed and is being used by some states to evaluate the success of test stations and inspectors relative to a range of individual performance indicators. A few states are also beginning to use triggers software to evaluate equipment performance.

Generally, states are concerned that any public release of specific details regarding their triggers software could potentially result in inspection stations learning how the data are being analyzed. This could in turn lead to stations that might otherwise be identified by the software as problem performers intentionally modifying their behavior to avoid detection. As a result, it is often difficult for states to learn from one another in this area.

*Throughout this report the word "state" is used to refer to the governmental entity in charge of an I/M program, although some programs are actually administered at the county or municipal level.

**So named because it involves identifying a level for various program variables (e.g., failure rate as a function of model year) at which a more detailed investigation is "triggered."

Similarly, no guidance has been available to-date from EPA on how to integrate trigger approaches into automated ET systems and VIDs.

EPA has become interested in developing guidance to aid states in designing, implementing, and using such triggers. Accordingly, Sierra was issued Work Assignment 3-03 to develop and provide EPA with draft recommended guidance aimed at assisting states in using triggers and statistical process control (SPC) analysis to identify potential problem inspection stations, inspectors, and test systems. This report was prepared in response to the work assignment. Specifically, the report addresses such issues as the types of possible triggers and "control" charts, analytical issues to be considered in developing trigger calculations, methods for reporting the results (including control charting), and potential uses of the trigger results. It also provides recommendations regarding the structure of an effective overall QA/QC system.

A number of states were contacted and subsequently provided information that was used in preparing this report. Per Sierra's discussions with the individual states, the general approach that is used herein is to refrain from attributing detailed information regarding specific triggers to particular states or I/M programs. State staff were also reluctant to discuss problems encountered in implementing existing triggers systems or other sensitive issues, due to concerns regarding possible negative consequences to their programs of releasing this information. Therefore, as agreed with the states, this information is also presented without attribution.

The discussions contained in the report concentrate on VID-based decentralized program implementation issues, since this is the expected primary focus of any triggers software. Notwithstanding this, many of the analytical techniques and system approaches described herein may also have application to centralized inspection programs. In addition, the same techniques could be applied to decentralized programs that involve data collection via retrieval of computer diskettes from the individual test systems (assuming the retrieved data are subsequently entered into some sort of overall database).

Organization of the Report

Section 2 provides more detailed background information and a general overview of the issues and concepts discussed in the report. It provides an overview of the extensive discussions on these items contained in subsequent sections.

Section 3 describes potential station and inspector triggers, and is aimed at the full range of existing I/M test procedures and program designs. A complete description of each trigger, its purpose, and the data required for the trigger is provided.

Section 4 is similar to Section 3, but is focused on equipment-related triggers that are designed to identify instances of excessive degradation in performance among the full range of existing I/M test equipment.

Section 5 contains separate "menu" listings of all station/inspector and equipment triggers, as well as matrices that show those station/inspector and equipment triggers that are generally applicable to each type of existing I/M test or program design. The section is designed as an easy reference guide that the various types of I/M programs can use to more quickly focus on those triggers that are relevant to their program.

Section 6 presents additional detail on reporting methods, as well as the overall structure of an effective QA/QC system. The section also includes detailed discussions regarding the use of VID messaging and control charts as reporting tools. In addition, it provides further insight into other potential uses of the triggers results.

Section 7 describes possible analysis approaches, analytical issues, tools, or specialized skills that may be needed and are available to implement a triggers program or individual triggers, system security, and other related issues. The section also presents recommended formulas to use in the trigger calculations and an example method for weighting trigger results to produce composite ratings.

Section 8 discusses the need to incorporate some method of performance feedback into the system to evaluate the effectiveness of selected triggers in correctly identifying poor performers, and fine-tune the triggers to maximize their effectiveness in identifying true problem performers.

Section 9 provides a list of references used in the project.

###

2. BACKGROUND AND GENERAL OVERVIEW

"Triggers" software is a powerful QA/QC tool that has the potential to provide significant benefits, including improved identification of underperforming stations and inspectors, advanced diagnosis of impending test system problems, a possible reduction in the required level of station and/or equipment audits, and ultimately improved program performance at a lower oversight cost. The overall objective of this report is to describe how states can conduct triggers and other related analyses to produce a manageable number of figures and tables for I/M staff to use in improving program performance. The exact scope (e.g., number and type) of such outputs will depend on the size of the program and management staff, and the degree to which the staff is interested in reviewing and using such triggers results.

In keeping with the goal of producing easy-to-understand results that would be most useful to I/M management staff, all trigger methods presented in the report are standardized to a trigger index scale of 0-100, with scores closer to zero indicating poor performance and scores closer to 100 indicating high performance. This approach ensures that all trigger results can be compared on an equal basis.

Triggers

California, which was the first state with a decentralized I/M program to implement a VID and an accompanying ET system, also pioneered the development of software designed to perform statistical analyses of program data. This involves the use of triggers software that was developed by the California Bureau of Automotive Repair (BAR). Triggers software typically includes multiple algorithms, each of which is designed to evaluate the success of test stations and inspectors relative to an individual performance indicator. For example, a very simple indicator would be the I/M failure rate recorded by a station's inspection system, relative to the network-wide failure rate. Some states modify the individual triggers incorporated into the software package on an on-going basis, as deemed necessary through field experience (e.g., in determining which are the most effective in identifying problem performance) and other factors.

Each of the triggers algorithms is run on a routine basis on the contents of the test record data files that are automatically transmitted to the VID by the individual test systems. For each of the performance indicators, individual trigger scores are computed and then translated into index numbers for every station in the program. Stations with low test volumes and low subgroup volumes (i.e., for certain model years) are normally excluded from the applicable analyses to ensure that statistically valid results are produced. The resulting index numbers are used to rank all the I/M stations in order of performance.

Details regarding the triggers software currently in use are not publicly available. Generally, the I/M programs are concerned that any public release of these details could potentially result in inspection stations learning how the data are being analyzed. This could in turn lead to stations that might otherwise be identified by the software as problem performers intentionally modifying their behavior to avoid detection without actually performing proper inspections. As a result, it is often difficult for states to learn from one another in this area, with the overall objective of developing the most effective triggers system possible. Similarly, no guidance has been available to date from EPA on how to integrate trigger approaches into automated ET systems and VIDs.

Equipment-Related Triggers - Over the last 18 months, Sierra Research (Sierra) has been assisting some states in beginning to use triggers to track equipment performance based on the calibration results recorded on the test systems. The I/M programs in these states span nearly the full range of program designs and test procedures.* The intent of these efforts is to develop methods for identifying test systems that need to be serviced or replaced before the systems begin producing inaccurate test results. As an initial step, several parties involved in various aspects of I/M programs were contacted to determine if any other decentralized programs were actively reviewing or using calibration data for the envisioned purpose. The parties that were contacted and the results of those contacts are summarized below.

1. BAR staff indicated that they had only recently begun to look at the calibration files and had not developed any calibration file tracking or analysis tools.[1]**

2. Snap-On Diagnostics was contacted since the company has provided equipment to most decentralized I/M programs in the U.S. Snap-On indicated it was unaware of any states that had reviewed or were in the process of reviewing calibration data.[2]

3. MCI Worldcom was also contacted because it has developed and manages VIDs and electronic transmission systems, including the collection of calibration data, for several states. MCI Worldcom staff indicated that they had not processed any requests for calibration data or analysis of calibration records from states in which the company operates VIDs (i.e., they had never been asked for calibration data from any of the states).[3]

The above contacts indicate that no one is making use of the calibration records in any significant way. All of the parties that were contacted had the same reaction—they had

*This includes programs with (1) both centralized and decentralized inspection facilities; (2) two-speed idle (TSI), acceleration simulation mode (ASM) and transient loaded mode tailpipe tests; (3) underhood checks, functional gas cap and OBDII testing; and (4) contractor and state-operated VIDs and ET systems.

**Superscripts denote references listed in Section 9.

not realized that no one uses the calibration files collected automatically in all BAR90/BAR97 programs. It thus appears that the states with which Sierra has been working are the first to attempt to use the contents of these files to improve the performance of their decentralized emissions testing program.

Despite the lack of previous efforts in this area, equipment-related triggers have significant potential for identifying impending problems with individual system components such as gas analysis benches, dynamometers, gas cap testers, etc. These problems are expected to occur independently; thus, the development of a series of individual triggers aimed at each of the various components appears most promising.

Repair performance could potentially be the subject of a third set of triggers. However, the amount and quality of repair-related data typically recorded in vehicle inspection programs are low, in which case it becomes very difficult to generate meaningful results. No repair-related triggers are addressed in this report.

Control Charts

A QA/QC element that is used in at least some centralized I/M programs is statistical process control (SPC). SPC software is used to develop what are referred to as "control charts." A control chart is a statistical quality control tool, developed by W.A. Shewhart in the 1930s, that has historically been used in manufacturing processes to identify and correct excessive variation (i.e., that occurs outside of a stable "system of chance causes"). A process is described as "in control" when a stable system of chance causes seems to be operating. When data points occur outside the control limits on the charts, the process is considered "out of control." These points indicate the presence of assignable causes of variation in the production process (i.e., factors contributing to a variation in quality that should be possible to identify and correct). Control charts work best when used in connection with processes that produce relatively homogeneous data; the inherent "scatter" in more heterogeneous data makes control charting less effective in this environment.

Control charts are used to track both "variables" and "attributes." As implied by their name, variables are recorded data that vary by number or degree (e.g., the number of certificates issued to a vehicle or the amount of HC measured in the vehicle's exhaust). Attributes are data records that are either "yes" or "no" (or "pass" or "fail"), such as whether a vehicle passes the I/M test.

A key idea in the Shewhart method is the division of observations into what are called "rational subgroups." These subgroups should be selected in a way that makes each subgroup as homogenous as possible, and gives the maximum opportunity for variation from one subgroup to another. It is also important to preserve the order in which the observations are made in recording data on the control charts.

The use of control charts is included in section 85.2234, IM240 Test Quality Control Requirements, in EPA's IM240 & Evap Technical Guidance.[4] This guidance indicates that control charts and SPC theory should be used "to determine, forecast, and maintain performance of each test lane, each facility, and all facilities in a given network." As implied by these provisions, SPC analysis is best suited to the evaluation of equipment-related problems that are likely to develop over time as components deteriorate or fail. A key advantage of control charts is their ability to visually display trends in performance over time.

There are several reasons why SPC analysis is not suitable for evaluating station or inspector performance. Although I/M test results may appear homogeneous when very large samples are analyzed, heterogeneous results (which are incompatible with SPC analysis) can occur with smaller samples. This is caused by several factors, including differences in vehicle age, vehicle and engine type (e.g., passenger cars versus trucks), the emissions standards to which individual vehicles are certified when new, mileage, types of emissions control equipment on the vehicles, the geographic area in which a station is located (which can have a significant impact on the distribution of test vehicles), and the level of preventive or restorative maintenance that has been performed on each vehicle.

Some of these factors can be eliminated through data subgrouping (i.e., individual subgroups can be selected based on vehicle type, model year range, certification standard, and mileage). Such subgroupings, however, geometrically expand the number of control charts that would be generated, thus making it much more difficult for I/M program staff to easily evaluate inspection station or inspector performance. In addition, other factors (e.g., level of vehicle maintenance) that can have a tremendous effect on emissions cannot be accounted for in the selection of subgroups. This is because there is no way to identify and split test vehicles into appropriate subgroups based on such factors (e.g., those that have been subjected to various levels of maintenance). Finally, SPC analysis is only effective in identifying changes in performance over time. An inspection station that consistently performs improper tests will not be identified.

As noted above, SPC analysis is best suited to the evaluation of equipment-related problems that are likely to develop over time as components deteriorate or fail. By design, equipment performance should be homogeneous among test systems and over time, which means that control charts would be an excellent means of tracking and identifying degradation in performance among individual test systems. This has been done to some extent by the existing centralized contractors based on the provisions of EPA's IM240 guidance; however, the details of how they have implemented SPC analysis are not publicly available.[5] In addition, little attempt has been made to implement this QA/QC element on a decentralized basis.

State Survey Results

As an initial element in this project, state I/M programs believed to potentially be operating triggers systems or conducting SPC analysis were contacted for two primary purposes. First, the states were asked if they were willing to provide detailed information on their triggers systems, etc., that could be used in developing EPA guidance on this subject. As part of this effort, a clear understanding was reached with each state regarding the extent to which any information it provided would be incorporated into the guidance. The general approach that was agreed upon was to refrain from attributing detailed information regarding specific triggers to particular state I/M programs.

Second, discussions were held with staff from each I/M program regarding what they believed should be included in the guidance, with the overall objective being to maximize the utility of the resulting document. General items of discussion included the following:

- Topics that each state felt should be included in the guidance.
- The level of detail that should be provided.
- Approaches to structuring the guidance and presenting the information in the most usable format.

Results from the discussions with the individual states have been incorporated into the contents of this report. Overall findings from these discussions[6] include the following:

1. Only a few programs are running station/inspector triggers and even fewer (only those working with Sierra) are running equipment triggers.
2. States that are running station/inspector triggers are typically doing so on either a monthly or quarterly basis.
3. Those states (or their VID contractors) that are running triggers are doing so almost entirely through the ad-hoc querying or pre-defined reporting applications built into the VID. Trigger system outputs consist almost entirely of simple tabular listings/rankings. Existing systems are relatively simple and do not include much analytical or reporting functionality.
4. The states generally think their triggers programs are working well and that it is relatively easy to identify cheating stations. Therefore, there may be limited interest among these programs in implementing additional triggers or enhanced functionality described in this report.
5. Some states have delegated the entire QA/QA system to a program management contractor and may therefore never even see the triggers results; e.g., the contractor reviews the results and uses them to target covert audits on an independent basis. Other states, even some with management contractors, are much more involved in the design and operation of the entire QA/QC system.
6. One state has currently suspended much of its use of triggers. According to this state, program management is focused primarily on working with and counseling stations to improve their performance, rather than taking a harder enforcement stance.

7. States are generally interested in what triggers other programs are running and what information they hope to get from these triggers. One state expressed particular interest in seeing a robust way to compare emissions readings using software that does not require a large amount of computing resources.

Analytical Approaches and Issues

There are a number of possible approaches that can be used to perform triggers analysis or other statistical analyses of I/M program data designed to identify potential performance problems with stations, inspectors, and test systems. A key issue in deciding which approach to use is whether these analyses will be incorporated into VID software or run external to the VID.

States that have attempted to conduct data-intensive analyses on the VID have found this approach to be problematic due to significant limitations with the data analysis applications incorporated into the VID and a substantial resulting draw-down in system resources.[6] This has led most programs to run such analyses on the backup or "mirror" VID typically included in the system design. Other programs are considering performing these analyses using commercially available statistical analysis software to avoid these problems and take advantage of the robustness of such software.

In addition to the general issue of how the overall triggers or SPC system will be designed and run, there are a number of specific analytical issues that need to be considered in developing the system. These include ensuring the use of statistically significant minimum sample sizes, normalizing results to a consistent basis for relative comparisons, assigning weightings to multiple trigger results in developing a composite trigger result, and assuring an adequate range of trigger results to facilitate identification of poor performers. All of these issues are addressed in Section 7.

Reporting

Possible reporting of triggers and other results includes both graphical and tabular outputs. The simplest approach would be to simply output tabular lists of individual triggers scores for each of the applicable inspection stations, inspectors or test systems, which is what most states are currently doing. This type of approach would also be most compatible with using a database application to run triggers on the VID or another similar database (e.g., the backup VID). Adding graphical output capability to a database application would be more difficult and is likely to require some type of third party software.

Notwithstanding the fact that most states are producing only tabular summaries, Sierra's and its state clients' experience to date shows that graphical outputs of trigger results are critical to:

1. Determining if the trigger has been properly designed;

2. Determining whether the resulting index scores are being correctly calculated and show sufficient range to allow the poor performers to be easily identified; and

3. Getting a quick understanding regarding the number of inspection stations that appear to need further investigation.

To illustrate this issue, the histograms shown in Figures 1 and 2 were produced using trigger calculations developed by Sierra and actual test results from an enhanced I/M program. The two figures are based on a trigger index scale of 0-100, with scores closer to zero indicating poor performance and scores closer to 100 indicating high performance. Both figures show a wide range of distributions in the actual index scores. .

Figure 1

Histogram of Offline Test Frequency Index by Station

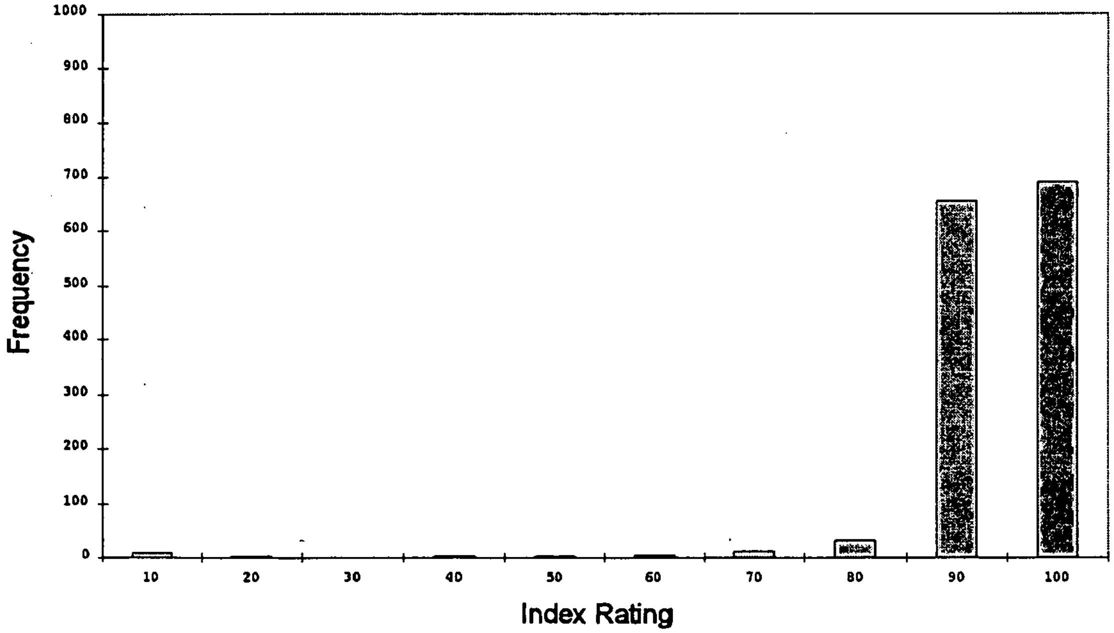


Figure 2

Histogram of Untestable on Dynamometer
Index Frequency by Station

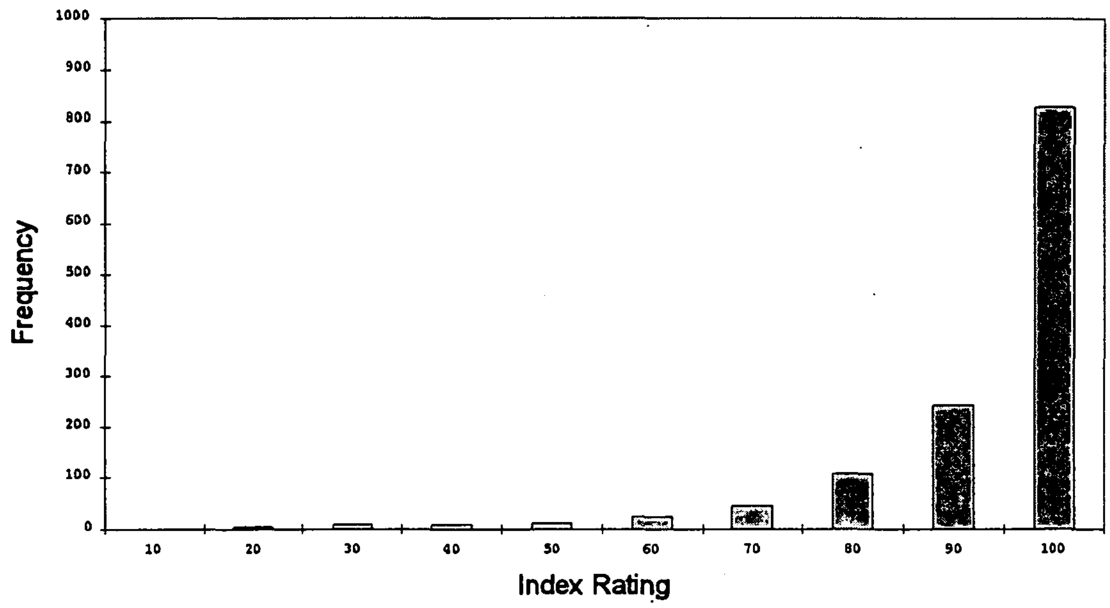


Figure 1 presents results for an offline test trigger. Most inspection stations have relatively high index scores, meaning they performed mainly online tests. The results also suggest that an index score of 80 or less could be used to identify the small number of potential problem facilities falling below this threshold.

The figure also shows that some stations have very low index scores (i.e., in the 0-20 range). These stations are clearly the worst performers in terms of offline testing and are therefore the ones that should be first targeted for follow-up investigation to these trigger results.

Figure 2 presents results for a dynamometer untestable trigger. The trigger tracks the rate of occurrences in which an inspector indicates that a vehicle that is supposed to receive a loaded mode emissions test cannot be tested on a two-wheel drive (2WD) dynamometer due to factors such as the presence of full-time all-wheel drive or no-switchable traction control. This typically causes the vehicle to receive a less stringent idle or two-speed idle (TSI) test.

As with the offline trigger, most stations have relatively high index scores, meaning they have performed idle/TSI tests in lieu of the loaded mode test on a relatively small fraction of vehicles. It also shows that the dynamometer untestable index ratings slowly tail-off along a fairly even curve below the 80-100 range. A statistical threshold such as 3 sigma below the mean can be selected as a starting point to identify those stations that have most likely misidentified an excessive number of vehicles as untestable, thus triggering a further investigation of inspection stations receiving these ratings.*

While the above histograms illustrate the ability of graphical outputs to convey trigger results in an easily understandable manner, electronic and hard copy tabular outputs are also considered essential. Such outputs allow I/M program management staff to determine why certain test systems received low trigger index scores. For example, Table 1 contains a tabular output for one enhanced test system** that was produced for an *Average HC Emissions Score* trigger. Based on the low emissions scores recorded by this system, it was identified as a poor performer. One concern could be that such a station tested only new vehicles and therefore could have a valid reason for relatively low emissions scores. Differences in model year distributions are accounted for in how the trigger scores are calculated; however, reviewing model-year-specific results provides an additional assurance regarding this issue. As shown in the table, average HC emissions scores recorded by the test system are very low for all vehicle model years. Network-wide average emissions scores and analyzer/network ratios are also shown for reference.

This type of output, which shows that the test system is recording extremely low emissions scores relative to the network average, is valuable in providing additional insight into the causes of low trigger scores. Producing electronic data files containing such information allows a text editor, or spreadsheet or database application, to be used to easily display such tabular information for specific stations and test systems.

The volume of detailed results produced by a single triggers run can be tremendous. This makes it very cumbersome to print and review hard-copy information. Despite this, more limited summary information in hard-copy format may also be useful for some applications such as comparing the performance of a small subset of stations (e.g., the worst performers).

Another reporting element described previously is the use of control charts for equipment-related trigger results. Possible SPC/control chart presentations are discussed in Section 6.

*These results are based on the rate of inspector entries of "dynamometer untestable" for vehicles shown in the EPA I/M Lookup Table as 2WD testable; i.e., in which the inspector overrides the test system's selection of a loaded mode test. However, many programs ignore the 2WD testable information contained in the Lookup Table. There is no "override" frequency that can be used as a trigger. Dynamometer untestable entries in these programs would therefore include both legitimate (e.g., for AWD vehicles) and improper testing. Thus, unless this trigger can somehow be designed to account for expected high dynamometer untestability rates at certain inspection stations (e.g., Subaru dealerships that routinely test AWD vehicles), the results must be carefully interpreted to avoid falsely targeting such stations for further investigation.

**Trigger results can be produced for individual test systems, stations, or inspectors. Stations may have more than one inspector and multiple test systems. Station-based results are shown in the two figures, while Table 1 shows results for an individual test system.

Table 1
Sample HC Emissions Score Trigger Results

Model Year	Number of Vehicles Tested	Test System Average HC Emissions (ppm)	Network Average HC Emissions (ppm)	Ratio of Test System to Network Average Emissions
1981	2	4.5	184.9	0.02
1982	7	9.6	133.0	0.07
1983	2	9.5	137.0	0.07
1984	11	6.6	111.9	0.06
1985	8	5.4	126.5	0.04
1986	30	4.9	99.6	0.05
1987	8	8.6	101.9	0.08
1988	34	5.9	81.5	0.07
1989	11	6.3	80.3	0.08
1990	26	6.4	65.1	0.10
1991	2	3.5	64.3	0.05
1992	12	7.2	51.6	0.14
1993	3	2.7	52.3	0.05
1994	13	15.8	38.3	0.41
1995	2	6.5	31.6	0.21
1996	8	3.8	19.2	0.20
1998	4	8.8	10.4	0.84
Overall	183	6.8	49.2	0.14

Using the Results

There are a number of ways in which triggers or other statistical analysis results of I/M program data can be used. The most obvious include using (1) station/inspector results to better target overt and covert audits, enforcement actions, and other administrative procedures (e.g., mandatory inspector retraining); and (2) equipment results to identify test systems that need auditing, servicing, or other attention.

There are also less obvious uses. One specific use that could provide significant benefits would be to link the trigger results to the VID messaging capabilities inherent in most if not all VID-based programs. Depending on the exact features of this functionality, the VID can be programmed to send electronic messages to the I/M test systems either the next time they connect or at a preset date and time. These messages are then displayed to the inspector.

Using this messaging capability to notify the stations and inspectors of trigger results has considerable merit. For example, it could be used to alert them to degrading equipment performance and the need for service. Messages could also be sent to stations that are identified as poor performers as a proactive approach to improving performance rather than or in addition to targeting the stations for auditing and other administrative actions. The messages could be tailored so that they provide sufficient information to demonstrate that questionable activities have been detected without detailing exactly how these activities were detected.

Various messaging approaches such as the above are discussed in more detail in Section 6. Other potential uses of the analysis results that are also addressed in the section include the following:

1. Identifying and documenting a history of problems with test system components or individual brands of test systems;

2. Evaluating the impact of poor station performance on program benefits; and

3. Evaluating trends in both inspection and equipment performance over time.

Performance Feedback

It is critical that the efficacy of the selected triggers in correctly identifying poor performers be verified through some type of feedback loop. One method for accomplishing this is to track the success rate of these identifications relative to the results found through subsequent investigations of stations and inspectors (e.g., covert audits) or test systems (e.g., overt equipment audits). A second method would be to conduct a "post-mortem" review of the trigger results for test systems that are installed at stations found to be conducting improper tests through other means, to determine how the station could have been identified through a triggers analysis. This type of feedback is needed to fine-tune the triggers and maximize their effectiveness in identifying true problem performers. Performance feedback is discussed in more detail in Section 8.

###

3. STATION AND INSPECTOR TRIGGERS

This section contains a comprehensive list of recommended station/inspector triggers aimed at the full range of existing I/M test procedures, equipment and program designs. Some of these triggers are currently being run by certain states, while others are only in the conceptual stage. A description of each recommended trigger and the data required to run it is provided. Additional details regarding how each of the listed triggers should be calculated and performed are included in Section 7.

The recommended items are considered the most effective and feasible triggers. A separate list of additional triggers that some states are either currently running or considering is also provided, along with brief summaries of why each of these items is not included on the recommended list.

In addition to the recommended items, an individual state may want to include other triggers depending on its exact I/M program design and the specific data that are being collected on a regular basis. The triggers that are presented below will hopefully provide insight into additional performance parameters that could be tracked.

Future program changes may also lead to the need to add or modify various triggers. For example, the OBDII-related triggers that are included below are obvious recent additions to the list of typical triggers that have been used by some of the states. These triggers will take on added importance in tracking inspection performance as more states begin to implement OBDII checks on a pass/fail basis.

Recommended Triggers

Recommended triggers that are described below include the items listed below. (Asterisks indicate when model year weighting is recommended to ensure comparability between stations and inspectors.)

- Test abort rate
- Offline test rate
- Number of excessively short periods between failing emissions test and subsequent passing test for same vehicle
- Untestable on dynamometer rate (for loaded mode programs)
- Manual VIN entry rate
- Underhood visual check failure rate*
- Underhood functional check failure rate*

- Functional gas cap check failure rate*
- Functional vehicle pressure test failure rate*
- Emissions test failure rate (all tailpipe test types)*
- Average HC emissions score*
- Average CO emissions score*
- Average NO or NOx emissions score (for loaded mode programs)*
- Repeat emissions
- Unused sticker rate (for sticker programs)
- Sticker date override rate (for sticker programs)
- After-hours test volume
- VID data modification rate
- Safety-only test rate (emissions test is bypassed in program with safety testing)
- Non-gasoline vehicle entry rate (when this would result in the bypass of the emissions test)
- Non-loaded mode emissions test rate (loaded mode test is bypassed in favor of less stringent idle or TSI test)
- Frequency of passing tests after a previous failing test at another station
- Average exhaust flow (for mass-based transient test programs)
- Exhaust flow change between failing test and subsequent passing test (for mass-based transient test programs)
- Frequency of no vehicle record table (VRT) match or use of default records (for loaded mode programs)
- Drive trace violation rate (for loaded mode programs)
- Average dilution correction factor (for programs that have incorporated dilution correction into their test procedure)
- RPM bypass rate
- OBDII MIL key-on engine-off failure rate (for OBDII testing)*
- OBDII MIL connection failure rate (for OBDII testing)
- Overall OBDII failure rate (for OBDII testing)*
- VIN mismatch rate (for OBDII testing)
- PID count/PCM module ID mismatch rate (for OBDII testing)
- Lockout rate
- Waiver rate
- Diesel vehicle inspection-related triggers

The meaning of many of the above triggers is obvious; e.g., the failure rate triggers are all based on comparing the failure rate recorded by a specific test system to the network rate. Others (e.g., the *Repeat Emissions* triggers) are not as clear, but are fully described below and in Section 7 (which contains the associated calculation methodology). As noted for some triggers, they apply only to certain types of programs (e.g., loaded mode or OBDII). Section 5 contains a station/inspector triggers matrix that can be used by individual states to easily determine which triggers are applicable to their program.

The above listing shows there are a wide variety and number of triggers that can be performed. The most comprehensive existing triggers systems employ only about 20 triggers on an on-going basis, since states and VID contractors have found that it is fairly easy to identify under-performers using some of the more basic triggers (e.g., failure rates, offline test rates, etc.).[6] As noted previously, some states also rotate their active triggers to ensure that stations do not catch on to how they are being identified and as needed to focus on items of particular concern. For example, if I/M program managers become concerned that some stations or inspectors may be attempting to perform fraudulent after-hours tests, this trigger could be activated to track such occurrences.

One approach would be to initially implement a triggers system using only a limited number of fairly basic performance parameters. As experience is gained in running the initial triggers and it is determined how well they identify poor performers, the system can be expanded as desired. Flexibility will need to be incorporated into the initial system design to allow for such future expandability.

The term "station" is used for uniformity in the following descriptions, except in cases that apply specifically to inspectors. However, each of the following triggers could be run on either a station or inspector basis using the unique license identification (ID) numbers that have been assigned to each station and inspector, which are typically recorded in the vehicle test records and other applicable data files.

1. **Test Abort Rate** – This trigger tracks the frequency of tests that are aborted by the station. A high frequency of aborted tests indicates that the station either is having problems running proper tests or may be intentionally aborting failing tests in an attempt to fraudulently pass vehicles. For this trigger to be performed, test aborts must be recorded and transmitted to the VID; this does not occur in some programs.

2. **Offline Test Rate** – This trigger tracks the frequency of tests that are performed on an offline basis. A high frequency of offline tests indicates the station is attempting to evade proper test procedures by conducting tests in which some of the constraints have been removed (due to the tests' offline nature), or there is some sort of equipment problem that is causing the station to have trouble connecting to the VID. This latter issue is expected to be much less of a concern. Test records must indicate whether each test was performed on- or offline for this trigger. It is also critical that the network be properly designed to collect all offline test results as soon as possible after reconnection to the VID, and a standard process be in place to collect offline test results from stations that do not conduct any online tests for an extended period.

3. **Number of Excessively Short Periods between Failing Emissions Test and Subsequent Passing Test for Same Vehicle** – Some programs have tried to look at the length of time between tests or of the tests themselves to identify potential instances of clean piping or other test fraud. However, these factors can be influenced by many perfectly legitimate factors (e.g., inspector experience and competence, how a shop is set up to handle repairs and retests, etc.). This trigger instead focuses on the number of short periods between a failing and a subsequent passing test on the same vehicle, which is considered a better indicator of possible cheating. It is also not focused on short periods between failing initial tests and subsequent failing retests, since this could simply be caused by an inspector trying to do a very simple repair to see if he/she can get a failing vehicle to pass. Accurate starting and end test times

must be recorded in the test record for this trigger to be performed.[*] The triggers software must also be capable of identifying subsequent tests for the same vehicle. While conceptually simple, this can be computationally intensive (it requires reordering all test records captured by the VID in chronological order).

4. **Untestable on Dynamometer Rate** – This trigger is applicable only to loaded mode programs in which vehicle chassis dynamometers are used. It tracks the frequency of tests in which a vehicle is recorded as being untestable on the dynamometer, which is a method often used to fraudulently avoid the loaded mode test. As noted regarding Figure 2 in Section 2, some programs record instances of inspector overrides of the test system's selection of a loaded mode test for vehicles shown in the EPA I/M Lookup Table as 2WD testable. If so, the rate of such overrides can be tracked. More typically, however, there is no comparison of whether vehicles are 2WD-testable (as listed in the vehicle lookup table that resides on the test system software) versus the type of tests actually conducted. The rate of dyno-untestable entries in these programs could still be tracked, but they would include both legitimate (e.g., for AWD vehicles) and improper testing. The trigger would therefore need to be designed to account for expected high dynamometer untestability rates at certain inspection stations (e.g., Subaru dealerships that routinely test AWD vehicles) or the results must be carefully interpreted to avoid falsely targeting such stations for further investigation.

5. **Manual VIN Entry Rate** – Most test systems include scanners designed to read bar-coded vehicle identification numbers (VINs), to minimize data entry errors that often occur with manual VIN entries. Test software typically prompts inspectors to scan the bar-coded VIN, but then allows them to manually enter the VIN if it cannot be scanned. A high frequency of manual VIN entries may indicate an attempt to get around the required inspection protocol and thereby falsely pass vehicles. To ensure that bar-coded VINs are available on all applicable test vehicles, results for this trigger should be based on test data for 1990 and newer vehicles only. The method of VIN entry must be included in the test record for this trigger. Excessive manual VIN entry rates can also result from equipment problems (e.g., a non-functioning scanner), a factor that should be considered in evaluating the results of this trigger.

6 **Underhood Visual Check Failure Rate** – A low failure rate for any visual checks performed in the program (most typically as part of TSI tests) may indicate the station is performing inadequate visual inspections or is falsely passing vehicles. Older models with higher mileage accumulation rates typically exhibit higher visual failure rates due to emission control system deterioration and/or tampering by vehicle owners. This trigger (as well as others) therefore need to be model year weighted to

[*] Although appearing intuitively simple, actual experience has shown this to be somewhat more complicated. For example, one state recently identified some tests in which the start times were <u>later</u> than the end times. This was caused by clock problems in certain test systems, combined with a network design in which the system clock is reset to the VID clock after an electronic connection is made to the database. Prior to the connection, an incorrect, slow start time was recorded for the test. The clock was then reset upon connection to the VID and prior to the recording of the test end time, which in some cases led to a corrected end time that was earlier than the incorrect start time.

eliminate possible bias due to differences in demographic factors among stations. The only data required for the trigger are overall visual inspection pass/fail results.

7. **Underhood Functional Check Failure Rate** – If any underhood functional checks (e.g., of the air injection or exhaust gas recirculation [EGR] systems) are performed in the program, a low failure rate may indicate that the station is performing inadequate inspections or is falsely passing vehicles.* This trigger is very similar to the visual check failure rate trigger. It needs to be model year weighted to eliminate possible bias due to differences in demographic factors among inspection stations, and the only required data are overall pass/fail results for the functional inspection.

8. **Functional Gas Cap Check Failure Rate** – A low functional gas cap failure rate may indicate fraudulent tests (e.g., use of a properly functioning cap or the calibration standard in place of the actual vehicle gas cap). This trigger needs to be model year weighted. The only required data are overall pass/fail results for the gas cap check.

9. **Functional Vehicle Evaporative System Pressure Test Failure Rate** – A low functional vehicle pressure test failure rate may indicate fraudulent tests (e.g., use of the calibration standard in place of the actual test vehicle). This trigger needs to be model year weighted. The only required data are overall pass/fail results for the vehicle pressure test.

10. **Emissions Test Failure Rate** – A low failure rate may indicate that clean piping (i.e., in which a clean vehicle is tested in place of motorists' vehicles) or other fraudulent activity is occurring. This trigger needs to be model year weighted.** The only required data are overall pass/fail results for the emissions test.

11. **Average HC Emissions Score** – A low average initial emissions score, when adjusted for possible station-specific biases, is considered a good way to identify possible stations that are performing clean piping. Because older vehicles have higher emissions on average, this trigger needs to model year weighted. HC emissions scores for all initial tests are needed to compute the trigger results.

12. **Average CO Emissions Score** – A low average initial emissions score, when adjusted for possible station-specific biases, is considered a good way to identify possible stations that are performing clean piping. Because older vehicles have higher emissions on average, this trigger needs to model year weighted. CO emissions scores for all initial tests are needed to compute the trigger results.

*The performance of inadequate underhood functional checks has previously been identified as a particular problem in some decentralized I/M programs (e.g., California's Smog Check program).

**Some stations (e.g., dealerships) will test cleaner vehicles on average and thus have lower emissions failure rates. Model year weighting is designed to minimize the effect of this type of bias on the trigger results. Nonetheless, states will need to consider such factors in interpreting the results from this and other triggers such as those involving average emissions scores.

13. **Average NO or NOx Emissions Score** – This trigger applies only to loaded mode programs, since NO/NOx is not measured in TSI or single mode (curb) idle testing. A low average initial emissions score, when adjusted for possible station-specific biases, is considered a good way to identify possible stations that are performing clean piping. Because older vehicles have higher emissions on average, this trigger needs to be model year weighted. NO or NOx emissions scores for all initial tests are needed to compute the trigger results.

14. **Repeat Emissions** – This trigger is aimed at identifying stations with an abnormally high number of similar emission readings, which is a probable indication that clean-piping is occurring. The trigger is designed to assess the degree of repeated emissions scores among all test results recorded by a specific test system and is considered particularly powerful at identifying suspected clean piping. For this reason, most of the states that have triggers systems are either already running or actively interested in this type of trigger.[6] However, the approaches adopted to date are relatively simplistic and do not appear to be the most effective methods for identifying this type of fraudulent performance. An alternative method, which was recently developed and copyrighted by Sierra, involves using statistical cluster analysis to identify similar sets of emissions scores and group them for study.[*] Cluster analysis works by organizing information about variables so that relatively homogenous groups, or "clusters," can be formed. To visualize how cluster analysis works, consider a two-dimensional scatter plot. Cluster analysis will attempt to identify a locus of points by "drawing" circles on the plot so as to fit the maximum number of points within each circle. On a three-dimensional plot, the circle becomes a sphere in order to fit data along all three dimensions. While it becomes increasingly difficult to visualize how this process works as the number of variables increases, cluster analysis can cluster items along many different dimensions (i.e., address many variables at the same time).

    Using this technique, it is possible to cluster HC, CO, NO (or NOx), and CO_2 emissions scores recorded during loaded-mode emissions tests. HC, CO, and CO_2 scores from idle or TSI emissions tests can also be clustered. Newer vehicles (e.g., the five most recent model years) should be excluded to minimize any potential bias in the results. These vehicles would be expected to have a predominance of low emission scores, thus leading to a likely increase in the frequency of clustered scores. By excluding newer models from the analysis, identified clusters are expected to provide a truer indication of questionable emissions behavior. Model year weighting, similar to that used in other triggers, could also potentially be used. However, doing so would significantly reduce the number of test results included in each cluster analysis set, thus compromising the statistical validity of the results. It is also believed that such weighting is not needed if newer vehicles are excluded.

[*] This is believed to be the first use of statistical clustering as an I/M trigger. States that wish to pursue the use of statistical clustering should be aware that it has been copyrighted by Sierra and they will need to pay a copyright fee if they wish to use the method. The size of the fee will depend on the particular application. Interested states can contact Sierra (at 916-444-6666) for more information.

Cluster analysis is more computational intensive than the other triggers recommended in this report. For this reason, SAS®[*] or another appropriate computer programming language should be used to perform the clustering process. This enables the user to define the "radius" of the circles in order to control how tightly grouped readings must be in order to fall into the same cluster. Different cluster radii can be used iteratively until reasonable results are achieved. Results from a cluster analysis completed recently by Sierra for one enhanced ASM program are described below to illustrate how the analysis is performed and the type of results that are produced.

Cluster analysis tends to produce evidence of inspection stations that are engaging in clean piping in a couple of different ways. The first and most powerful indicator of potential instances of clean piping is the number of larger clusters generated for each station. In the example analysis, a cluster was considered large if it contained six or more inspections, and represented at least 1% of an analyzer's total initial test volume.[**] A sizeable number of larger clusters may be an indication that a variety of vehicles were used for the clean piping at an analyzer.

Table 2 lists the five analyzers found to have the greatest number of large clusters in the example analysis. Based on the results recorded for all other analyzers, the maximum number of clusters found in the worst-performing analyzers was significant. For comparison, the table also contains the number of initial ASM tests, the number of those tests with failing results, and the initial ASM failure rate recorded for each of the analyzers during the analysis period. All of the failure rates shown in the table are well below average, with two analyzers showing less than a 0.5% ASM failure rate. These results are obviously suspect and clearly support the use of clustering as a powerful statistical tool to identify potential fraudulent performers.

The *Emissions Test Failure Rate* trigger would have also identified these egregious cases. However, there may be other factors (e.g., pre-inspection repairs, test vehicle mix, etc.) influencing station-specific results that would tend to obscure the results of a simple failure rate trigger and reduce its effectiveness in identifying likely cases of clean piping. The *Repeat Emissions* trigger is based on less ambiguous test results (i.e., similarities in actual emissions scores) and is therefore considered a much better approach to this issue.

[*] SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

[**] The example analysis was performed on an analyzer basis; however, it can also be conducted on a station basis if the program includes a number of high-volume stations with multiple analyzers.

Table 2
Five Analyzers with Greatest Number of Large Clusters

Number of Large Clusters	Initial ASM Tests	Initial ASM Failures	Initial Failure Rate
17	483	12	2.5%
14	811	10	1.2%
11	897	19	2.1%
10	389	1	0.3%
9	914	4	0.4%

Another indicator of potential clean piping is the size of the largest cluster relative to the analyzer test volume. This statistic may indicate instances in which a facility is using the same vehicle repeatedly for clean piping. Table 3 lists the five analyzers having the highest percent of total test volume in the largest single cluster. In this context, the term "largest" means the cluster containing the most inspections. The actual size (radius) is the same for all clusters.

Table 3
Five Analyzers with Highest Percent of Total Test Volume in Largest Single Cluster

Cluster Label	Number of Tests in Largest Cluster	Total Number of Tests	Percent of Tests in Largest Cluster
a8	8	25	32.0%
a8	41	153	26.8%
a5	48	193	24.9%
a1	11	65	16.9%
a4	9	57	15.8%

As shown in the table, almost one-third of one analyzer's initial test volume, and roughly one-quarter of two other analyzers' initial test volumes, was found to fit into a single cluster. These findings are considered a strong indication that the analyzers are likely being used for fraudulent clean piping. Using the same cluster size criteria, 94% of all stations have a maximum of 1% of tests in the largest cluster for pre-1993 models.

The "cluster label" shown in the table is a unique alphanumeric title assigned by the analysis software to each cluster found in the data, with the first letter identifying the type of test generating the result (i.e., "a" = ASM). The subsequent number is assigned in sequential order by the software to each cluster found for an individual analyzer. Thus, the two "a8" cluster labels shown in the table represent the eighth cluster found for each of these two analyzers. Test type identifiers of "t" (for Transient Test clusters), "i" (for Low or Curb Idle clusters), and "h" (for High or 2500 RPM Idle clusters) can also be used. Only ASM results were produced in the example analysis; however, the triggers software can be easily configured to analyze other test results as well.

As with all the triggers results, results from both of the above cluster indicators can and should be further scrutinized rather than simply assuming that a station acted improperly. One approach to this more in-depth investigation would be to review the individual test results that make up the clusters. These can be found in the data file produced by the analysis software, which contains detailed results organized by station ID number, analyzer ID number, technician ID number, cluster number, test date and time, and individual emissions results for each test included in a cluster.

Table 4 presents a sample of cluster results for the analyzer shown as the apparent worst performer in Table 2. Due to the size of a complete set of cluster results, only a portion of the results for the analyzer is shown.

As the table shows, there are multiple cluster results from this analyzer that appear quite suspicious. Clustered emission results near zero are not that uncommon, due to the existence of clean, primarily newer vehicles in the in-use fleet. 1993 and newer models were therefore excluded from the example analysis to minimize the occurrence of valid near-zero clusters. This means that a high number of valid near-zero clusters for a single analyzer is unlikely. In addition, large clusters with higher (but still passing) emissions results is even more statistically improbable. The existence of such clusters is thus considered even more indicative of potential fraud.

Using the results shown for cluster a57 as an example, it is statistically unlikely that a single analyzer would have been used to test six different vehicles with emissions scores in the relatively narrow ranges of 134-144 parts per million (ppm) HC and 111-208 ppm NO. Identifying this type of emissions pattern in multiple clusters recorded by the same analyzer clearly justifies further investigation.

Table 4
Analyzer-Specific Cluster Results, Example 1*

Station	Analyzer	Technician	VIN	Time	Date	HC	CO	NO	CO2	Cluster
				12:20:41	20000801	123	0.10	93	15.3	a5
				12:07:18	20000809	124	0.07	61	15.3	a5
				15:09:04	20000810	118	0.07	91	15.3	a5
				12:13:37	20000814	123	0.06	76	15.3	a5
				12:54:38	20000814	126	0.09	117	15.3	a5
				12:20:17	20000826	118	0.08	96	15.3	a5
				13:36:01	20001030	125	0.08	90	15.3	a5
				12:02:20	20001130	132	0.08	109	15.3	a5
				11:40:56	20000701	70	0.03	20	15.3	a53
				11:12:41	20000801	62	0.02	49	15.3	a53
				10:28:28	20000811	68	0.03	18	15.3	a53
				9:10:36	20000829	65	0.05	27	15.3	a53
				15:54:01	20000926	70	0.02	18	15.3	a53
				14:09:28	20001002	73	0.05	28	15.3	a53
				10:01:43	20001021	67	0.02	50	15.3	a53
				9:02:02	20001025	61	0.02	26	15.3	a53
				14:17:49	20000920	55	0.09	677	15.3	a54
				11:53:16	20001208	53	0.03	632	15.3	a54
				10:43:11	20000805	95	0.07	149	15.3	a56
				14:16:40	20000823	90	0.05	40	15.3	a56
				14:39:36	20000829	97	0.07	89	15.3	a56
				8:23:45	20000901	88	0.08	70	15.3	a56
				11:29:32	20000926	88	0.04	53	15.3	a56
				11:10:24	20000929	91	0.06	35	15.3	a56
				15:12:26	20001103	92	0.05	85	15.3	a56
				11:22:40	20001207	98	0.08	43	15.3	a56
				16:04:49	20001207	90	0.05	133	15.3	a56
				14:59:32	20000821	141	0.14	208	15.2	a57
				12:31:10	20000830	141	0.13	153	15.2	a57
				13:48:03	20001103	131	0.13	157	15.2	a57
				12:56:45	20001130	144	0.09	138	15.3	a57
				15:25:05	20001201	134	0.12	172	15.2	a57
				13:46:27	20001206	141	0.11	111	15.2	a57
				11:05:37	20000920	34	0.01	128	15.3	a59
				12:14:18	20001026	35	0.02	137	15.3	a59
				12:38:01	20001030	27	0.01	156	15.3	a59
				10:15:12	20001107	31	0.07	116	15.3	a59
				9:56:16	20001117	40	0.02	135	15.3	a59
				9:02:05	20001228	37	0.00	175	15.3	a59

* Station, analyzer and technician ID numbers, and VINs, have been removed due to confidentiality concerns.

One caution that needs to be emphasized in looking at the cluster results produced in the example analysis is that the results of this technique have not yet been validated. No attempt has been made to further investigate actual inspection performance at the stations using those analyzers identified as potential problems through clustering (i.e., to determine whether fraudulent inspections are actually being performed). This same caution applies to all of the recommended triggers when a state first begins a triggers program. The success of each trigger that is implemented in identifying actual under-performers (i.e., facilities in which inspections are either being conducted poorly or fraudulently) needs to be evaluated and "calibrated" through follow-up investigations, as part of the state's efforts to develop an optimized on-going triggers system. This issue is described in more detail in Section 8.

The manner in which the other triggers are structured to identify potential underperformers is conceptually easy to understand; i.e., it would be expected that analyzers with the lowest emissions failure rates would be highly indicative of potential problems. This is not nearly as clear for the cluster results since it may be possible that multiple test vehicles simply have very similar emissions scores.

This is not expected for two reasons: (1) large numbers of vehicles typically do not exhibit such clustered emissions; and (2) only a very small fraction of analyzers are showing this behavior. If it were a common phenomenon, more analyzers would be expected to produce similar results. Despite this, caution should be exercised in confirming the validity of this technique prior to adopting it as a viable method of identifying instances of fraudulent testing.

There are also several ways the cluster analysis can be adjusted to identify potential problem analyzers. This includes adjustments to the following cluster criteria:

1. The cluster radii;

2. The number of tests that must be included within the radii for a cluster to be considered large; and

3. The percentage of total test volume resident in a single cluster that is deemed to be excessive.

For example, the cluster radii can be defined more tightly to minimize the likelihood that the analysis is simply identifying vehicles with similar emissions. It is noted, however, that, if anything, the criteria assumed in the example analysis are considered conservative, since they identify a very small number of possible clean-piping analyzers. For example, only seven analyzers were found to have more than six large clusters (i.e., which contain five or more tests). These criteria may therefore actually need to be loosened to best identify all potential instances of clean piping. Some type of feedback mechanism between the statistical cluster results and actual enforcement results clearly needs to be established to calibrate this trigger. This could include both (1) follow-up investigations into facilities identified as potential clean-pipers based on various cluster criteria; and (2) after-the-fact analysis of test results from facilities that have been found through other means to be conducting clean-piping, to determine what cluster criteria would have properly identified the facility.

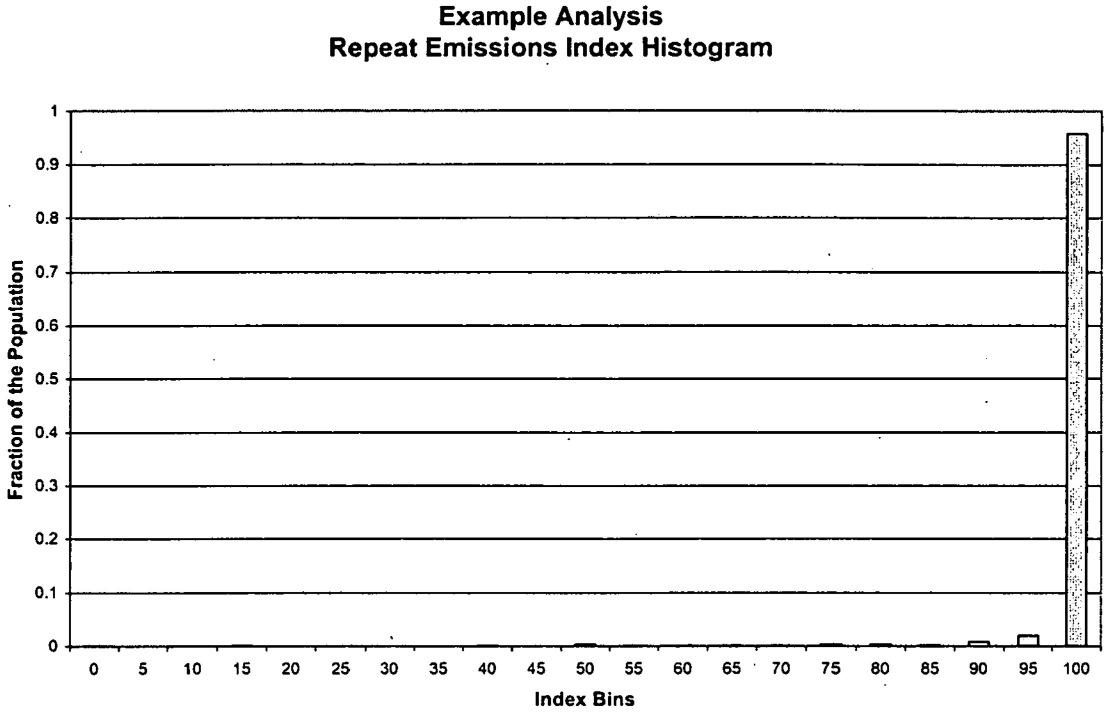
Another factor that could be considered is the time and date of clustered readings. In some cases, emission results will repeat within close chronological proximity of each other. Table 5 shows one such case for an analyzer that was also listed in Table 1. Of the five inspections included in cluster a13 for the analyzer, three of them occurred in the same afternoon on supposedly different vehicles. This finding is extremely suspect and is considered strong evidence of potential clean piping.

Table 5
Analyzer-Specific Cluster Results, Example 2

Time	Date	HC	CO	NO	CO_2	Cluster
15:33:52	11/02/2000	30	0.16	284	12.1	a13
16:21:57	11/02/2000	37	0.18	209	12.0	a13
18:17:39	11/02/2000	31	0.17	219	12.1	a13
11:00:28	11/04/2000	31	0.19	227	12.2	a13
14:44:24	12/01/2000	33	0.18	242	12.1	a13
13:10:28	12/23/2000	28	0.15	249	12.0	a13

Figure 3 shows the distribution of index numbers for this trigger that was produced in the example analysis using the above formula. The figure demonstrates that the vast majority of analyzers did not exhibit much, if any, clustering based on the criteria used to define the clusters in the example analysis. Those that had extremely low index scores were so few in number that they are barely noticeable on the figure. As noted previously, this appears to support the viability of this triggers technique, since abnormal test results have clearly been recorded for those analyzers with low index scores.

Figure 3



15. **Unused Sticker Rate** – This trigger applies only to programs involving sticker-based enforcement, in which unique stickers are issued to each passing vehicle and the disposition of all stickers is automatically tracked in the inspection software. Such software usually allows inspectors to indicate a particular sticker has been voided due to one of several possible reasons, with these results then recorded in a separate data file. Typical possible entries include V (voided), S (stolen), M (missing), and D (damaged), with inspectors allowed to enter any of these reasons. For the purposes of this trigger, all such entries are summed to compute total unused stickers. A high frequency of unused stickers indicates potential sticker fraud. To run this trigger, unused sticker results recorded in the separate data file need to be accessed by the analysis software. This can be problematic in some cases, since files containing unused sticker data may not be normally transmitted to the VID or the data in these files may be stored in less accessible database tables.

16. **Sticker Date Override Rate** – This trigger applies only to certain programs involving sticker issuance to each passing vehicle, in which (a) sticker expiration dates are automatically determined by the inspection software, (b) inspectors are allowed to manually override an automatically issued expiration date, <u>and</u> (c) instances of such overrides are recorded in either the regular test record or a separate sticker override data file. While obviously not applicable to many programs, the potential for this functionality to be misused in a fraudulent manner makes this trigger very important for relevant programs. To run the trigger, sticker override results must be accessible, which can be a problem if they are recorded in separate data files

that are not normally transmitted to the VID or the data in these files are stored in less accessible database tables.

17. **After-Hours Test Volume** – A high frequency of passing tests that occur after "normal" business hours may indicate fraudulent tests, since some programs have found that stations or individual inspectors are more likely to perform such tests in the off-hours.[6] This trigger is run by first defining the after-hours period, such as beginning at 7:00 p.m. and ending at 5:00 a.m.* While most of the listed triggers are based on the frequency (or rate) of occurrence, this trigger is instead based on the total number of passing tests that are started during the pre-defined after-hours period. This approach was selected by one state that is running this trigger to avoid the possibility that a high number of after-hours tests for a particular station could escape detection simply because the station performs a lot of inspections.[6] If after-hours test rates were instead tracked, this would tend to reduce the apparent significance of after-hours tests at such high-volume stations.

18. **VID Data Modification Rate** – Programs with online testing are designed to minimize the need for manual entry of vehicle identification and other information (e.g., to which type of emissions test a vehicle is subject). This is done by automatically sending such information from the VID to the test system at the start of each test in response to an initial vehicle identifier (e.g., VIN or vehicle registration number) entered by the inspector. The inspector is typically allowed to modify some or all of the VID-provided information, i.e., to correct inaccurate information. At least some programs track the frequency of changes to the data sent down by the VID.** A high rate of such changes is an indication that a station may be trying to cheat, e.g., by changing model year in order to get a vehicle exempted from the emissions test. These types of inspector changes would need to be recorded to run this trigger.

19. **Safety-Only Test Rate** – A number of vehicle inspection programs incorporate both emissions and safety inspections, with some including biennial emissions tests and annual safety tests. In such a program, a station may fraudulently attempt to bypass emissions inspection requirements by indicating a vehicle is only subject to safety testing (i.e., it is the "off-year" of the biennial emissions inspection cycle for the vehicle). A high frequency of safety-only tests, when model year weighted to adjust for station-specific biases, may be an indication of such fraudulent activity. The only required data are the type of test (emissions versus safety) that each vehicle received.

20. **Non-Gasoline Vehicle Entry Rate** – This trigger is aimed at identifying occurrences of fraudulent testing in which a fuel type that is not subject to emissions testing is intentionally entered for a test vehicle. Such fuel types could include Diesel (if these are not emissions tested), electric, and certain alternate fuels. A high frequency of

* The start and end times of the after-hours period should be user-configurable in the triggers software to allow for future adjustment as experience is gained with this trigger.

**One state records any occurrences in the test record of inspector changes to model year, GVWR, emission test selection, or fuel type values sent down from the VID.

non-gasoline tests, when model year weighted to adjust for station-specific biases, may be an indication of such fraudulent activity. The only required data are the fuel type that was recorded for each test vehicle.

21. **Non-Loaded Mode Emissions Test Rate** – In programs that include a loaded mode emissions test (e.g., IM240, ASM, etc.), stations may also attempt to evade this test in favor of a less stringent idle or TSI test. While similar to the untestable-on-dynamometer trigger (which is aimed at identifying the same behavior), this trigger is somewhat broader in that it looks just at the rate of non-loaded emissions tests rather than specifically at whether the inspector has indicated the vehicle cannot be tested on a dynamometer. For example, some programs include allowable low-mileage exemptions in which a vehicle that is not driven much (e.g., 5,000 miles per year) is allowed to receive an idle or TSI test in place of the normal loaded mode test. This type of legitimate exemption criterion could provide an avenue for unscrupulous stations to fraudulently bypass the loaded mode test. A high frequency of non-loaded emissions tests, when model year weighted* to adjust for station-specific biases, may therefore be an indication of fraudulent activity. The only required data are the type of emissions test that each test vehicle received.

22. **Frequency of Passing Tests after a Previous Failing Test at Another Station** – This trigger is aimed at identifying stations that may be offering to pass retests regardless of repairs, therefore attracting more retest business than expected. A high frequency of such passing retests may be an indication of fraudulent activity. Data needed to perform this trigger include accurate test times and overall test results. The triggers software must be capable of identifying subsequent tests for the same vehicle. As noted under a previous trigger, this can be computationally intensive since it requires reordering all test records captured by the VID in chronological order. It also requires establishing criteria for demarcating those tests included in each separate inspection cycle.

23. **Average Exhaust Flow** – Exhaust volume data are collected and recorded on a second-by-second basis in centralized and at least some decentralized transient emissions test programs.** Exhaust flow below expected levels for a test vehicle may indicate that the sample cone is not properly located. This could occur due to either unintentional improper cone placement (i.e., the inspector simply does a poor job of placing the cone) or an intentional, fraudulent action designed to not capture all the exhaust, thus resulting in lower mass emissions scores.***

*Most loaded-mode program designs include idle or TSI tests on older (e.g., pre-1981) model year vehicles. The trigger therefore needs to be designed to either look just at those model years that are subject to loaded mode testing or incorporate model year weighting to eliminate any bias in the results.

**Even if these data are collected in a particular program, they may not be transferred to the central database. Exhaust volume data for each transient test (an overall average flow for the test is acceptable) must be available at the VID for this trigger to be feasible.

***The centralized I/M contractors typically incorporate some type of fuel economy check or other software algorithm into their inspection lane software that is designed to identify occurrences of incorrect cone placement and abort such tests. Such an algorithm may not, however, be included in decentralized mass-

(continued...)

Because of the ease with which cheating or even unintentional improper testing could occur in this manner, it is critical that a method be implemented to prevent and/or detect poor placement of the sample cone. Incorporating a fuel economy or similar check in the test system software will serve as such a preventive measure. Under this type of approach, if measured fuel economy for a test exceeds a pre-defined maximum threshold for the vehicle, the test must be restarted.

This type of preventative approach may be relatively meaningless, however, depending on how the pre-defined maximum thresholds are established (i.e., they may be set high to avoid unnecessary retests but then prevent very few questionable tests). In addition, some programs may not include such functionality. Implementation of a method for detecting such occurrences through subsequent analysis is therefore considered a very important trigger for the transient test. Under this trigger, average exhaust flow rates would be compared after normalizing the results for vehicle test weight and engine displacement. Both of these parameters have a significant effect on exhaust flow. They are included in the I/M Lookup Table previously developed for EPA and are typically passed down to the test system in loaded mode programs.

24. **Exhaust Flow Change between Failing Test and Subsequent Passing Test** – Another effective exhaust flow-related trigger is to look at the change in recorded readings between a failing transient test and a subsequent passing test. Variation in exhaust flow from test to test should be less than 10%, barring major changes in vehicle condition. As a result, a significant decrease between a failing test and a subsequent passing test is a very good indication that the inspector conducted a fraudulent test in order to pass the vehicle. This trigger therefore looks at the frequency of excessive decreases in exhaust flow between such tests, with a 20% decrease initially used as the criterion to flag these occurrences. The average change in flow for such tests is also computed. While not used in the trigger computation, this result can be looked at on a stand-alone basis since average decreases of more than 10% are also be considered strong evidence of possible fraudulent inspections.

25. **Frequency of No Vehicle Record Table (VRT) Match or Use of Default Records** – This trigger applies to loaded mode programs that incorporate an electronic vehicle record table (VRT) containing dynamometer test parameters (e.g., vehicle test weight and horsepower settings) typically obtained from the EPA I/M Lookup Table. The proper test parameters contained in the VRT are automatically accessed by the test software based on the vehicle identification information obtained from the VID or entered by the inspector. If no match is found to a specific VRT record, the test software will instead use existing default test parameters for the test vehicle. These defaults may consist of either a single set of test parameters that vary by vehicle body style and the number of cylinders, or model-year-specific sets of test parameters that can also be obtained from the I/M Lookup Table.*

***(...continued)
based test systems.

*The latter approach is more accurate; however, it has not been incorporated into some programs.

Relatively low VRT match rates have been seen in several programs. Reasons for this include inconsistencies in make/model naming conventions among I/M Lookup Table entries, state vehicle registration records and pre-existing make/model lists incorporated into the test software (which have been copied from older software); bugs in VID and test software design and operation; and errors in the manual entry of vehicle identification information. The programs are continuing to address these problems with the intent being to maximize the VRT match rate to the extent possible in order to ensure that vehicles are being properly tested. This trigger is aimed at tracking and using abnormally high no-match rates to identify poor or fraudulent performance in entering correct vehicle identification information. It can also be used to identify vendor-specific problems in programs involving multiple equipment vendors. Data needed to perform the trigger include a recorded indication for each loaded mode test of how the dynamometer loading parameters were determined (e.g., input source, VRT record number, etc.).

26. **Drive Trace Violation Rate** – Loaded mode test software includes speed excursion and/or speed variation limits that may not be exceeded during the transient or steady-state (ASM) drive trace. If they are exceeded, the test is aborted and the inspector prompted to re-attempt to perform the test. Most if not all programs also record the number of drive trace violations that occur during test performance. This trigger is aimed at evaluating station performance in complying with the applicable drive trace limits. A high drive trace violation rate is an indication that inspectors are either having a hard time driving the trace or attempting to effect emissions scores by intentionally modifying how vehicles are being driven. Depending on the nature of the drive trace, older vehicles may have a harder time complying with specified speed excursion and variation limits. Model year weightings are therefore used to address this issue. Data needed to perform this trigger include information on the number of drive trace violations that occurred in attempting to perform each recorded test. This information may not be recorded in all loaded mode programs.

27. **Average Dilution Correction factor (DCF)** – Many programs now incorporate dilution correction into their steady-state (e.g., idle, TSI or ASM) test procedures. This involves the use of software algorithms that compute the amount of dilution occurring in tailpipe emissions from the raw CO and CO_2 readings, and automatically correct reported concentrations to a non-diluted basis. A higher-than-average DCF means excessive dilution is occurring and may be evidence of an attempt to falsely pass a vehicle by not inserting the exhaust probe all the way into the tailpipe, thus reducing the measured emissions concentrations.* However, care must be taken in performing this trigger to ensure that the results are not biased by (a) the presence of air injection systems (AIS) on certain vehicles or (b) differences in the average age of test vehicles (e.g., older vehicles are more likely to have higher DCFs due to a greater frequency of exhaust leaks and certain emissions-related defects). The first factor is

* Although possible with BAR90-grade and earlier analyzers, this will not work with BAR97-grade analyzers due to the incorporation of dilution correction. However, inspectors who do not know this and previously used this technique to conduct fraudulent idle/TSI tests may still try to cheat by not completely inserting the tailpipe probe on steady-state tests, including ASM tests.

addressed by identifying all AIS-equipped vehicles on the basis of recorded visual or functional check entries and excluding them from subsequent analysis.* The remaining records are then model year weighted to remove this source of bias. Data needed to perform this trigger include AIS identification information and DCF values for each test. Many programs may not record the AIS data.

28. **RPM Bypass Rate** – Engine RPM must be maintained within specified limits in most steady-state test programs, and is typically recorded in the test record. At least some programs allow the inspector to bypass the RPM measurement (e.g., if the RPM probe is not working), with the test record indicating if this occurred. This trigger would track the rate of such RPM bypasses. An abnormally high rate would indicate either equipment (RPM probe) problems, or poor or fraudulent performance on the part of inspectors. Data needed to perform this trigger include a recorded indication of any RPM bypasses that have occurred.

29. **OBDII MIL Key-On Engine-Off (KOEO) Failure Rate** – A low KOEO failure rate may indicate inadequate OBDII inspections (e.g., the check is not actually being performed in order to complete the test faster) or fraudulent results. This trigger is model year weighted (only 1996 and newer vehicles are included) to eliminate possible station-specific bias. Data needed to perform the trigger are KOEO pass/fail results for each test.

30. **OBDII MIL Connection Failure Rate** – A high non-connect rate may indicate (a) inadequate OBDII inspections (i.e., problems in locating the connector); (b) fraudulent inspections, in which an inspector indicates a non-connect simply to pass this check**; or (c) a workstation problem (e.g., with the connector cord). This trigger is model year weighted (only 1996 and newer vehicles are included) to eliminate possible station-specific bias. Data needed to perform the trigger are MIL connection pass/fail results for each test.

31. **Overall OBDII Failure Rate** – A low failure rate may indicate the equivalent of clean piping, in which a vehicle known to pass the OBDII inspection is tested in place of motorists' vehicles. This trigger is model year weighted (only 1996 and newer vehicles are included) to eliminate possible station-specific bias. Data needed to perform the trigger are overall OBDII pass/fail results for each test.

32. **VIN Mismatch Rate** – While not currently required in the OBDII datastream, VIN can be obtained via the OBDII data link for many vehicles by accessing the OEM enhanced diagnostic data. One scan tool manufacturer (EASE Diagnostics) has indicated that it can obtain VINs from Chrysler, GM, and most Ford models.[7] EASE is in the process of adding this functionality for Toyota models, which it believes to

* This presumes that the AIS-related entries in the test records are correct. While a small fraction of entries (most likely under 10%) may actually be wrong due to inspector entry errors, the small degree of error resulting from incorrect entries is considered acceptable.

**This would obviously only be the case if the test software passes or waives the vehicle (e.g., in the case of backup tailpipe testing) through the OBDII test if a connection cannot be made.

include roughly 70% of all OBDII-compliant vehicles sold in the U.S. However, EASE also indicated that other scan tool manufacturers may not have achieved this same level of VIN coverage. In addition, the California Air Resources Board (CARB) is currently updating its OBDII regulations, one element of which would require all vehicle manufacturers to embed the VIN in the non-enhanced data stream beginning with the 2004 or 2005 model year. EPA is also in the early stages of considering revisions to the federal OBDII regulations, one element of which may well be a similar VIN requirement aimed at all federally certified vehicles.

This trigger is aimed at comparing the VIN obtained from the VID or entered at the test system with that contained in the OBDII datastream, for as many vehicles as possible. An abnormally high rate of mismatches between the two VIN entries would be an indication of poor or fraudulent* performance. To perform this trigger, the test software must be configured to obtain the VIN via the OBDII data link and record this in the test record (in addition to recording the VIN that is obtained from the VID, or through bar-code scanning or manual entry).

33. **OBDII PID Count/PCM Module ID Mismatch Rate** – This trigger is similar to the VIN mismatch rate trigger. PID Count and PCM Module ID address are vehicle parameter information that can currently be obtained from the generic OBDII datastream. While the combination of these two parameters may not be absolutely unique for every vehicle make/model, they are sufficiently unique to identify likely instances of clean scanning.

    To perform this trigger, the test software must be configured to obtain these data via the OBDII link and record them in the test record. An electronic lookup table containing correct PID Count and PCM Module IDs for each applicable make/model combination would also need to be developed and included in triggers software. A high rate of mismatches between the data contained in this table and those contained in the test records would be an indication of likely clean scanning.**

    As described in more detail in Section 7, EASE is presently developing such a table. It may also be possible to obtain this information from the individual vehicle manufacturers; however, states are likely to find this to be a time-consuming and relatively difficult approach to this issue. A third approach that states could also pursue is to begin collecting this information on a regular basis as part of all OBDII tests and subsequently analyze the resulting data to populate the required lookup table. After the table has been created,*** the full trigger can be implemented.

*This includes scanning a known clean vehicle or an OBDII simulator (or "verification tester") in place of the actual test vehicle, a fraudulent practice that has been labeled as "clean scanning."

**The best known available OBDII verification tester is manufactured by EASE. This unit has intentionally been programmed with a PCM Module ID address of FD, an address that is not used in actual OBDII vehicles, to facilitate the detection of occurrences in which a unit is used to conduct fraudulent tests.

***Alternatively, these data could be added to the existing VRT.

34. **Lockout Rate** – Decentralized test system software normally includes a series of "lockout" criteria that automatically prevent the system from being used to conduct official I/M tests if one or more of the criteria are violated. Some test system lockouts are cleared by I/M program staff only after they verify that the underlying cause of the lockout has been corrected.* Egregious or repeated problems that trigger lockouts may lead to additional enforcement or other administrative actions against the station. While individual programs often utilize slightly different lockout criteria, they typically cover the general areas listed below. Some programs incorporate most or all of these, while others only include a more limited set of lockouts.

    - Test system security-related tampers that are designed to prevent unauthorized access to sensitive portions of the test system;
    - Equipment calibration failures or other errors (some of these are self-clearing; i.e., they are cleared automatically upon test system recalibration);
    - Expired station licenses;
    - No-VID contact limits that are designed to prevent stations from conducting offline tests for an unlimited period of time;
    - File corruption errors or an invalid software version number;
    - Lockouts set by state staff at the test system or from the VID if the station is found to be violating program procedures; and
    - Lockouts set by the management or VID contractor if the station has not paid required fees.

    This trigger tracks the rate of security and other station-initiated lockouts (i.e., those caused by an action of the station) that are set on the test system.** An abnormally high lockout rate is an indication that attempts are being made to circumvent some of the automated protections built into the test system. To run this trigger, lockout occurrences must be recorded and transmitted to the VID on a regular basis. However, this functionality is not built into the ET system software in many decentralized programs.

35. **Waiver Rate** – In recent years, most decentralized programs have moved to limit the ability to issue vehicle waivers to only the program management office or its

*Certain lockouts are self-clearing (as described below), and some programs also incorporate service-related lockouts that can be cleared after required equipment maintenance/repair by field service representatives (FSRs).

**Non-station-initiated lockouts (e.g., calibration failures and file corruption errors) are not included since they are not a good measure of station/inspector performance. Some lockout records contain the reason for each lockout and can therefore be used in conjunction with more refined triggers that track only tamper-related lockouts.

designated representatives. However, a few programs still allow individual inspection stations to issue certain types of repair cost or other waivers to vehicles that meet specified waiver issuance criteria. This trigger, which is relevant to these latter programs, is aimed at tracking the rates of such waiver issuance. An abnormally high waiver rate is an indication that a station may be using this functionality to bypass vehicle inspection and repair requirements. Data needed to run this trigger include the number of waivers issued (which is often recorded in a separate data file) and the number of tests performed by each station.

36. **Diesel Vehicle Inspection-Related Triggers** – States that include emissions inspections of Diesel vehicles in their programs may also choose to establish separate triggers aimed at the performance of these tests.* Many of the above triggers (e.g., *Offline Test Rate*, *Emissions Test Failure Rate*, *Safety-Only Test Rate*, etc.) are directly applicable to Diesel emissions tests. Possible additional Diesel-specific triggers include light-duty and heavy-duty average Diesel opacity scores, which would be similar to the average emissions score triggers described above. If both light- and heavy-duty Diesel vehicles are inspected in a program, their average opacity scores should not be combined but instead tracked with separate triggers due to the likely fundamental differences in test procedures, opacity pass/fail standards, and emissions performance.

Additional Triggers

Other triggers that are being used or at least considered by some states[6] are described below, along with brief explanations of why these triggers are not included on the recommended list provided above.

1. **High Failure Rate or Excessive Emissions Readings** – These triggers are the opposite of the failure rate and average emissions triggers included in the previous list of recommended triggers. They are focused on identifying abnormally high rather than excessively low failure rates and emissions readings. They are not recommended as standard triggers since they are focused more on consumer protection (i.e., to identify stations that may be falsely failing vehicles to increase repair-related revenues), than identifying cheating and maximizing program effectiveness. Either type could be run on an ad-hoc basis to identify such stations if a state considers this to be a particular problem in its program. A high failure rate trigger would appear to be more appropriate than an excessive emissions readings trigger for this purpose.

*Diesel emissions test results should not be combined with results from gasoline vehicle tests in performing the triggers described previously, since there will be fundamental differences in the two sets of results (e.g., in failure rates, etc.). Many programs may choose to simply omit Diesel test results from their QA/QC tracking system, since these tests are a very small subset of total inspection volumes. However, there are some programs that have separate licenses for Diesel vehicle test stations, which may want to include a separate set of Diesel-only triggers.

2. **Test Volume, Test Time, or Time between Tests** – This type of trigger is aimed at identifying stations that are performing an abnormally high volume of tests or very quick tests. These types of results are considered possible indications that a station is clean piping vehicles, not performing full inspections, or otherwise conducting fraudulent tests to maximize test revenues and improperly pass failing vehicles. However, as noted previously in discussing the *Frequency of Excessively Short Periods between Failing Emissions Test and Subsequent Passing Test for Same Vehicle* trigger, these factors can be influenced by many perfectly legitimate factors (e.g., inspector experience and competence, how a shop is set up to handle repairs and retests, etc.). These triggers are therefore considered less effective in identifying questionable performance.

3. **Invalid VIN Rate** – This trigger is aimed at identifying stations that are having difficulty entering valid VINs, i.e., they have a relatively poor match rate with VINs recorded in the state vehicle registration records and at the VID. This is another approach to tracking VIN entries that can be used to identify stations that are intentionally altering the inspection process in order to bypass or influence certain program requirements.* However, the concern with this trigger is that state vehicle registration databases may have a relatively high degree of VIN errors. As a result, using the VINs obtained from the registration database as the benchmark against which the station entries are judged may be problematic. For this reason, the *Manual VIN Entry Rate* trigger described previously is considered a better approach to the same issue.

4. **Safety-Related Failure Rates** – Programs that include safety inspections may also include one or more triggers aimed at identifying poor or fraudulent performance in conducting these tests. The most basic of these simply involves tracking the overall safety failure rate, with the only required data being pass/fail results for this part of the test. This trigger needs to be model year weighted since older vehicles typically have more safety-related defects. Programs may also choose to add additional triggers aimed at tracking the pass/fail rates of various subelements of the safety inspection (e.g., the brake check). Because of the large number of individual items typically checked in a safety inspection, it is likely that only a few of these (e.g., those that the state believes stations may not be performing) would be tracked. One such element of interest is the steering/suspension check.** However, an initial triggers analysis in one state found a very low rate of actual steering/suspension defects. This in turn made it very difficult to differentiate between stations that were improperly passing vehicles and those properly performing the check; i.e., the failure rate data did not contain a sufficient range to accurately identify poor performers from good stations. All safety inspection-related triggers were omitted from the list of

*If a non-matching VIN is entered, the VID does not transmit any vehicle identification information to the test system and the inspector is instead allowed to manually enter this information. Certain false information can be entered to affect the test process; e.g., a passenger car could be identified as a sport utility vehicle (SUV) to trigger the assignment of less stringent emissions standards.

**Proper steering/suspension system operation is critical to ensuring the vehicle can be safely operated; however, checking this is time consuming. A station may therefore skip or perform it in a cursory manner and then enter a pass into the test system.

recommended items due to this factor (which applies to many of the individual safety inspection elements) and because this report is focused on emissions inspection-related QA/QC procedures. States could add them, however, following the same approach as described in the report for the *Emissions Test Failure Rate* trigger.

5. **Rate of Excessive Exhaust Flows** – Rather than simply comparing relative average exhaust rates among stations as described in the *Average Exhaust Flow* trigger, an additional step that could be added to this trigger would be to compare average exhaust flows to minimum acceptable exhaust flow thresholds. This would require adding a lookup table to the triggers software and populating it with the applicable thresholds. Station-specific rates of exceedances of the threshold rather than average flow rates could then be used as the trigger criteria. This would be a more robust approach to identifying possible cone misplacement; however, the required lookup of threshold values would significantly increase the implementation and operation complexity of the trigger. For this reason, it has not been included on the list of recommended triggers.

###

4. EQUIPMENT TRIGGERS

This section is similar to the previous one, but is instead focused on equipment-related triggers that are designed to identify instances of excessive degradation in equipment performance. It includes a comprehensive list of triggers aimed at the full range of existing decentralized I/M test equipment.* A description of each recommended trigger and the data required to run it is provided. Additional details regarding how each of the listed triggers should be calculated and performed are included in Section 7.

An important point is that these triggers are run using data obtained from the electronic calibration data files recorded by the test systems and uploaded to the VID. However, recent acceptance testing efforts conducted in several programs[8] have found significant errors relative to the test system specifications developed for each program in how:

- Calibration data are collected and calculated, and
- Calibration files are formatted and populated.

This is likely a result of the fact that much less attention has been focused on the calibration files (relative to the test record files) in the past. However, to perform the equipment triggers, all applicable data fields must be properly formatted and populated with accurate data. Given the high number of calibration-related errors that have been found in multiple programs, it appears probable that they are common to most decentralized programs. Any existing program that decides to develop and implement a set of equipment triggers should recognize that a significant acceptance testing, debugging, and verification process will be necessary as part of this development effort to identify and resolve all such calibration-related problems.

While a significant effort may be required to implement a system of equipment triggers, the possible benefits justify this effort. As described in more detail both below and in Section 6, equipment trigger results can be used to maximize the effective use of available resources in performing overt equipment audits and test system maintenance/repairs. Potentially, equipment triggers can be used to reduce the frequency of required equipment audits, thus providing a significant savings in program management costs. While required by federal regulation,[9] few enhanced decentralized

*As noted previously, using triggers to track decentralized equipment performance and identify possible problem test systems or components that need to be serviced or replaced before they begin producing inaccurate test results is relatively new and not in widespread use programs are believed to be performing on-site equipment audits of every test system at least twice a year. By tracking test system performance through the use of triggers, these equipment audits can be targeted at questionable units as opposed to routinely auditing all test systems in the program. This may allow audit frequency to be reduced below federally mandated levels without raising concerns regarding a resulting degradation in overall network performance.

As detailed below, certain triggers also have the potential to substantially reduce the use of calibration gas, resulting in a significant cost savings to the inspection stations. Given both this possible benefit and the potential reduction in the frequency of routine equipment audits, it is strongly recommended that states seriously consider adding equipment triggers to their current QA/QC procedures.

Recommended Triggers

The following recommended triggers are described below.

- Analyzer calibration failures
- Leak check failures
- Low calibration gas drift (CO and NO)
- Analyzer response time (CO and NO)
- NO serial number (evaluates service life of NO cell)
- Calibrations per gas cylinder
- Dynamometer coast down calibration failures
- Changes in dynamometer parasitic values
- Gas cap tester calibration failures
- Vehicle pressure tester calibration failures
- Opacity meter calibration failures
- O_2 response time (for VMAS™ programs)
- Excessive differences in VMAS™ flow (during "hose-off" flow checks)

Each of these triggers is fully described below and in Section 7 (which contains the associated calculation methodology). As noted, the various triggers apply only to certain types of programs (e.g., loaded mode or transient/VMAS™). Section 5 contains an equipment triggers/test matrix that can be used to determine which of the triggers are applicable to a particular type of program.

The above listing shows there are a number of triggers that can be performed, particularly to assess analyzer performance. States may choose to initially run only a small subset of these triggers that are aimed at tracking equipment performance of particular concern (e.g., NO cell performance). As experience is gained in running the initial triggers and using the results to target equipment problems, the system can be expanded as desired. However, flexibility will need to be incorporated into the initial system design to allow for such future expandability.

Each of the following triggers is run on a test system basis using the test system ID numbers that are typically recorded in the calibration and other applicable data files. The triggers are listed separately by the type of equipment component to which they apply; however, they numbered sequentially for ease of reference.

Analyzer-Related

1. **Analyzer Calibration Failures** – An usually high rate of analyzer calibration failures is an indication that the analyzer is going out of calibration in less than the three days typically allowed between successful calibrations in decentralized test systems.* This could result in vehicles being tested with equipment that is out of calibration. Data needed to perform this trigger are the analyzer pass/fail results recorded in the calibration data files.

   Some equipment manufacturers have incorporated a 2-point calibration method into the decentralized test system software used in many programs.[10,11] This approach will tend to mask any bench calibration problems. It involves calibrating the bench at both the low and high cal points, rather than calibrating it at the high point and checking for proper readings at the low point. (The latter method is required by the BAR97 specifications and recommended in the EPA ASM guidance[12]). This means that the calibration curve of the bench is never independently checked during the standard 3-day calibrations, but simply reset based on the low and high calibration gas values. The result is that these test systems can never fail their 2-point calibrations, which will bias the results of this trigger. Sierra has recommended to its state clients that the calibration procedure follow EPA guidance in which an independent check of the bench calibration, using the low gas, does occur. This would remove the bias from this trigger.

2. **Leak Check Failures** – An usually high leak check failure rate could indicate the analyzer has a leaky probe that has not been repaired, but has instead only been patched to pass the leak check. Unless dilution correction is incorporated into the test software, this may lead to sample dilution and low emissions scores. In addition, a leaky probe may cause gas audit failures. (Audit gases are commonly introduced to the analyzer through the probe). Data needed to perform this trigger are the leak check pass/fail results recorded in the calibration data files.

3. **Low Calibration Gas Drift** – The larger the difference between the actual calibration gas concentration and the low "pre-cal" gas readings, the more likely it is that the analyzer is experiencing drift. Although it is typically required to be recalibrated to account for such drift once every three days, continued drift is an indication that the analyzer may be in need of service. This trigger therefore tracks such drift on a pollutant-specific basis (e.g., for HC, CO or NO). States could run this trigger for all three pollutants or choose to focus on a lesser number.

*The 3-day calibration frequency applies to the HC, CO and NO channels. The O_2 sensor has a separate, 7-day required calibration frequency.

Another possible use of the triggers results would be to determine if a 3-day period between calibrations is appropriate or if the interval could be extended without a negative impact on analyzer accuracy. If differences between the low calibration gas concentrations (or "bottle" values) and the actual readings are essentially zero, there is little need to calibrate the bench every three days. The current interval between calibrations was developed based on a study conducted for BAR in the late 1980s.[13] The quality of the analytical benches may have improved since then, and analysis of calibration data may indicate less frequent calibrations are now possible. Such an approach has the potential for cost savings to I/M stations, in both the cost of calibration gas and technician time required to perform the calibrations.*

Currently, there is no known source of data on the effect of extending calibration frequency on analyzer performance; i.e., whether calibrations can be extended beyond the current 3-day period (or how far they can be extended) without a significant negative impact on analyzer accuracy. For states interested in pursuing this issue, there appear to be two possible alternatives. The first would be to ask the equipment and analyzer manufacturers if they have conducted any related testing, and whether they can provide any data or other information that provides insight into this issue.

The second would be to conduct a controlled study on one or more in-use analyzers. This study would extend the calibration frequency and evaluate the effect on analyzer drift. Obviously, care would need to be taken to ensure that vehicles are not being tested using inaccurate equipment. The most feasible approach to ensuring this appears to be to perform a daily audit on the equipment each morning, in order to determine the amount of drift that is occurring over time. If any audit shows an excessive amount of drift, the analyzer would be recalibrated and the analysis period restarted. In order to perform the study, a state would need to work with one of the manufacturers to disable the required 3-day calibration frequency on the selected equipment. There are clearly several logistical issues that would need to be overcome in performing such a study; thus, obtaining and analyzing any available data that have already been collected related to this issue is obviously the preferred approach.

Another issue that must be addressed in considering an extension of the current 3-day required calibration frequency is whether it should be replaced with (1) simply another required frequency (e.g., once per week); or (2) some sort of feedback loop

*The number of gas calibrations that can be typically performed using a cylinder of calibration gas ranges from 20-40, depending on the brand of analyzer. This means 2½-5 cylinders are needed on an annual basis, assuming the analyzer regularly passes its periodic 3-day calibrations (which total roughly 100 required calibrations/year). Calibration gas currently costs $53-80 per low or high gas cylinder, plus $25 for a cylinder of zero air (this would not be required in test systems that incorporate zero air generators). Assuming a zero air and two (low and high) cal gas cylinders are required, the total cost would be $131-185 for all three cylinders. Combining this with the number of required cylinders shown above results in a total annual cost range of $328-925. If required calibration frequency can be extended to once per week, this would cut the number of required calibrations roughly in half, resulting in an annual cost savings per analyzer of $164-463. Additional savings would also occur due to reduced technician time in performing the calibrations. While not huge, the savings is not inconsequential, particularly when considered on a program-wide basis involving 1,000 or more stations.

that can be used to identify when analyzer performance is degrading to the point that more frequent calibrations are required, which would then trigger a return to more frequent calibrations. Under the latter approach, such a feedback mechanism could be implemented either at the test system level (i.e., it would be programmed into the software) or at the VID level. Analyzer drift would need to be tracked at one of these levels and used to determine when to trigger an analyzer calibration.

While there may be a fair degree of effort involved in figuring out how best to get the answers to the questions discussed above, this effort may be justified by the cost savings that the stations could realize if the calibration frequencies on properly operating analyzers can be reduced.

4. **Analyzer Response Time** – Most decentralized test system hardware and software have been developed to meet equipment specifications modeled after the BAR90 or BAR97 specifications, or EPA ASM guidance. Accordingly, their analyzer calibration procedures include response time checks of the CO, NO, and O_2 channels.* Reference (or benchmark) T_{90} values** must be stored in memory for each of the channels when the analyzer bench is first manufactured or installed in the test system (in the case of a bench replacement). Actual T_{90} values are then measured during each calibration and compared to the reference values. If the difference in these values exceeds 1 second for any of the channels,*** a message is displayed indicating that the analyzer should be serviced. If it exceeds 2 seconds for CO or NO (but not for O_2), the analyzer fails the gas calibration and is locked out from use. The software also includes similar T_{10}-related CO and NO response time checks (which track analyzer performance in measuring how long it takes to return to the T_{10} point when calibration gas flow is replaced with zero air).

   A slow response time could be an indication of the needed service due to a buildup of residue inside the analyzer. Response time data may also be used to predict when an analytical bench will fail or require service. While a loss in CO or NO response time in excess of 2 seconds will result in analyzer lockout, a lesser degradation in performance may also mean the analyzer is not capable of accurately measuring sharp increases and decreases (or "spikes") in emissions readings. This could be of particular concern to programs that are using BAR97-grade analyzers in combination with VMAS™ flow measurement equipment to conduct transient loaded mode emissions tests. This trigger is therefore aimed at tracking the average difference between actual and reference T_{90} values for CO and/or NO. Each pollutant would be

*HC response time is not tracked separately, since HC and CO are measured using the same analyzer bench in some test systems (i.e., this is true of the Horiba analytical benches that are used in SPX test systems). The Sensors analytical benches that are used by ESP, WorldWide, and Snap-On incorporate separate measurement tubes for each of the three gases (HC, CO and CO_2).

**This is the period of time between when the analyzer first begins to respond to the introduction of a calibration gas and when it reaches 90% of the final gas concentration reading.

***For reference, the BAR97 certification procedures include maximum allowable response times of 3.5 seconds for HC and CO, 4.5 seconds for NO, and 8 seconds for O_2.

tracked separately through separate triggers.* Yet another trigger could also track the average difference between actual and reference T_{90} values for the O2 sensor.** This latter trigger is discussed in more detail under the section on VMAS™-related triggers.

To run any of the response time triggers, both the reference and actual T_{90} values applicable to the pollutant(s) of interest must be recorded in the calibration file. The BAR97-grade analyzers used in programs other than California are believed to all contain the response time check functionality described above, since other states typically require these analyzers to be BAR97-certified. However, not all states include the reference and actual T_{90} values in their calibration file specifications. They are instead stored internally and are not transmitted to the VID. States that decide to implement one or more response time triggers will need to ensure that these data are being recorded in the calibration files.

5. **NO Serial Number** – The NO cell has a finite lifetime and therefore needs to be replaced regularly to ensure its performance continues to be adequate. Cell lifetime is typically 6-12 months, although it may be as long as 15 months for the newest generation of NO cell.[14] Standard maintenance practices typically call for replacing the cell every 12 months to ensure its continued performance. The NO cell serial number is usually recorded in the calibration record and thus values can be tracked to determine how long a particular serial number has been active. Trigger results can then be used to target needed audits or service of analyzers with NO cells that are approaching or exceeding their expected 12-month lifetime.

   To run this trigger, the NO cell serial number must be recorded in the calibration record. It also requires that the manufacturers' field service representatives (FSRs) be diligent in updating the serial number when they replace NO cells (i.e., recording of the same serial number for an extended period of time could either be due to non-replacement of the cell or because the FSR neglected to update the serial number when the cell was replaced). One state that attempted to implement this trigger encountered considerable difficulty related to this issue.[6] In fact, it recently decided to modify the trigger to pull the required serial number data from the FSR logs that are recorded electronically on their test systems due to the inability to get the FSRs to record these data in the calibration files. This type of modification is a good example of the type of adjustments that a state may need to make in implementing certain triggers to address particular software specifications or other issues unique to the program.

6. **Calibrations per Gas Cylinder** – An unusually high number of calibrations conducted per gas cylinder would indicate the analyzer is not being properly calibrated (i.e., an insufficient amount of gas is being used per calibration), or a station changed the gas cylinder but neglected to change the gas values from the old

* NO needs to be tracked separately in a loaded mode program to evaluate NO cell performance.

** While this could be tracked to evaluate the performance of the O_2 sensor, it is not considered very important <u>except</u> in programs involving VMAS™ flow measurement, since O_2 is not a pass/fail pollutant.

gas cylinders to those of the new cylinders.* Conversely, a low number of calibrations per gas cylinder changes could indicate there is a leak in the system. This trigger is aimed at tracking this by checking for changes in the entry of calibration gas cylinder serial numbers. Under this approach, the number of calibrations in which the same cylinder serial number is recorded for a particular test system can be compared to the overall average per serial number for all test systems.

Different brands of test systems can use significantly different amounts of gas per calibration. Therefore, in programs involving multiple brands of test systems, brand-specific differences in the number of calibrations per gas cylinder must be accounted for in performing this trigger to eliminate bias due to equipment brand. Calibrations per cylinder should also be tracked for the high and low calibration gas cylinders.

In addition to identifying possible problem test systems that appear to be using too little calibration gas, stations with analyzers that are identified as using too much gas (i.e., they are getting less than 75% of the average number of calibrations from any gas cylinder) should be contacted and told that the unit may have a calibration gas leak and they should check their gas line connections. This can be done by programming the VID to either (a) notify I/M program staff of all such occurrences, so that they can notify the station; or (b) automatically send such a message to the test system via the messaging capabilities programmed into most VIDs. Linking this type of automated messaging to the triggers results is discussed in more detail in Section 6.

Dynamometer-Related

7. **Dynamometer Coast Down Calibration Failures** – An usually high rate of coast down failures is an indication that the equipment is experiencing problems and may require service. This could result in vehicles being tested with equipment that is out of calibration. Data needed to perform this trigger are the coast down pass/fail results recorded in the calibration data files.

8. **Changes in Dynamometer Parasitic Values** – Parasitic losses are determined after all failing coast downs following the procedures included in the BAR97 specifications and EPA ASM guidance. A high rate of change in recorded values over sequential

*The latter cause would mean that all vehicles are being tested with mis-calibrated equipment, which could lead to false fails or false passes (depending on the direction of the improper calibration).

parasitic determinations is an indication of an equipment problem and/or an inconsistent parasitic loading being applied to test vehicles. This trigger tracks the occurrence of what are considered excessive changes in calculated parasitic losses, as recorded in the calibration file. It is recommended that the trigger be initially defined to consider a change of more than 10% between sequential parasitic determinations.*

Evaporative Test-Related

9. **Gas Cap Tester Calibration Failures** – An usually high rate of gas cap tester calibration failures is an indication that the equipment is experiencing problems and may require service. This could result in vehicles being tested with equipment that is out of calibration. Data needed to perform this trigger are the gas cap tester pass/fail results recorded in the calibration data files.

10. **Pressure Tester Calibration Failures** – An usually high rate of vehicle evaporative system pressure tester failures is an indication that the equipment is experiencing problems and may require service. This could result in vehicles being tested with equipment that is out of calibration. Data needed to perform this trigger are pressure tester pass/fail results recorded in the calibration data files.

Diesel Opacity-Related

11. **Opacity Meter Calibration Failures** – An usually high rate of opacity meter calibration failures is an indication that the equipment is experiencing problems and may require service. This could result in vehicles being tested with equipment that is out of calibration. Data needed to perform this trigger are the opacity meter pass/fail results recorded in the calibration data files.

VMAS™-Related**

12. **BAR97 Analyzer O_2 Response Time** – This trigger was previously mentioned under the *Analyzer Response Time* trigger. However, it is repeated here due to its importance related to ensuring adequate VMAS™ performance. In programs using these units, a mixture of the vehicle exhaust and dilution air is routed through the VMAS™. Total flow through the unit is measured and combined with a calculated

* While 10% is recommended as the initial value to be set for determining when the difference in parasitic values is considered excessive, this value should be user-configurable in case future adjustment is found to be needed. For example, subsequent feedback on this trigger could indicate that a large fraction of dynamometers are experiencing this rate of change, thus necessitating an adjustment in the threshold value.

** Neither of the following triggers are now being used in the three inspection programs that have incorporated VMAS™ flow measurement technology into their test procedures. However, Sierra has worked with one of the states and the VMAS™ manufacturer, Sensors, to develop these triggers, which are expected to be implemented in the future.

dilution factor to determine vehicle exhaust flow. The dilution factor is determined by comparing:

(a) Oxygen concentration in a raw exhaust sample that is extracted via a tailpipe probe and routed to the BAR97 analyzer, where the oxygen concentration is measured by the O_2 sensor in the analyzer; and

(b) Oxygen concentration in the dilute flow through the VMAS™ unit, which is measured by another O_2 sensor incorporated into the unit itself.

The accuracy of both O_2 measurements is critical to properly determining exhaust flow. Any exhaust flow measurement error will in turn result in inaccurate mass-based emissions measurements in the transient loaded mode tests that are performed using the VMAS™. As a result, tracking the performance of the BAR97 O_2 sensor becomes much more important than in steady-state tests (where oxygen measurements are primarily collected to be provided to repair mechanics for possible diagnostic purposes).

As noted previously, reference and actual T_{90} values for the O_2 must be recorded in the calibration file to perform this trigger, which is not the case with all programs using BAR97-grade analyzers. It is unknown whether all three programs that are currently using VMAS™ flow measurement and BAR97 analyzers to conduct transient testing are recording these values in the calibration file.

13. **Excessive Differences in VMAS™ Flow** – A reference "hose-off flow value"* is determined for each VMAS™ unit at the time of manufacture or when installed. By comparing this value to actual hose-off flow measurements that are performed during subsequent in-use calibrations by inspectors, it is possible to determine if there has been any degradation in equipment performance. (Sensors has recommended that these calibrations be performed weekly.[15]) A high rate of excessive differences between the reference hose-off flow value and values recorded during hose-off flow checks is an obvious indication that there is an equipment problem. It is recommended that the trigger be initially defined to consider a difference in the reference and actual hose-off flow values of more than 10% as excessive.**[15]

To perform this trigger, test system calibration functionality must include the required performance of periodic hose-off flow checks, and reference and actual hose-off flow

*This is the flow rate through the VMAS™ unit with the intake and exit hoses removed.

**Based on input provided by Sensors, 10% is recommended as the initial threshold to use in determining when the difference in measured versus reference flow values is considered excessive; however, it should be user-configurable. Program-specific flow data should then be collected and analyzed to determine if this threshold value should be adjusted.

values must be recorded in the calibration file. Although this is not currently being done in any of the programs that are currently using VMAS™ flow measurement and BAR97 analyzers to conduct transient testing, it is scheduled for future implementation in at least one of the programs.

###

5. TRIGGER MENU AND MATRICES

As a supplement to the detailed descriptions contained in Sections 3 and 4, this section provides a series of summaries of the recommended station/inspector and equipment triggers, which is designed to serve as a quick reference guide for users.

Trigger Menus

Tables 6 and 7 respectively, list all of the recommended station/inspector and equipment triggers. They include the following summary information for each trigger:

1. Its focus (i.e., the type of performance it is designed to evaluate).

2. Qualitative rating (low, medium, or high) of its effectiveness in identifying poor performers relative to other triggers with the same focus. A high rating indicates that the trigger is very effective in identifying poor performers relative to other triggers.

3. Qualitative rating (low, medium, or high) of the ease with which its results can be interpreted relative to the other triggers. A high rating indicates that the trigger results can be interpreted relatively easily.

4. Qualitative rating (low, medium or high) of how easily the trigger can be implemented. A high rating indicates that the trigger can be implemented relatively easily.

5. The data required to run the trigger.

6. Any other pertinent comments that should be considered in deciding whether to implement the trigger.

Table 6
Station/Inspector Triggers Menu

Trigger	Focus	How Effective*	How Easy to Use*	How Easy to Implement*	Needed Data	Other Comments
Abort rate	Fraud/poor performance	M	H	M	% test aborts	Aborts may not be recorded
Offline rate	Fraud	M	H	H	% offline tests	
Short periods between tests rate	Clean piping	H	M	M	Test times	
Dyno untestable	Fraudulent bypass of loaded test	H	H	H	% untestable entries	
Manual VIN	Fraud/poor performance	H	H	H	% manual VIN entries	
Visual failures	Fraud/poor performance	M	H	H	% visual failures	
Functional failures	Fraud/poor performance	M	H	H	% functional failures	
Gas Cap failures	Fraud	M	H	H	% gas cap failures	
Pressure failures	Fraud	M	H	H	% pressure failures	
Tailpipe failures	Fraud	M	H	H	% tailpipe failures	
Average HC	Clean piping	M	H	M	HC scores	
Average CO	Clean piping	M	H	M	CO scores	
Average NO	Clean piping	M	H	M	NO scores	
Repeat emissions	Clean piping	H	M	L	HC/CO/NO scores	
Unused stickers	Sticker fraud	H	H	H	% unused stickers	
Sticker overrides	Sticker fraud	H	H	H	% sticker date overrides	
After-hours tests	Test fraud	M	M	M	Test times	
VID data modifications	Test fraud	H	H	H	% changes to VID data	May not be recorded
Safety-only tests	Fraudulent bypass of emissions test	M	H	H	% safety-only tests	
Non-gasoline tests	Fraudulent bypass of emissions test	M	H	H	% non-gas vehicles	
Non-loaded tests	Fraudulent bypass of loaded test	M	H	H	% non-loaded emissions tests	

Table 6
Station/Inspector Triggers Menu

Trigger	Focus	How Effective*	How Easy to Use*	How Easy to Implement*	Needed Data	Other Comments
Passing test after failing test rate	Fraud	M	H	M	% subsequent passing tests after failing test	
Average exhaust flow	Exhaust cone misplacement	H	M	M	Average exhaust flow	
Exhaust flow change	Exhaust cone misplacement	H	M	M	Average exhaust flow	
No VRT match/ defaults	Vehicle entry errors/fraud	H	H	M	% VRT non-matches/use of defaults	
Drive trace violations	Drive trace fraud/poor performance	M	H	M	% drive trace violations	
Average DCF	Tailpipe probe misplacement	H	H	M	DCF values	
RPM bypass rate	Fraud/poor performance	H	H	H	% RPM bypasses	
OBDII KOEO failures	OBDII fraud	H	H	H	% KOEO failures	
OBDII connection failures	OBDII fraud	H	H	H	% OBDII connection failures	
Overall OBDII failures	OBDII fraud	M	H	H	% overall OBDII failures	
VIN mismatch rate	Clean scanning	H	H	L	Regular and OBDII VINs	Have to add to OBDII datastream
PID count/PCM module ID	Clean scanning	M	H	L	OBDII PID count/PCM module ID	Lookup table of reference values must be developed
Lockout rate	Fraud	M	H	H	# lockouts	May not be transmitted to VID
Waiver rate	Fraud	H	H	H	% waivers	
Opacity failures	Diesel fraud	M	H	H	% opacity failures	

* The entries in these columns reflect L(ow), M(edium) or H(igh) ratings.

Table 7
Equipment Triggers Menu

Trigger	Focus	How Effective*	How Easy to Use*	How Easy to Implement*	Needed Data	Other Comments
Analyzer cal failures	HC/CO/NO accuracy	M	H	H	% failures	
Leak check failures	HC/CO/NO accuracy	M	H	H	% failures	
Low CO cal gas drift	HC/CO accuracy	H	H	M	Gas readings	
Low NO cal gas drift	NO accuracy	H	H	M	Gas readings	
CO response time	HC/CO accuracy	H	H	M	T_{90} values	
NO response time	NO accuracy	H	H	M	T_{90} values	
NO serial number	NO accuracy	L	M	L	Serial numbers	FSRs must record data
Cals per high cylinder	Cal gas usage	H	M	L	Cylinder numbers	Normalize by analyzer brand
Cals per low cylinder	Cal gas usage	H	M	L	Cylinder numbers	Normalize by analyzer brand
Coast down failures	Dyno accuracy	H	H	H	% failures	
Excessive dyno parasitic changes	Dyno accuracy	H	H	M	Parasitic values	
Gas cap cal failures	Tester accuracy	H	H	H	% failures	
Pressure test cal failures	Tester accuracy	H	H	H	% failures	
O_2 response time	O_2 accuracy	H	H	M	T_{90} values	Direct effect on overall accuracy
Excessive VMAS™ flow changes	VMAS™ flow	H	H	L	VMAS flows	Flow checks not yet implemented
Opacity meter cal failures	Opacity accuracy	H	H	H	% failures	

* The entries in these columns reflect L(ow), M(edium), or H(igh) ratings.

Trigger/Test Matrices

Tables 8 and 9 have been developed as respective guides to what station/inspector and equipment triggers are generally applicable to each type of existing emissions tests. Whether a particular trigger can be used in an individual program will also depend on the data that are being collected by the test systems..

Table 8
Station/Inspector Triggers Applicable to Existing Emissions Tests

Trigger	Idle/ TSI	ASM	Transient/ CVS*	Transient/ VMAS™	Gas Cap	Vehicle Pressure	OBD II	Diesel Opacity**
Abort rate	●	●	●	●	●	●	●	●
Offline rate	●	●	●	●	●	●	●	●
Short periods between tests rate	●	●	●	●		●	●	●
Dyno untestable		●	●	●				○
Manual VIN	●	●	●	●	●	●	●	●
Visual failures	●	○	○	○	○			
Functional failures	●							
Gas cap failures					●			
Pressure failures						●		
Tailpipe failures	●	●	●	●				
Average HC	●	●	●	●				
Average CO	●	●	●	●				
Average NO		●	●	●				
Repeat emissions	●	●	●	●				
Unused stickers	●	●	●	●	●	●	●	●
Sticker overrides	●	●	●	●	●	●	●	●
After-hours tests	●	●	●	●	●	●	●	●
VID data modifications	●	●	●	●	●	●	●	●
Safety-only tests	●	●	●	●	●	●	●	●
Non-gasoline tests	●	●	●	●	●	●	●	
Non-loaded tests		●	●	●				○

Table 8
Station/Inspector Triggers Applicable to Existing Emissions Tests

Trigger	Idle/ TSI	ASM	Transient/ CVS*	Transient/ VMAS™	Gas Cap	Vehicle Pressure	OBD II	Diesel Opacity**
Passing test after failing test rate	●	●	●	●	●	●	●	●
Average exhaust flow			●	●				
Exhaust flow change			●	●				
No VRT match/ defaults		●	●	●				
Drive trace violations		●	●	●				○
Average DCF	●	●		●				
RPM bypass rate	●	●						○
OBDII KOEO failures							●	
OBDII connection failures							●	
Overall OBDII failures							●	
VIN mismatch rate							●	
PID count/PCM module ID							●	
Lockout rate	●	●	●	●	●	●	●	●
Waiver rate	●	●	●	●	●	●	●	●
Opacity failures								●

Note: The ● symbol means the trigger is applicable to the listed test. The O symbol means it may be, depending on the specific test procedures being used.

*This test type is not explicitly discussed in the report since it is currently being conducted only in centralized programs.

**This test type includes non-loaded snap idle tests as well as other, loaded mode tests that some states are either using or considering using to test primarily light-duty Diesel vehicles. The O entries in this column are applicable only to loaded mode Diesel tests, except for the entry for the RPM bypass rate trigger (i.e., RPM is measured as part of a snap idle test).

Table 9
Equipment Triggers Applicable to Existing Emissions Tests

	Idle/		Transient/	Transient/	Gas	Vehicle		Diesel
Trigger	TSI	ASM	CVS	VMAS™	Cap	Pressure	OBDII	Opacity
Analyzer cal failures	●	●	●	●				
Leak check failures	●	●		●				
Low CO cal gas drift	●	●	●	●				
Low NO cal gas drift		●	●	●				
CO response time	●	●	●	●				
NO response time		●	●	●				
NO serial number		●		●				
Cals per high cylinder	●	●	●	●				
Cals per low cylinder	●	●	●	●				
Coast down failures		●	●	●				
Excessive dyno parasitic changes		●	●	●				
Gas cap cal failures					●			
Pressure test cal failures						●		
O_2 response time*				●				
Excessive VMAS™ flow changes				●				
Opacity meter cal failures								●

*This trigger could also be run in Idle and TSI programs; however, it is not needed since accurate O_2 measurement in these programs is not critical to equipment performance.

###

6. REPORTING METHODS/POTENTIAL USES

Triggers software is a powerful QA/QC tool that has the potential to provide significant benefits, including improved identification of underperforming stations and inspectors, advanced diagnosis of impending test system problems, a possible reduction in the required level of station and/or equipment audits, and ultimately improved program performance at a lower oversight cost. To achieve this potential, it is essential that the triggers software be combined with effective reporting methods and an overall comprehensive QA/QC plan aimed at improving inspection and equipment performance. This section addresses these topics by building on the reporting issues described in the preceding sections. It also describes other potential uses of the trigger results.

A key issue is how to present meaningful and easily understood summaries of the large amount of information that is typically produced by a triggers system. The system needs to be designed to minimize the amount of time and effort that program management staff have to spend creating and reviewing this information, thus freeing them up for actual auditing, other enforcement activities, etc. In designing the reporting system, care must be taken to avoid creating a system that is a reporting and paperwork nightmare. It must instead be focused on producing easy-to-understand results that can be used by I/M management staff to improve program effectiveness.

Web-Based Outputs

Two of the states contacted by Sierra indicated that their trigger results are available via the web to any authorized state or contractor staff.[6] The website is security-protected (e.g., by password, etc.) to avoid access by unauthorized parties. Typical website outputs include detailed report listings in either PDF-formatted tabular listings or Excel comma-separated (*.csv) files suitable for further analysis. On at least one of the websites, the analyzer ID number acts as a hyperlink that takes the user to the data for that analyzer, thus providing easy access to the desired results.

This type of reporting functionality clearly facilitates state access to the triggers results. It is particularly well-suited to large programs, and those involving a program management or VID contractor, in which multiple parties (e.g., the state environmental agency, state DOT, state DMV, and applicable state contractors) with numerous staff need to access the results. One of the states with an existing web-based reporting system also provides authorized EPA staff with direct access to the site.[6]

Graphical Outputs

While the tabular listings described above are similar to what is being produced by other existing triggers systems, these listings have some shortcomings. In a system with multiple triggers, a large volume of trigger ratings are produced that may require considerable time to interpret and use. It is therefore recommended that states consider adding graphical outputs such as histograms to their triggers systems. As illustrated by Figures 1 and 2 in Section 2, such histograms are very helpful in determining:

1. If the trigger has been properly designed;

2. Whether the resulting index scores are being correctly calculated and show sufficient range to allow the poor performers to be easily identified; and

3. The number of inspection stations that appear to need further investigation.

The power of the graphical summaries in meeting the above objectives is ably demonstrated by the results shown in Figures 1 and 2. The first two items are considered of particular importance. Many of the recommended triggers have never been run; the accuracy of others may be seriously compromised if defects exist in how test or calibration results are being recorded in the specified data files. Without looking at such figures, it is difficult to tell by simply reviewing columns of numbers whether a particular trigger has been correctly designed and resulting scores are being accurately calculated.

Conversely, an improper trigger design or an inaccurate calculation is likely to be readily apparent when looking at the type of histogram shown in the two figures. It is absolutely essential that the triggers be properly designed and their calculations checked prior to the rollout of their use in identifying questionable performers. For this reason, it is highly recommended that graphical output capability be included in the triggers system. This capability can also be added relatively easily. As discussed in Section 7, the two figures were created by importing the requisite trigger index scores into an Excel spreadsheet. This capability, while not typically included in existing VID application software, could certainly be added or the histograms could easily be generated external to the VID.

Graphical outputs other than Excel-generated histograms can also be used to convey this information, using software with which the user is most familiar. The key is to show the distribution of trigger index scores in such a way as to allow the above three determinations to be made as easily as possible. Results can be graphed for each of the triggers listed in Sections 3 and 4 using whatever software and graphical format is selected for this purpose. The same approach should be used to display all trigger indices to allow easy comparison of the distribution of index scores among the different triggers.

Control Charts

As described previously, control charts have historically been used to track the degree of variation occurring in production processes over time. One key advantage they provide is their ability to visually display trends in performance over time in a manner that can be interpreted fairly easily. While control charts are not feasible for use with station and inspector trigger results for the reasons discussed in Section 2, they should be considered for use in reporting equipment trigger results. In fact, recommended use of control charts in tracking I/M equipment performance over time is incorporated into EPA's IM240 & Evap Technical Guidance. Although this has led at least some centralized contractors to use them to track equipment performance over time, they have not yet been applied to decentralized inspection networks. Given their ability to visually display trends in performance over time, use of selected equipment control charts is a recommended addition to the QA/QC tools that can be used by I/M program management to track and improve test system performance.

Using control charts to track trigger results on an individual basis for each test system in a decentralized inspection network will result in large numbers of charts. Sixteen different equipment triggers are recommended in Section 4.[*] A number of decentralized programs include more than 1,000 licensed stations and test systems at present. Based on an example network of 1,000 test systems, 16,000 control charts would be generated on a regular basis (e.g., monthly) if they were generated for each recommended trigger.

Not all control charts would be generated for each test system, since the SPC software can be programmed to "ignore" in-control charts. Despite this, taking the time to create, review, understand, and interpret the results from out-of-control charts still represents a significant task. It would clearly be counterproductive to generate so many charts that program staff spends the majority of its time creating and reviewing the control charts, rather than using the chart results to investigate those test systems that appear to have problems. For that reason, it is strongly recommended that the control charts be limited to a manageable number that are of clear value in identifying potential problems.

Notwithstanding this recommendation, a number of the equipment triggers are aimed at various test system hardware components whose performance is fairly independent from one another. It therefore appears worthwhile to track most of the equipment triggers, except for those that are duplicative in that they are aimed at monitoring the performance of the same equipment (i.e., the NO response time and NO serial number triggers both are focused on NO cell performance).

Coupling control charts with a subset of the recommended equipment triggers appears to be a very effective QA/QC management approach. By design, equipment performance should be homogeneous among test systems and over time. Control charts would therefore be an excellent means of tracking and identifying degradation in performance among individual test systems. The trigger calculation methodologies described in

[*] This includes running multiple versions of some triggers (e.g., response time) to track pollutant-specific test system performance.

Section 7 are also designed to minimize any inherent variability that does exist in the data being analyzed (e.g., due to performance differences among brands of test systems), thus further ensuring homogeneous data. In addition, this approach ties the two processes (triggers and control charting) into a single combined QA/QC tool that is focused on the same data, making it easier for program management to understand and interpret the results from each process.

Recommended Control Charts – It is recommended that a series of user-selectable control charts be used to graphically show historical network performance for a user-specified period. This would default to the most recent triggers analysis period unless the user selects a different reporting period using available query criteria or another suitable front-end application. Third-party software could then be used to create and print control charts for certain of the recommended triggers.

Recommended trigger variables are shown in Table 10 (no attributes are recommended for tracking). Twelve variables are shown, along with the equipment component being tracked by each variable. Note that two triggers each are included for tracking HC/CO bench and NO cell performance, and calibration gas usage.

Table 10
Recommended Equipment Performance Control Charts

Equipment Component Being Tracked	Trigger Number[1]	Trigger Description
HC/CO bench	3	Low calibration gas drift - HC
	4	HC response time
NO cell	3	Low calibration gas drift - NO
	4	NO response time
Calibration gas usage	6	Calibrations per low range cylinder
	6	Calibrations per high range cylinder
Dynamometer	8	Changes in dynamometer parasitic values
Gas cap tester	9	Gas cap tester calibration failures
Pressure tester	10	Pressure tester calibration failures
Opacity meter	11	Opacity meter calibration failures
VMAS™ O_2 sensor	12	O_2 response time
VMAS™ flow measurement	13	Excessive differences in VMAS™ flow

[1] The numbers shown in this column refer to the numbered equipment triggers contained in Section 4 of the report.

Individual programs should initially evaluate incorporating each of these triggers into the control charting process, before settling on a single trigger for each component. This would thus ultimately result in nine sets of control charts being created on a regular basis. However, only certain of the charts would apply to each program (e.g., few programs are using VMAS™ units and no decentralized programs are conducting vehicle pressure tests). Combining this factor with a system design in which the user normally only sees out-of-control charts (i.e., for test systems whose performance exceeds specified statistical limits) results in what is considered a reasonable level of information for program staff to review.

To keep the number of charts to a manageable level, the system should be designed to print only out-of-control charts on a routine basis. As with the tabular trigger ratings, authorized users could be prompted at the beginning of each month to enter a password to trigger the automatic printing of all out-of-control charts. In addition, authorized users could generate an ad hoc query for all or selected (e.g., by test system or station) additional charts for the previous month or other periods, based on an inputted reporting period.

Overall QA/QC Approach

An additional element that could be included in the reporting of QA/QC information is to link the trigger results to the VID messaging capabilities inherent in most if not all VID-based programs. Typically, the VID can be programmed to send electronic messages to the I/M test systems either the next time they connect or at a preset date and time. These messages are then displayed to the inspector.

Using this messaging capability to notify stations and inspectors of trigger results has considerable merit. For example, it could be used to alert them to degrading equipment performance and the need for service. Another specific equipment-related example that was mentioned previously would be to send a message to any test system found to be getting significantly less than the average number of calibrations from any gas cylinder, indicating that the unit may have a calibration gas leak and its gas line connections should be checked.

Messages could also be sent to stations that are identified as poor performers as a proactive approach to improving performance rather than or in addition to targeting the stations for auditing and other administrative actions. The messages could be tailored so that they provide sufficient information to demonstrate that questionable activities have been detected without detailing exactly how these activities were detected. Such messaging appears to have significant potential to affect changes in inspector and station behavior at a low cost to the program, particularly compared to the high costs associated with an extensive overt and covert audit program. The best approach to this issue is likely to be some graduated program in which messaging is used as an initial step in attempting to improve inspection performance, and auditing and other enforcement activities are pursued in egregious cases (e.g., involving strong evidence of clean piping) or if such messaging is unsuccessful in improving performance.

Station Counseling - This type of graduated method would directly address a view that was expressed by one of the states contacted by Sierra.[6] Although the state has a station/inspector triggers system in place, it currently makes limited use of the system due to concerns about taking an overly authoritarian or hard-line position with under-performing inspection stations. The state is instead involved at present in more of a coaching role aimed at improving station performance.

There are, however, some drawbacks with this latter approach. First, it can take substantial resources to administer such a counseling-based program, particularly in large programs that have a significant fraction of underperforming stations. Most states are not currently expending sufficient resources on QA/QC activities in their I/M programs,[8] relative to both the level of effort required by federal regulation and that needed to minimize loss in program benefits due to inspection performance and equipment problems. Given this past history, it is expected that these states would also be unable to provide sufficient resources for this type of counseling approach to be effective in reaching and affecting the behavior of all or even a majority of the underperforming stations. Instead, it is likely to be focused only on a relatively few stations, and while looking good from a public relations standpoint, have little real effect on improving the program.

Second, it is unlikely to be effective in eliminating intentional fraud in the program. The stations engaged in such activities are unlikely to begin performing proper inspections unless they believe they will be caught if they continue to cheat. A counseling approach therefore only works if a number of fraudulent performers have previously been caught and these results have been publicized so that the remaining stations know about the state's success in catching these offenders. Even if this has happened in the past, this fear of punishment will only tend to restrict future fraud until such time as a few stations attempt and are successful at cheating in a relatively blatant manner. For the "carrot" of such a counseling program to work, there must also be a "stick" behind it that makes stations consider carefully before they decide to commit fraud.

Recommended Graduated Approach - Given the above background, it is recommended that states consider implementing a QA/QC approach aimed at improving inspection performance that includes the following series of graduated steps:

- Step 1 – VID messaging;
- Step 2 – Station/inspector overt audits and counseling;
- Step 3 – Station/inspector covert audits; and
- Step 4 – Stringent enforcement actions.

Under this approach, some level of overt and covert audits would be performed on a random basis to function as a deterrent against attempted fraud. If stations know the state is looking over their shoulders, they are less likely to attempt to perform improper inspections. In addition to these random audits, Steps 1-4 would be implemented against stations identified by the triggers system as being underperformers relative to the other stations in the program.

Step 1 would be to send VID messages to stations found to be under-performing except in egregious cases such as those involving strong evidence of clean piping. (Such cases might be immediately advanced to Step 3 or 4.) Such VID messaging is described in more detail below. If such messages are not successful in improving the behavior of a particular station within a certain time frame, the station would be advanced to Step 2. This step would involve targeting the station for overt audits and/or counseling (depending on what type of policy a particular state wants to pursue against such underperformers).

Stations whose performance cannot be improved through Steps 1 and 2, or which are advanced immediately to Step 3, would be targeted if possible for covert auditing under Step 3.* If covert audits cannot be successfully performed against a particular station or are not successful in improving station performance, the station would be targeted for more stringent enforcement actions under Step 4. This could involve mandatory retraining of inspectors, license suspension or revocation, and other actions.

VID Messaging - There are a number of issues to be considered in deciding how best to use VID messaging as a QA/QC improvement tool. The first is whether to automate at least some messages. Under this approach, the VID could be programmed to automatically send preset messages to a test system if certain trigger results are found. (This would also require the triggers to be run on an automatic basis; e.g., monthly, on the VID itself.) An example of this type of messaging would be sending a message to a station that it check the test system for calibration gas leaks. Obviously, such an automated system would be suitable only for messaging that is 100% accurate. Continuing the above example, a message to check for gas leaks could be sent if the contents of the calibration files clearly show that the test system is using calibration gas at a much higher rate than other systems. It would also be acceptable to send a station a message that its failure rate is much lower than other stations if the test results show that to be true. However, sending a station a message that it has been identified as possibly clean piping vehicles simply because the trigger results show it as a potential clean piper would not be acceptable.

A second approach to messaging, if the triggers are run on the VID, would be to program the system to automatically run the triggers software and provide the I/M program staff with a list of stations identified as poor performers or with possible equipment problems, based on preset criteria. The staff could then decide which stations should receive messages or other follow-up activities.

If the triggers are run external to the VID, predefined messages could be developed and sent to stations based on information input to the messaging system. In the above

*Detailed guidance regarding overt and covert station/inspector audits is outside the focus of this work assignment and is therefore not included in the report. However, one caution in particular regarding covert audits is that many programs have found them to be relatively ineffective in collecting solid evidence of wrong-doing. Many stations cheat only when testing vehicles owned by known customers or friends; others that may be engaged in more widespread fraud have developed methods intentionally designed to avoid detection by covert audits. Covert audits therefore work better as a deterrent against fraudulent behavior by the vast majority of stations than in producing hard proof of fraud by the hard-core cheaters.

example of the calibration gas leak, a predefined message could be sent to all test systems that program staff enter into the system or select from a table of active test systems, based on the triggers results that are run external to the VID.

Another VID message-related issue is exactly what messages to send to stations identified as potential underperformers by the individual triggers. Given the large number of possible triggers and the likelihood that various states will want to adopt somewhat different approaches to dealing with the stations, specific recommended messages are not included in this report. In general, however, to effect a meaningful change in station behavior, messages should meet the following criteria:

1. Be relatively specific. If a message is overly generic, stations will not think they have actually been identified but rather the state is just trying to "jawbone" all the stations into improving their performance. The message should also read as if it is being sent just to that individual station, even if the same message is in fact transmitted to a number of stations.

2. Supply enough information to convince a station that the state knows exactly what the station has been doing. The message should convince the station that the state is really looking over its shoulder.

3. Contain completely accurate information. Any misinformation contained in a message targeted at a particular station will make it totally worthless, and may also compromise the integrity of the entire messaging approach (e.g., if stations begin to talk among themselves about inaccuracies in the directed messages). For this same reason, all messages should be totally predefined. This eliminates the possibility that program staff may make mistakes on ad-hoc messages and thereby compromise this element.

The final issue is regarding the criteria that are to be used in deciding when to send each pre-defined message. These criteria should also be pre-defined to avoid inaccurate decisions by program staff. They should also be designed so that there is no question whether a message applies to the station to which it is being sent.

Equipment Problems - The above approach addresses the issue of inspection performance, but does absolutely nothing to address equipment problems. Working with stations and inspectors to improve inspection performance will do little good if the test systems they are using produce inaccurate results. Therefore, any such counseling approach needs to be combined with some type of separate effort aimed at assuring that test systems are performing properly. To date, decentralized inspection programs have largely relied on the equipment certification process (supplemented by various degrees of pre-rollout acceptance testing) and the required periodic calibrations programmed into the test systems (e.g., 3-day gas and dynamometer calibrations) to supply this assurance. Although regularly scheduled equipment audits would also help in this regard, many programs are not performing a high frequency of such audits.[8]

While it is largely unpublicized outside the I/M industry, this approach has had only limited success in assuring proper equipment operation. Rigorous acceptance testing conducted by a few states has turned up substantial equipment problems.[8] This includes equipment being sold as BAR97-certified but that is actually noncompliant with the BAR97 specifications because the test systems have been modified from the configurations certified by BAR. While rigorous acceptance testing led to better test systems in these programs, similar corrections are unlikely to have occurred in other programs that incorporated a lesser level of acceptance testing prior to rollout.

Every decentralized program that has attempted to initiate equipment audits involving BAR-specified or EPA-recommended audit tolerances has also immediately run into very high levels of audit failures.[8] The typical reaction to these findings has been to reduce the number of audits being performed, decrease the stringency of the audit standards, and/or make the audit results advisory rather than forcing the equipment manufacturers to correct the problems.[8] These actions have been taken due to the acknowledgment that if a high fraction of test systems were to be locked out, it would cause a significant furor and loss of public credibility in the I/M programs. Certain states and the equipment manufacturers are also working on addressing these problems.[8] However, allowing the equipment problems to go uncorrected to at least some extent at present in decentralized programs across the U.S. means that some fraction of vehicles are currently being tested with questionable test systems. Moreover, many states may not be fully aware of corrections that BAR has insisted the manufacturers make in their BAR97 compliant systems. For example, while all enhanced decentralized test systems are typically required to be BAR97 certified, BAR recently released Addendum 8 to the BAR97 specifications. While most of the contents of Addenda 1-8 are software-related changes specific to California, the manufacturers may not have included any hardware performance improvements incorporated into the addenda into test systems used in other programs.

The final equipment-related concern is regarding the required periodic calibrations incorporated into the BAR97 and other states' test software. It is an accepted fact in the I/M industry that such periodic calibrations are an effective way to assure that decentralized test systems continue to operate within specification. It is assumed that if a test system cannot pass the required calibrations it is locked out until it can be serviced, thereby preventing inaccurate equipment from being used to test vehicles. However, there is growing evidence that this may not be the case. Information obtained during Sierra's triggers survey of the states[6] and through other means[8] reveals the following concerns in this regards:

1. States that have focused recently on reviewing calibration files (e.g., in acceptance testing and as a prelude to implementing equipment-related triggers) have found that (a) the files do not meet required specifications and (b) some required calibration functionality (e.g., certain response time checks) is even missing from the software.

2. Calibration records indicate that some failing test systems are simply calibrated over and over again until they pass calibration.

3. Test systems that routinely pass 3-day gas calibrations have been found to fail in-use audits at rates as high as 40% when audited using independent gases.

4. Required internal dynamometer feedback functionality (which is supposed to alert test systems to any problems that develop during testing) has been found to be missing in some test systems.

5. Newly developed hardware components (e.g., VMAS™ units) have been rolled out without what Sierra considers relatively essential calibration functionality.*

When they were first developed, the BAR97 equipment specifications pushed the envelope on the degree of performance expected out of decentralized test systems (i.e., they were relatively technology forcing). It is now clear that the equipment, manufacturers were able to obtain BAR97 certification only by using finely tuned, highly optimized test systems for verification testing. In retrospect, these systems do not appear to have accurately represented the level of performance achievable with a typical production unit, particularly after it has been exposed to several months or years of in-use service.

As a result, there are a number of in-use test systems in California that have problems passing audits. If this is the case in California, it should come as no surprise that other states with test systems that may not even be fully BAR97 compliant are seeing similar or worse problems. This raises the obvious question of whether these test systems are accurately inspecting vehicles.

Given this background, it is clear that states with decentralized programs should be devoting considerable resources to tracking and assuring adequate equipment performance to the extent possible. There are three primary ways this can be done. The first is by performing a rigorous acceptance testing process prior to equipment rollout and not allowing test systems to be rolled out until all problems identified in the testing have been satisfactorily resolved. The second is to implement a comprehensive equipment audit program that is designed to ensure all test systems are audited on a regular basis. However, even if such audits are performed at the level required by federal regulation (i.e., twice per year), this means it will be six months between the detailed performance assessments of each test system. In addition, few if any states are currently performing this level of audits due to the substantial resources that would be required.[8]

The third method for tracking equipment performance is to implement a series of equipment triggers that provide detailed data on the performance of each test system once per month. Combining this type of system with fewer but better targeted on-site audits will allow a state to truly assess how all test systems are performing in accurately testing vehicles. Equipment triggers are therefore considered an absolutely essential QA/QC element that should be implemented by all ET/VID-based decentralized programs.

*The two VMAS™-related triggers described in Section 4 incorporate elements (e.g., hose-off flow calibrations) that have been specifically developed to address this issue.

Other Potential Uses

In addition to the uses described above, the triggers results can also be used for other purposes. Some potential uses are summarized below.

Document Equipment Problems - The triggers results can be used to identify and document a history of problems with individual components or brands of test systems. This can be an effective way to assemble information for use in negotiating with equipment manufacturers and vendors regarding the need to upgrade or replace certain components. For example, if trigger results show a high frequency of calibration failures (e.g., due to a high incidence of multiple attempts at passing the required three-day calibrations) across the network or in one brand of test system relative to the others in the network, the information can be useful in getting the applicable equipment manufacturers to address the problem.

Evaluate Program Benefits - Triggers results can also be used to evaluate the impact of poor station performance on program benefits. By looking at both the fraction of test systems identified as poor performers and the fraction of tests conducted by such systems during a given period, it is possible to get a qualitative sense of the magnitude of the impact on overall program benefits. For example, if a small percentage of test systems are so identified and the systems were used to test an even smaller fraction of the inspection fleet, it can be concluded that the potential adverse effect on program benefits is fairly minor. Note that it should not be assumed that all stations identified as questionable performers are in fact performing improper testing, nor that all tests conducted at such stations during the analysis period were flawed. However, these results can be used to roughly estimate the impact on the program.

Evaluate Performance over Time - The triggers results can also be used to track and evaluate trends in both inspection and equipment performance over time. This subject has already been discussed in the context of using equipment performance-related control charts for reporting purposes. However, the underlying triggers data can also be used to track trends in both individual station and test system performance, as well as overall network performance. Note that care must be taken in how this is done. Trigger results typically evaluate the performance of individual test stations or test systems relative to the network average. Therefore, simply tracking average index scores or a similar metric over time can be deceptive since both individual test system and network performance will change.

In order to evaluate trends in inspection performance over time, it is necessary to fix the standard that is being used for benchmark comparisons. For example, a 0% offline rate can be used as the benchmark standard for the *Offline Test Rate* for the purpose of trends analysis. Similarly, 0% calibration failure rates can be used for benchmark purposes on all the various component-specific calibration failure rate triggers. Using what is defined as "perfect performance" as the benchmark ensures that this performance standard will be fixed in the future, thus allowing a realistic assessment of station and equipment improvement in performance against such a benchmark.

Each of the recommended trigger calculation methodologies includes the computation of the network average for the specified trigger index. These network results can be combined with the approach described above to evaluate the improvement in network performance over time.

###

7. ANALYSIS APPROACHES AND TOOLS

This section describes possible trigger-related analysis approaches, analytical issues, tools or specialized skills that may be needed and are available to implement a triggers program or individual triggers, system security, and other related issues. Specifically, information is provided on the following topics:

- System design and analysis issues;
- Recommended system design;
- Issues related to how to integrate the triggers system with the VID;
- Analytical tools and skills needed or available to perform the recommended trigger analyses;
- Trigger analysis approaches based on manual transcription of data; and
- Specific calculation methodologies for each trigger listed in Sections 3 and 4.

Each of these topics is discussed in detail below.

System Design and Analysis Issues

The basic format of triggers programs currently being used in I/M programs involves the use of a range of performance indicators, each of which is based on the analysis of I/M test records and other data (e.g., calibration records) that are automatically recorded by the workstations. For each performance indicator, individual trigger scores are computed and then translated into index numbers for every station in the program. Both mean and median values (by station and for the network as a whole) for the various triggers may be computed. States can choose to do this to see if there are significant differences in the results. Stations with low test volumes and low subgroup volumes (i.e., for certain model years) are normally excluded from the applicable analyses to ensure that statistically valid results are produced. The resulting index numbers are used to rank all the stations in order of performance.

This approach is used to avoid the use of relatively superficial statistics (e.g., pass rates, average test times, etc.) that could be significantly affected by the vehicle population being tested at particular stations. For example, a low failure rate is one type of trigger used by some states that compares average failure rates at each station weighted by the age of vehicles inspected by the station. This eliminates biases that might otherwise affect the index numbers assigned to stations such as dealerships (which would test mostly newer vehicles) or those located in low-income neighborhoods (which would be expected to test relatively older vehicles).

Assuming that stations are arranged in descending rank order from "best" to "worst," those stations that fall below a certain predetermined index threshold are considered potential under-performers. For each trigger, this threshold could be based on where an individual station ranks either (1) in relation to others (e.g., the bottom 50 worst-ranked stations are considered potential under-performers); or (2) relative to an index level of performance that has been determined to be unacceptable (e.g., all stations that rank below 50 out of a possible 100 points on a particular index could be considered under-performers). More sophisticated statistical means (e.g., identifying stations that lie more than 3 standard deviations [3 sigma] outside the mean) can also be used.

Each of these approaches has merit. For example, a preset number of "worst" performers could be used to identify and eliminate the most egregious cases of incompetent or fraudulent behavior. In particular, this approach may make sense if a state has limited resources to pursue enforcement actions and wants to maximize the effectiveness of these resources. It may also be the best approach to identifying truly fraudulent activities, as opposed to inspection problems that are related more to a lack of competence on the part of inspectors. Its disadvantage is that there may be more than 50 stations (or whatever number is chosen as the threshold) that are worthy of further investigation. Adopting a fairly arbitrary cut-off thus leads to the possibility that a state may miss identifying other stations and inspectors that should be investigated as well.

Using more sophisticated statistical methods would also identify outlier stations and could be structured to limit the stations that are identified for follow-up investigation to a reasonable number. States that are running triggers indicate, however, that such statistical approaches are not really needed; i.e., less complex approaches have been found to be successful in identifying problem stations.[6]

Focusing on all stations that fail to achieve what is considered an acceptable level of performance could be used to improve the performance of the overall network of certified stations. This approach is likely to identify more of a mixture of problems (i.e., related to both fraud and competency) than the above methods; however, it has two possible disadvantages. First, establishing a proper threshold for what is considered acceptable may be difficult, particularly when the triggers program is first implemented. Without experience in looking at the various metrics being used for the triggers and seeing how the range of values varies over time and across the network, it is hard to judge what is truly acceptable performance. Second, it could result in the identification of numerous stations as potential under-performers, leading to difficulty in deciding how to use available enforcement resources in investigating all of these shops. This problem is compounded if a large number of triggers (e.g., 20 or more) are used, each of which would result in separate rankings, and by the fact that separate station and inspector rankings could be developed.

Recommended System Design

Since most states have little or no experience in creating and interpreting trigger rankings, the system should be kept as simple as possible at the beginning. This is clearly the case with the existing triggers systems, which are relatively simplistic at present. States (or their VID contractors) that are running triggers are doing so almost entirely through the ad-hoc querying or pre-defined reporting applications built into the VID.[6] If any triggers analysis has been exported from the VID, it is still being carried out in a database environment. Trigger system outputs consist primarily of simple tabular listings/ratings, with some systems designed to produce data files that are subjected to further external analysis using either Microsoft Excel or Access.[6] The only trigger-related graphic results identified in this study were those developed by Sierra for two of its state clients.

The existing systems clearly do not include much of the analytical and reporting functionality described or recommended in this report. Although this may appear quite inconsistent, it is emphasized that the contents of the report reflect recommendations aimed at realizing the full potential of a combined VID/triggers system. This can be contrasted with the design and operation of the current systems, which largely reflect the infancy of triggers systems. While it certainly makes sense to initially begin with a very simple approach similar to that being used by some states at present, this report presents a road map that can be used by interested states to expand the system over time into a more complex (and useful) QA/QC tool. States that have begun to explore some of the elements (e.g., using equipment triggers) included in the report have already discovered the additional benefits they can provide.[6] Thus, while many of the elements described in this and the next section are well beyond all of the existing systems, they are considered parts of the "ultimate" triggers system that a state can work toward if interested.

To take advantage of the best elements of each of the approaches described above and to avoid I/M staff having to sort through and interpret multiple rankings, it is recommended that these approaches be combined into a single trigger system that contains the following elements:

1. Individual analysis results for each trigger would be computed for each station and compared to a specified standard to produce an index score based on a possible 0-100 point range;

2. A station's index scores for each trigger would be combined to produce an overall weighted score based on predetermined weightings;

3. Those stations not meeting the established minimum acceptable performance threshold for the overall weighted trigger score would be identified as potential under-performers; and

4. A prioritized list of stations would be generated and used to target future audit, enforcement or other administrative actions.

Note that the development of composite scores under item #2 above would not be done for the equipment triggers. Problems with individual hardware components such as the gas analysis benches, dynamometer, gas cap checker, etc., are expected to occur independently; thus, the development of a composite equipment rating system would be expected to produce fairly meaningless results.

Each of above elements and its impact on the design of the triggers system are discussed below.

Index Scores – A key issue in using available VID data to develop useful trigger results is determining the most appropriate standards of comparison (i.e., on both the low and high ends) to use for each trigger in developing an index score for that item. This may be fairly easy for some items; for others it is much more complex. For example, a trigger based on the rate of aborted tests can be easily compared to a "high end" standard of 0% aborts. A station with 0% aborts would therefore be assigned an index score of 100. However, even with this relatively simple example, there is a question of how the low end of the index range would be determined, i.e., what level of aborts would equal an index score of 0.

This exercise becomes even more complicated with other possible triggers. For instance, it is much harder to identify both the low and high standards of comparison for a trigger that uses the length of time between a vehicle failing a test and it subsequently being given a passing test as a way to identify potential instances of clean piping. Rather than attempting to use absolute limits in such a case, it may be better to base it on the performance of all the stations, i.e., through a comparison to mean or median network averages. This same approach could be used in setting the low end in the above example involving test abort rate.

One problem with this approach is that if a majority of stations are performing poorly on a particular indicator, only the poorest performers would be identified as possible problem stations. (However, this problem is self-correcting over time as long as the worst stations/mechanics are being removed from the program or rehabilitated.) Another concern is that if all stations are performing within a narrow range, this approach would assign substantial significance to very small changes in performance. For example, if all stations have average tailpipe emissions failure rates (corrected for differences in test fleet distribution) on initial inspections that range from 8% to 10%, a difference of 0.1% in the average failure rate would be equivalent to 5 points on a rating scale of 0-100.

Once the low and high ends of each trigger value are identified, it is relatively simple to translate the corresponding value for a particular station into an index score based on a possible 0-100 point range. In all cases, a high index score is considered "good" (i.e., it indicates proper performance) and a low index score is considered "bad" (i.e., it indicates questionable performance).

Trigger Weightings – For the triggers program to be useful and effective in identifying problem stations, its results must be easy to understand and use in targeting available audit and enforcement resources. Producing up to 20 lists of station rankings on a regular (e.g., monthly) basis will not meet this objective. To be truly useful, the number of resulting trigger lists should be kept to an absolute minimum. While this could be interpreted to mean that a single list should be produced, certain combinations of triggers are expected to be indicative of very different types of problems. For example, it may be possible to combine triggers that are aimed at identifying clean piping, but these should probably not be combined with triggers designed to track potential sticker fraud.*

As noted previously, station and inspector performance can be evaluated separately. At least initially, however, it makes the most sense to focus primarily on the station rankings. If a problem is identified with a particular station, subsequent investigation can determine whether a particular inspector is the cause of the problem. Users should therefore be able to generate separate station or inspector rankings for all of the station/inspector triggers. This functionality should be included as part of initial system development, since it is likely to be relatively easy and less costly to incorporate up-front rather than adding it later. If it is included, the system should be configured to allow users to decide whether to activate the inspector-specific rankings.

Clearly there is a wide range of possible data elements that can be analyzed to produce indicators of the level of station, inspector, and equipment performance. Depending on the nature of the underlying data, it is expected that the resulting indicators will have varying degrees of effectiveness in identifying problems. As a result, it is recommended that individual weights be assigned to each of the triggers used to produce the various ranking lists, resulting in an overall weighted result for each set of rankings. Detailed recommendations regarding how trigger weightings can be developed, including an example that illustrates how such weightings could be used in a program, are presented at the end of this section.

Minimum Thresholds – Until a state begins to run its desired triggers, the range of the expected index scores (on both an unweighted and a weighted basis) is unknown. This makes it very difficult to determine minimum acceptable thresholds for each weighted trigger score in advance. On the other hand, Figures 1 and 2 demonstrate that for at least some triggers it may be relatively easy to determine suitable thresholds based on initial triggers runs and the level of resources available to pursue follow-up actions aimed at questionable performers (i.e., stations, inspectors and test systems). A series of initial runs for each of the desired triggers should therefore be performed, and the results used to set the minimum thresholds. Other available information should also be considered in setting the thresholds. For example, if a significant number of service calls are being received by the state, management contractor, or equipment manufacturers, this is a fairly good indication of widespread equipment problems.

*Some states may instead choose to simply use single (unweighted) trigger rankings aimed at particular problems of concern or interest in their programs, or use some combination of single and combined rankings. The system should be designed to allow this type of flexibility.

Prioritized List – Once potential problem stations are identified by comparison to the applicable thresholds, each triggers list will be prioritized based on the weighted index scores. Prioritized lists will then be generated that identify the problem stations in reverse order from their scores. For example, the station that has the lowest inspection performance index number would be listed first, since it is considered the most likely to be having actual problems. The list will also display the index score for each of the listed stations, the network-wide mean and median scores, and the minimum acceptable score used as the threshold for identifying possible problem stations.

As experience is gained in using the trigger results, it is likely that future changes will be needed. Once the system is implemented, it is important to assess its performance in correctly identifying problem stations. If a high fraction of stations identified as having potential problems in fact turn out to be achieving adequate performance, the system will need to be "fine tuned" to better identify true poor performers. It is therefore essential that some sort of feedback process be established prior to implementation to assess the efficacy of the initial settings programmed into the system in identifying potential problem stations. This issue is discussed in more detail in Section 8.

In the event of the need for fine-tuning the triggers system, possible modifications could include some combination of changes to the data elements/triggers used to construct each weighted rating list, the weightings of the individual triggers, and the minimum acceptable thresholds. The system should be designed to accommodate these changes as easily as possible (e.g., by storing the "criteria" data in separate, easily updated files or database tables).

Another recommended design element is the capability to establish user-specific settings. Different users (e.g., staff from various state agencies and any management contractor) should have the ability to use different settings to optimize the system based on the particular issues of interest or under investigation. This element is recommended on the basis of the recognition that there may be some fundamental differences in how the various groups of users wish to use the resulting triggers information. For example, a state environmental agency may be interested in tracking very different inspection elements than a state transportation agency or a state DMV.

As noted previously, the capability to use either mean or median values for determining station scores should be included in the system design. Criteria also need to be specified to ensure that a sufficient number of data elements are being analyzed for each trigger to result in a statistically valid result. Rather than include more elaborate statistical criteria, it is recommended that a minimum of 30 records be required for each analysis element. Any stations that have fewer than 30 records in any specified analysis set would be treated as described below under the Calculation Methodologies section.

Because of the enforcement-related nature of the resulting information, security features are also needed to restrict access to the triggers system to authorized users only. It is expected that the specific security features programmed into the existing VIDs may vary considerably. In general, it is believed that extra care may be needed in restricting access to the triggers elements of these systems.

Federal I/M regulations[16] require that records audits be conducted on a monthly basis. It is therefore recommended that the triggers system be programmed to automatically produce the applicable lists of under-performers on the first day of each month, based on test and calibration data collected during the previous month. (States with existing systems typically generate triggers results on either a monthly or quarterly basis.[6]) To address the security concerns identified above, a message would be displayed at the beginning of each month and authorized users would be required to enter a password in order to trigger the automatic printing of each of these lists.* However, if the complexity and cost of this approach are considered excessive, the system could be designed to function only upon user input.

Separate lists would be printed if multiple types of separate performance evaluations (e.g., gasoline and Diesel vehicle inspection performance) are programmed into the triggers system. In addition, users should be able to create an ad hoc query (again based on the entry of a required password) for any of the programmed listings based on a reporting period input by the user. Authorized users should also have the ability to switch on routine monthly reporting of inspector listings, or conduct ad hoc queries to generate separate inspector listings. Users should also be able to generate trigger-specific listings if they wish to look at station performance on certain individual trigger items. Only trigger-specific listings would be generated for the equipment trigger items.

VID/Triggers System Integration

A number of possible approaches can be used to perform triggers and other statistical analyses of I/M program data designed to identify potential problem performers (i.e., stations, inspectors, and test systems). A key issue in deciding how the triggers system should be integrated with the VID is whether the trigger analyses should be conducted on or external to the VID. The advantage of running analyses directly on the VID is that all the data are already available. If the analyses are instead conducted external to the VID, a data set suitable for independent analysis (e.g., in ASCII format) would need to be created from the VID data and transferred to a separate computer for analysis.

However, VID-based trigger analyses have some inherent disadvantages as well. States that have attempted to conduct data intensive analyses on the VID have found this approach to be problematic for two primary reasons.[6] First, most VIDs are not set up with powerful data analysis tools, but instead rely mainly on pre-defined statistical reports or ad-hoc database queries to satisfy client data analysis needs. These tools may be sufficient for the standard reporting of program statistics and individual one-time investigations (e.g., of a vehicle's test history). They are not, however, designed to efficiently process the amount of data that is needed for triggers or other similar statistical analyses. Using the existing tools for this purpose can be cumbersome for the user unless such analyses have been clearly pre-defined and are readily available to users.

*This assumes that the triggers system is actively integrated into the VID, an issue discussed in more detail below. If the system is external to the VID, the display message could instead remind applicable state/contractor staff that it is time to complete the external analysis.

Database queries aimed at producing the desired triggers or other statistical analysis results can be developed using Structured Query Language (SQL). However, state staff contacted by Sierra have indicated that programming such SQL-based queries, particularly complex ones, can be very difficult.[6] (This issue is discussed in more detail below.) This raises the concern that it may be difficult to easily modify the tools to address special cases or evaluate slightly different elements than targeted by the pre-defined queries/reports.

A second problem is that the processing capacity needed to perform the statistical analysis typically causes a significant draw-down of system resources. This in turn causes substantial interference with the VID's main function – sending and receiving data to and from the inspection systems. In the extreme, it can lead to major response time problems and compromise the ability of the ET system and VID to adequately support on-line inspections. This problem in particular has led most if not all programs to run such analyses on the backup VID (also called the "mirror" or "shadow" VID) typically included in the system design.[6] (At least one program has even developed a second copy of the VID that is used for this type of data analysis.) Other programs are considering performing these analyses using some form of non-database application software.[6]

The above reasons provide a relatively strong argument for running the triggers external to the VID in some fashion. Software specifically designed for analyzing data may provide much more robustness and flexibility in dealing with the analytical issues inherent in triggers analysis. This is particularly true of the *Repeat Emissions* trigger, which is significantly more computationally intensive than the other triggers. According to one of the developers of the SQL-based VID used by Alaska, SQL is not well-suited for such analyses.

Based on Sierra's experience in using SAS software for a wide range of data analysis projects and the concerns expressed by state staff about their attempts to query the VID in their programs, it appears that structuring the triggers system (or more precisely the QA/QC statistical analysis system) to operate external to the VID will allow a wider range of analyses to be performed more efficiently. As noted above, the VID primarily serves as the data collection and storage element of the overall electronic transmission system. It is not designed to support intensive data analysis.

Any number of data analysis techniques and reports can be programmed into the VID. To date, analytical/reporting capabilities have generally lagged behind other VID implementation efforts, making it difficult to judge what analysis tools/functionality will ultimately be programmed into the VID and used in managing decentralized programs. There are many additional statistical analyses and reports that could be developed to aid states in assessing station and inspector performance. It is possible to use the VID's querying/reporting capabilities to address many of these; however, doing so may not be the most efficient approach.

To provide an example that illustrates this, one common way for stations to try to get around biennial emissions inspection requirements in a program that also involves annual safety tests is to simply change the vehicle model year from even to odd (or vice versa depending on the calendar year of inspection). This would have the effect of switching vehicles to safety-only tests. This type of fraudulent activity could be tracked using the *Safety-Only Test Rate* trigger described in Section 3, which could be constructed based on results obtained through pre-defined VID queries.

The problem with this approach is that there is no way of knowing for sure whether a particular station simply conducted a higher frequency of legitimate safety-only tests than the program average. While this example is a recommended inspection performance trigger, it illustrates the care that needs to be taken in interpreting the trigger results. Alternatively, a more precise approach to addressing this problem would be to compare the model year entered onto each test record with model year entries on corresponding vehicle records contained in the state motor vehicle registration database. This alternate approach would require additional programming and data processing resources that are likely to be difficult to attain in the VID environment. For example, data from the vehicle registration database would have to be somehow pulled into the VID environment. This could, however, be accomplished relatively efficiently using more robust statistical analysis software to analyze a combined stand-alone ASCII data set containing VID and registration vehicle records.

The above example demonstrates that it is possible to track QA/QC performance indicators using a variety of tools, including the triggers recommended previously and other analytical methods. Any of these approaches require the development of acceptable triggers or other statistical methods for identifying under-performing stations, inspectors and/or test systems, in order to guide state staff in deciding when to focus audit and other program management resources.

Given the above information, it appears preferable to run a triggers program external to the VID. States that are already running triggers on the VID or in another database environment should consider switching to a non-database-based analysis approach. It is also recognized, however, that some states may choose to continue to run triggers on the VID or a separate database since (1) this is they way they are currently doing it, and (2) the state and/or its VID contractor are most comfortable with and skilled in using this approach.

To the extent possible, therefore, recommendations contained in this report (e.g., regarding the use of VID messaging) are presented in a manner that allows their use regardless of where the analyses are performed. Information is also provided below on required and available software relevant to each of these approaches; i.e., either performing the analyses on the VID (or a backup version) or conducting them independently using a data set copied from the VID.. There is also a possible compromise approach in which available third-party SPC software and custom-built triggers software is "called" (accessed) by a VID application. This approach could allow states to continue to use the VID as the primary point of contact for such analyses and avoid the need to routinely generate the data sets needed for independent analyses. It could also take advantage of the analysis methods and equations incorporated into the custom triggers software (rather than having to construct a series of complex SQL queries), as well as the visual display capabilities (e.g., of control charts) offered by 3-party SPC software.

The concern with this approach is that it is likely to still cause problems in accessing needed system resources relative to those required to perform the VID's standard data transmission and storage functions. For this reason, the approach will only be feasible if linked to the back-up rather than the primary VID.

Analytical Tools and Skills

An important factor that needs to considered in deciding whether to implement a triggers system and exactly how such a system should be designed is the analytical tools and skills needed to implement and operate the system. Specific related issues include the following:

1. Any proprietary software or hardware that is needed to run an overall triggers program in general;
2. The availability and cost of any copyrighted or proprietary materials that are available to run the triggers program or individual triggers;* and
3. Specialized skills (e.g., programming in a certain language) needed to implement the triggers program or individual triggers.

These issues are relevant to the tools needed to (1) perform both the basic triggers and other statistical analyses; and (2) produce graphical and tabular reports of those results, including control charts. They are discussed below.

Overall Triggers System Software and Hardware - It is believed that almost all of the existing triggers programs are currently being run in database environments. Tools currently being used by the states for this purpose include Oracle Discoverer®, another SQL-based system and other database applications.[6] Each of these existing programs is presumed to be proprietary to the VID contractors involved in these programs.**

For states that wish to run the triggers system external to the VID, it is recommended that software specifically designed for analyzing data be used. As noted previously, Sierra uses SAS software for much of its data analysis activities. Since this software was originally developed for a mainframe environment, versions are available for almost all existing computer platforms. A state that decides to use SAS software for this (or any) purpose must purchase a copy of the software from the SAS Institute and pay a yearly license fee. There are a number of other types of data analysis software that can also be used for this purpose.

*To the best of Sierra's knowledge, no one is asserting patent rights to the basic concepts associated with a triggers program; however, states may be able to reduce the time and effort required to develop programs from scratch by relying on proprietary or copyrighted material already available.

**This is true in all programs for which Sierra has detailed knowledge.

Third-party SPC software that can be used to generate the recommended control charts is also available from numerous sources, including the SAS Institute. The available software packages encompass a wide range of functionality and complexity. An Internet or other search can be undertaken to identify suitable products involving both this and the data analysis software described above. It is recommended that states interested in adding this element to their QA/QC system consider purchasing an integrated product such as the SAS software that can be used to perform both data analysis and control charting.

Standard computer hardware can be used to run the triggers system and analysis software, which the exact hardware specifications dependent on the desired functionality and complexity of the software.

Available Copyrighted or Proprietary Materials - Most if not all of the existing VID contractors (e.g., MCI Worldcom, ProtectAir, and Testcom) have developed proprietary materials (i.e., software) that they are using to run triggers in which they are operating the VIDs.[6] The cost and availability of this software are unknown and it is not known whether the companies would be willing to sell or license its use absent a contract to also implement and operate the entire VID. If any such software is available, it would need to be customized for the program to which it would be applied. This is likely to be a fairly significant task given the typical program differences in test procedures, test and calibration data file structures, VID database structure, etc.

Sierra has developed copyrighted SAS software (which includes the *Repeat Emissions* trigger described previously) that is designed to run triggers on independent data sets provided by some enhanced inspection programs and produce both electronic tabular and graphical outputs (e.g., histograms). This software would need to be customized for application to another program. Such customized software is available through licensing or sale to interested states. The cost to license the software would depend on the number of desired triggers and their complexity and the degree to which additional reporting or other functionality is desired. Another option would be to contract with Sierra to develop and operate the custom software (i.e., analyze the data) on data sets provided by the state.

As mentioned in Section 3, the *OBDII PID Count/PCM Module ID Mismatch Rate* trigger would require the development of an electronic lookup table containing PID Count and PCM Module IDs for unique vehicle make/model combinations. EASE Diagnostics is working on a version of this table, but according to the latest information provided to Sierra it is not yet commercially available.[17] The likely licensing or purchase cost of this table is also unknown.

Needed Specialized Skills - A state that wishes to implement a triggers system faces three primary options:

1. It can hire a contractor to develop, implement, and operate the system, either as part of a VID implementation or on a stand-alone basis.

2. It can hire a contractor to assist in the development and implementation of the system, but operate the system itself.

3. It can develop, implement, and operate the system internally, starting either from scratch or by purchasing or leasing available software as described above.

In theory, if the state has a contractor develop the system, relatively few specialized internal programming skills should be needed regardless of whether the state or the contractor then operates the system. This should also be independent of whether the system is run in a database environment or using specialized data analysis software. If the system is well-designed, relatively unskilled users should be able to subsequently use it to run the specified triggers and produce reports, etc.

VID-Based Programming Skills - Actual experience may, however, be quite different from the above theory. States that have attempted to implement a VID-based triggers system have found that the implementation often lags well behind other, more important program objectives (e.g., proper integration of VID and test system operations).[8] The resulting triggers systems have been rolled out late and have to compete for contractor programming resources with other program elements. In addition, trigger system rollout is typically an iterative process involving the incorporation of various queries aimed at determining the best way to calculate effective trigger results. This means someone, either state or contractor staff, is often involved in constructing and reconstructing these queries. Even if contractor staff are responsible for this effort, states have found that they need some level of basic SQL query building skills to interact with contractor staff and review the results they produce.

Contractor-supplied SQL query building training for state staff is typically a part of a VID contract. However, these staff have found that there is a significant learning curve in picking up these programming skills to the extent needed to be able to construct the complex queries required for triggers and other statistical analyses of program data.[6] As noted above, SQL programming can be quite difficult. This programming involves a mixture of Boolean operations and English language phrases. Programmers typically use a query builder tool that includes query-by-example functionality to construct required SQL queries. Despite this, becoming proficient at SQL programming requires considerable experience in constructing such queries.

A state that chooses to develop its own VID-based triggers system would be faced with an even greater need to have in-house SQL programming expertise. Unless such internal programming skills exists, no state should attempt to implement this type of system on its own.

Data Analysis Software Programming Skills - A state that chooses to use SAS or another data analysis program will also need to have or acquire programming skills in this software if it wants to do anything other than run custom software developed by a contractor or obtained from another source (e.g., Sierra, as described above). Since triggers development is likely to be an iterative process, it is highly recommended that the state already have or acquire such programming skills rather than be dependent on a contractor or custom software provider to make any required modifications or enhancements to the triggers program.

Programming skills for SAS are considered somewhat easier to acquire than those for SQL; however, both can be extremely difficult for some people to learn. It is therefore recommended that any state that decides to implement a triggers system using either of these approaches have at least one staff member who is facile in the required programming language.

Motor Vehicle and Equipment-Related Engineering Knowledge - The trigger calculation methodologies recommended in this report are designed to make the results easily understandable; i.e., it should be easy to identify the worst performers without knowing much about the underlying data. However, using the trigger results as the first step in more in-depth investigations into the stations, inspectors, and test systems identified by the triggers system requires a range of knowledge covering the following elements:

1. The design of the I/M program, and all related policies and procedures. This element is fairly obvious yet its importance cannot be overstated. This would also include understanding the program's business rules and test procedures down to a very fine level of detail. For example, in reviewing the results of a *Frequency of No VRT Matches* trigger, program staff need to know exactly how the test software works to identify VRT matches, what inspector entries are possible in this part of the test, how they are interpreted by the test software and at the VID, how they might affect the trigger results, etc.

2. An understanding of motor vehicle emissions. This is needed to interpret trigger results involving average emissions scores or repeat emissions.

3. An understanding of how the test software works and how stations are actually conducting tests, which will help in interpreting the results of triggers that track the rate of anomalous test record entries (e.g., non-dyno testable, test aborts, manual VIN entries, etc.).

4. An understanding of how the test equipment works is needed to interpret the equipment trigger results. For example, results provided for the *Low NO Calibration Gas Drift* trigger by one state[6] showed the worst performer to be what appears to be a dead NO cell, which recorded a pre-cal reading of 10,000 ppm. This result is so bad that it widens the range of performance sufficiently to mask what appear to be many more problems with the NO cells. For example, an NO cell with an average pre-cal reading of 395 ppm (i.e., 30% above the average bottle value of 303 ppm) had a calculated index rating of 84.79. This rating is high enough to look good; however, the actual performance is quite poor. This illustrates the care that needs to be taken both in setting up the triggers index calculations and interpreting the equipment results. For example, readings that would bias the results, such as the 10,000 ppm NO value, could be ignored in computing the index ratings. However, any method used to exclude such results

needs to be well thought out. Those test systems that are excluded due to outlier results should also be identified as possible poor performers and targeted for follow-up action.

While state staff do not need to be experts in the areas of motor vehicles and emissions test systems, the more they know about these subjects the easier it will be for them to understand and use any of the recommended trigger results.

Control Chart and SPC Analysis Skills - If some level of control charting or other SPC analysis will be included as part of an overall QA/QC program, state (as well as any management contractor) staff will need to have sufficient expertise and training to correctly interpret the results displayed on the charts. The control charts will indicate whether each particular variable being tracked is out of control. They will not indicate whether this lack of control is due to poor performance or fraudulent actions by stations/inspectors, problems with test system hardware or software, or simply the heterogeneous nature of the test data.* While giving states an additional statistical tool to aid in assessing program performance is clearly beneficial, it is imperative that I/M program staffs be capable of and have experience in making the engineering judgements needed to interpret control charts and other SPC results, in order to correctly understand the potential root causes of variation in performance.

Manual Data Transcription

As part of the current work assignment, EPA directed Sierra to evaluate the extent to which any analysis approaches based on manual transcription of data could be used for trigger purposes. This could be done using two different approaches. Under the first approach, individual data elements for each test (e.g., emissions scores, pass/fail results for each inspection elements, etc.) would first be entered into an electronic database. The formulas presented later in this section could then be used to run those triggers for which the required data have been collected and entered electronically. The same approach could possibly be used to run any equipment triggers for which the required calibration data are available.

A second approach that would require less extensive data entry efforts would be to compute station-specific and network-wide results such as tailpipe and evaporative emissions test failure rates, and only enter these into an electronic database. While this would limit the triggers that could be run to those involving these parameters, it provides a way for a state to implement at least a minimal triggers program without spending huge resources on data entry.

*As noted previously, control charts are not recommended for use in tracking station and/or inspector performance. They should be used to track equipment performance only, while triggers and possibly other SPC results can be used to track station/inspector performance

One concern with any type of manual transcription-based triggers program is the accuracy of the underlying data. These data are likely to be subject to a significant degree of entry error, plus unscrupulous stations are able to enter false information with little control other than whatever level of auditing is being performed in the program.

Calculation Methodologies

General calculation formulas associated with each of the recommended triggers listed in Sections 3 and 4 are presented below. Unless otherwise noted, only use test results for gasoline-powered vehicles in completing the recommended calculations to avoid introducing potential bias into the results. Each of the formulas is designed to compute index numbers that range from 0-100, with a rating of 0 reflecting worst possible performance and a rating of 100 being assigned to the best performance for that trigger.

Use of the formulas is subject to any qualifications noted in the previous sections. The following descriptions also list any recommended user-configurable "factors" associated with each trigger (e.g., the period of time between inspections that is considered "excessively" short) that should be adjustable within the triggers system.

Individual states are expected to need to tailor the formulas to address program-specific issues (e.g., regarding what type of data are collected in the program and available for analysis). Some examples of such modifications are included below. Another common change that may be required would involve converting one or more of the formulas from the recommended format in which an index rating of 100 (i.e., best performance) is assigned to a 0% rate for the parameter that is being tracked to instead assign an index rating of 100 to the lowest rate recorded for an individual station.

For example, the formula shown for the *VIN Mismatch Rate* trigger involves assigning an index rating of 100 to a 0% mismatch rate. This is based on the presumption that a best performance rating should reflect zero mismatches between the VID obtained from the OBDII datastream and that obtained through the normal test process. However, OBDII-based VIN data are not yet being collected in any program. It is possible that subsequent implementation of this functionality will show that all stations in the network have some nominal level of VIN mismatches due to problems related to the OBDII data being captured from the vehicle or other factors outside the control of the station. If so, it would be appropriate to modify the formula to instead equate best performance to that achieved by the top-performing station. Similar adjustments may be needed in other trigger formulas.

Station/Inspector Triggers

1. **Test Abort Rate** – Assign an index number of 0 to the highest abort rate for all (Initial and After-repair) emissions tests recorded for an individual station. Assign an index number of 100 to a 0% abort test rate. The index number for a particular station will be computed as follows:

Index number = [(highest abort rate - station rate)/highest rate] x 100*

2. **Offline Test Rate** – Assign an index number of 0 to the highest offline rate for all emissions tests recorded for an individual station. Assign an index number of 100 to a 0% offline test rate. The index number for a particular station will be computed as follows:

   Index number = [(highest offline rate - station rate)/highest rate] x 100

3. **Number of Excessively Short Periods between Failing Emissions Test and Subsequent Passing Test for Same Vehicle** – Assign an index number of 0 to the highest number of occurrences in which the recorded difference between the end time of a failing emissions test and the start time of a subsequent passing test on the same vehicle is less than 10 minutes. This period is recommended as the initial value for determining when the time between tests is excessively short; however, it should be user-configurable to allow for possible future adjustment. Assign an index number of 100 to stations that have zero such occurrences. The index number for a particular station will be computed as follows:

   Index number = [(highest number - station number)/highest number] x 100

4. **Untestable on Dynamometer Rate** – Assign an index number of 0 to the highest rate of dynamometer-untestable entries recorded for all emissions tests conducted by an individual station. Assign an index number of 100 to a 0% untestable rate. The index number for a particular station will be computed as follows:

   Index number = [(highest untestable rate - station rate)/highest rate] x 100

5. **Manual VIN Entry Rate** – Assign an index number of 0 to the highest rate of manual VIN entries for 1990 and newer model year vehicles recorded for all tests conducted by an individual station. Assign an index number of 100 to a 0% manual rate. The index number for a particular station will be computed as follows:

   Index number = [(highest manual entry rate - station rate)/highest rate] x 100

6. **Underhood Visual Check Failure Rate** – Compute station-specific and network-wide visual check failure rates for all Initial emissions tests (based on the visual check inspection result contained in each Initial test record) for each separate model year included in test records for the analysis period.** Exclude any model year with fewer than 10 records from the station-specific analysis, and do not compute a trigger rating

*For this and all other formulas, if the value of the denominator is equal to zero the calculation should not be performed to avoid dividing by 0. Instead, all stations will be assumed to have an index number of 100.
**This is needed to weight the trigger results by model year.

for any station with fewer than 30 records total.* Index numbers will then be computed on a model-year-specific basis. Assign an index number of 0 to the lowest failure rate computed for an individual station for each model year. Assign an index number of 100 to the highest failure rate computed for an individual station for the same model year. Compute the index number for a particular station and model year as follows:

Index number = [(station failure rate – lowest average failure rate)/
(highest average failure rate – lowest average failure rate)] x 100

Model-year-specific index numbers for each station will then be averaged to compute an overall index number for the station. Model-year-specific index numbers that are missing due to a station testing fewer than 10 vehicles for those model years will be omitted from the calculation.

7. **Underhood Functional Check Failure Rate** – Use the same computational approach as described under the *Underhood Visual Check Failure Rate* trigger, except that the results will be based on the functional check inspection result contained in each Initial test record.

8. **Functional Gas Cap Check Failure Rate** – Use the same computational approach as described under the *Underhood Visual Check Failure Rate* trigger, except that the results will be based on the gas cap check inspection result contained in each Initial test record.

9. **Functional Vehicle Evaporative System Pressure Test Failure Rate** –Use the same computational approach as described under the *Underhood Visual Check Failure Rate* trigger, except that the results will be based on the pressure test inspection result contained in each Initial test record.

10. **Emissions Test Failure Rate** – Use the same computational approach as described under the *Underhood Visual Check Failure Rate* trigger, except that the results will be based on the overall emissions inspection result contained in each Initial test record.**

*Using 30 records/model year as the cutoff for the station-specific analysis, while proper from a statistical perspective, could result in many low volume stations being totally excluded from the triggers analysis. It is unlikely that these shops would test this many vehicles for most if not all model years subject to the program in a month. A lower cutoff of 10 records/model year was therefore selected to ensure that these stations would be included in the analysis. This is equivalent to applying a cutoff of 30 records to individual 3-year ranges of model years, which is judged to be a sufficient criterion to minimize bias due to differences in vehicle age distributions among inspection stations. A second cutoff criterion of at least 30 records total in overall station test volume during the analysis period will also be applied.

**This analysis should be limited to the primary emissions test method used in the program. For example, a program may test most vehicles using a transient loaded mode test, but subject older and non-dyno testable vehicles to an idle or TSI test. In this case, only the Initial transient emissions test results should be used in this triggers calculation. If desired, the program can also conduct the same calculation separately on the idle or TSI results.

11. **Average HC Emissions Score** – Compute station-specific and network-wide HC emissions scores on Initial emissions tests (involving the primary emissions test method used in the program) for each separate model year included in test records for the analysis period. Exclude any model year with fewer than 10 records from the station-specific analysis, and do not compute a trigger rating for any station with fewer than 30 records total. Index numbers will then be computed on a model-year-specific basis. Assign an index number of 0 to the lowest average HC emissions score on the Initial test computed for an individual station for each model year. Assign an index number of 100 to the highest average HC emissions score on the Initial test computed for an individual station for the same model year. Compute the index number for a particular station and model year as follows:

    Index number = [(station average HC emissions – lowest average emissions)/
    (highest average emissions – lowest average emissions)] x 100

    Model-year-specific index numbers for each station will then be averaged to compute an overall index number for the station. Model-year-specific index numbers that are missing due to a station testing fewer than 10 vehicles for those model years will be omitted from the calculation.

12. **Average CO Emissions Score** – Use the same computational approach as described under the *Average HC Emissions Score* trigger, except that the results will be based on the CO emissions scores on Initial emissions tests involving the primary emissions test method used in the program.

13. **Average NO or NOx Emissions Score** – Use the same computational approach as described under the *Average HC Emissions Score* trigger, except that the results will be based on the NOx (or NO) emissions scores on Initial emissions tests involving the primary emissions test method used in the program.

14. **Repeat Emissions** – This trigger involves using the statistical analysis approach described in Section 3 to identify stations with a high number of similar emission readings that may result from clean-piping. It is recommended that the index score for this trigger should be based upon the number of clusters.* The trigger calculations are structured such that a large number of clusters receives an index score closer to zero and a low number of clusters receives a high score. This is done by assigning an index number of 0 to the highest number of clusters found for an individual station. An index number of 100 is assigned to zero clusters. The index number for a particular station will be computed as follows:

    Index number = [(highest number of clusters - station number of clusters)/
    highest number of clusters] x 100

*States that decide to pursue statistical cluster analysis may want to try alternative methods for indexing the results, to see which approach is most effective in identifying actual clean pipers.

15. **Unused Sticker Rate** – Assign an index number of 0 to the highest rate of unused stickers recorded for an individual station.[*] Assign an index number of 100 to a 0% unused sticker rate. The index number for a particular station will be computed as follows:

    Index number = [(highest unused sticker rate - station rate)/highest rate] x 100

16. **Sticker Date Override Rate** – Use the same computational approach as described under the *Unused Sticker Rate* trigger, except that the results will be based on the rate of sticker date overrides recorded for an individual station.[**]

17. **After-Hours Test Volume** – Assign an index number of 0 to the station with the highest number of passing tests recorded in which the start time occurs during the after-hours period.[***] It is recommended that this period be initially defined as beginning at 7:00 p.m. and ending at 5:00 a.m, unless stations routinely perform tests within this time frame in a particular program. The start and end times of the after-hours period should be user-configurable in the triggers software to allow for future adjustment as experience is gained with this trigger. Assign an index number of 100 to stations that started no passing tests within the after-hours period. The index number for a particular station will be computed as follows:

    Index number = [(highest after-hours passing test volume - station volume)/ highest volume] x 100

18. **VID Data Modification Rate** – Compute station-specific and network-wide rates of occurrences of inspector changes to the vehicle identification data transmitted from the VID for all Initial tests for each separate model year included in test records for the analysis period.[****] Exclude any model year with fewer than 10 records from the station-specific analysis, and do not compute a trigger rating for any station with fewer than 30 records total. Index numbers will then be computed on a model-year-specific basis. Assign an index number of 0 to the highest rate of VID data modifications computed for an individual station for each model year. Assign an index number of 100 to a 0% VID data modification rate. Compute the index number for a particular station and model year as follows:

    Index number = [(highest VID data modification rate – station rate)/(highest rate)] x 100

[*] Unused stickers are those that are voided, stolen, damaged or othewise missing; thus, a high unused sticker rate is considered evidence of questionable performance.

[**] A high rate of sticker date overrides is also considered suspect, since it may indicate a station is illegally changing the dates assigned by the test system to aid motorists in evading program requirements.

[***] A high volume of after-hours passing tests is an indication that a station my be trying to conceal fruadulent behavior.

[****] This trigger can be used if the specified test record format includes a data field that is automatically filled to indicate whether the VID data were modified. If this is not available, a post-test comparison of the data contained in the record that transmitted from the VID would be required. Computational and logistical difficulties are expected to make this latter approach infeasible.

The model-year-specific index numbers for each station will then be averaged to compute an overall index number for the station. Model-year-specific index numbers that are missing due to a station testing fewer than 10 vehicles for those model years will be omitted from the calculation.

19. **Safety-Only Test Rate** – Use the same computational approach as described under the *Underhood Visual Check Failure Rate* trigger, except that the results will be based on safety-only test rates for all Initial tests for each separate model year subject to emissions testing.*

20. **Non-Gasoline Vehicle Entry Rate** – Use the same computational approach as described under the *Underhood Visual Check Failure Rate* trigger, except that the results will be based on rates of non-gasoline-fueled vehicle entries for all Initial emissions tests.

21. **Non-Loaded Mode Emissions Test Rate** – Use the same computational approach as described under the *Underhood Visual Check Failure Rate* trigger, except that the results will be based on non-loaded mode emissions test rates for all Initial emissions tests.

22. **Frequency of Passing Tests after a Previous Failing Test at Another Station** – Assign an index number of 0 to the highest rate of occurrences in which a station passes a vehicle that previously failed at a different test station during the same inspection cycle. Assign an index number of 100 to the lowest rate computed for an individual station. Compute the index number for a particular station as follows:

    Index number = [(highest rate of subsequent passing tests – station rate)/
    (highest rate – lowest rate)] x 100

23. **Average Exhaust Flow** – If average exhaust flow during the transient test is recorded in the primary test record, these data will be used in this trigger calculation. Otherwise, the first analysis step is to compute average exhaust flow for each transient test based on the second-by-second (modal) exhaust volume data recorded in the second-by-second test file. Because the relationship between exhaust flow and vehicle test weight is approximately linear, it is possible to normalize the average readings for each test to a single test weight. This is preferable to establishing test weights bins due to the large number of bins that would be required. Therefore, the next step is to normalize average exhaust flow for each test to an equivalent test weight (ETW) of 3,000 lbs (representing an average vehicle), using the following equation:

    Normalized exhaust flow = (actual exhaust flow x 3,000 lbs) / actual ETW

*Combined emissions and safety inspection programs often conduct safety-only tests on older vehicles that are not subject to emissions testing. Data from these tests should not be included in this trigger calculation.

The normalized data for each station will next be placed into four engine displacement bins to ensure that station-specific differences in average displacement among the test data do not bias the trigger results. The following criteria should be used to bin the data:

- Bin 1 = 0-2.5 liters (L)
- Bin 2 = 2.6-4.0L
- Bin 3 = 4.1-5.5L
- Bin 4 = 5.6L and greater

The next step is to compute station-specific average normalized exhaust rates for each of the above four engine displacement bins. Then compute index numbers on a bin-specific basis using the approach described below. Exclude any displacement bin with fewer than 10 records from the station-specific analysis, and do not compute a trigger rating for any station with fewer than 30 records total.

Assign an index number of 0 to the lowest average normalized exhaust flow reading computed for an individual station for each displacement bin. Assign an index number of 100 to the highest average normalized exhaust flow reading computed for an individual station for each displacement bin. Compute the index number for a particular station and displacement bin as follows:

Index number = [(station normalized exhaust flow – lowest exhaust flow)/
(highest exhaust flow – lowest exhaust flow)] x 100

Next average the displacement bin-specific index numbers for each station to compute an overall index number for the station. Displacement bin-specific index numbers that are missing due to a station testing fewer than 10 vehicles for those bins will be omitted from the computation.

24. **Exhaust Flow Change between Failing Test and Subsequent Passing Test** – If average exhaust flow has not been recorded in the primary test record, compute this parameter for each transient test based on the modal exhaust volume data recorded in the second-by-second test file. Because the trigger looks at the percentage change between a failing and subsequent passing test for the same vehicle, there is no need to bin or normalize the exhaust flow data. Instead, the percentage change in flow is computed for each paired set of test results. To do this, all passing transient retests are first identified. All corresponding immediately preceding failing tests are then flagged, and the percentage change in flow between each set of paired tests is computed using the following equation:

Percent change in flow = (flow for passing retest – flow for preceding failing test)/
flow for preceding failing test

Next compute station-specific rates of excessive changes in exhaust flow by dividing the number of paired tests that show more than a 20% decrease in flow[*] by the total number of such paired tests for each station. Assign an index number of 0 to the highest rate of excessive changes in exhaust flow computed for an individual station. Assign an index number of 100 to a 0% rate of excessive changes. Compute the index number for a particular station as follows:

Index number = (highest rate – station rate)/(highest rate) x 100

As discussed in Section 4, it is recommended that states also compute the average change in exhaust flow for all paired tests for each station; however, this result will not be used in the trigger computation.

25. **Frequency of No Vehicle Record Table (VRT) Match or Use of Default Records** – Assign an index number of 0 to the highest rate of no VRT match or use of default records[**] that was recorded for an individual station. Assign an index number of 100 to the lowest rate computed for an individual station. Compute the index number for a particular station as follows:

    Index number = [(highest rate of no VRT match or use of default records – station rate)/(highest rate – lowest rate)] x 100

26. **Drive Trace Violation Rate** – Assign an index number of 0 to the highest rate of drive trace violations that was recorded for an individual station. Assign an index number of 100 to the lowest rate computed for an individual station. Compute the index number for a particular station as follows:

    Index number = [(highest rate of drive trace violations – station rate)/ (highest rate – lowest rate)] x 100

27. **Average Dilution Correction factor (DCF)** – First identify all AIS-equipped vehicles on the basis of recorded visual/functional entries. Any test records that have other than an "N" recorded in any of the AIS fields will be excluded from subsequent analysis.[***] For a TSI test, the DCF values recorded in the test record for both the

[*] This criterion is recommended for initial use in flagging excessive decreases in exhaust flow between tests; however, it should be user-configurable in case future adjustment is needed. Also, as noted in Section 3, the average change in exhaust flow should be also computed and average decreases of more than 10% used on a stand-alone basis (i.e., independent of the results of this trigger) to identify possible fraudulent inspections.

[**] The exact data to be used for this trigger will depend on the specific contents of the test record. For example, in some programs VRT lookup is performed only if VID connection is unsuccessful or if no data are returned after connecting to the VID. For such programs, the trigger calculations would involve computing the number of occurrences for each station when an offline test was performed and a value of 0 was recorded for the VRT record ID (indicating that no match was found). For other programs, the calculations would need to be tailored differently.

[***] This presumes that the AIS-related entries in the test records are correct. While a small fraction of entries (most likely under 10%) may actually be wrong due to inspector entry errors, the small degree of

(continued...)

curb and high idle portions of all remaining tests should then be averaged. DCF values for the individual modes of an ASM2 test should also be averaged. Single DCF values recorded for a curb idle or single-mode ASM test can simply be used.

Model-year-specific average DCFs will be computed for each station and the overall network. More than one type of steady-state test may be conducted in a program (e.g., ASM tests may be conducted on most vehicles with TSI tests performed on older and non-dyno testable vehicles). If so, use the DCF values from all steady-state tests (regardless of test type) for this calculation. If TSI or ASM2 tests are included, average DCFs for each test should first be computed and then combined with DCF values for the other tests to produce overall average DCF values for each station and the network. Exclude any model year with fewer than 10 records from the station-specific analysis, and do not compute a trigger rating for any station with fewer than 30 records total.

Index numbers will then be computed on a model-year-specific basis. An index number of 0 should be assigned to the highest average DCF computed for an individual station for each model year. An index number of 100 should be assigned to the lowest average DCF computed for an individual station for the same model year. The index number for a particular station and model year will be computed as follows:

Index number = [(highest average DCF – station average DCF)/
(highest average DCF – lowest average DCF] x 100

The model-year-specific index numbers for each station will then be averaged to compute an overall index number for the station. Model-year-specific index numbers that are missing due to a station testing fewer than 30 vehicles for those model years will be omitted from the computation.

28. **RPM Bypass Rate** – Assign an index number of 0 to the highest RPM bypass rate that was recorded for an individual station. Assign an index number of 100 to the lowest rate computed for an individual station. Compute the index number for a particular station as follows:

    Index number = [(highest RPM bypass rate – station rate)/(highest rate – lowest rate)] x 100

29. **OBDII MIL Key-On Engine-Off (KOEO) Failure Rate** – Compute station-specific and network-wide rates of OBDII KOEO failures (based on the pass/fail result recorded in each Initial test record) for each separate 1996 and newer model year included in test records for the analysis period. Exclude any model year with fewer than 10 records from the station-specific analysis, and do not compute a trigger rating for any station with fewer than 30 records total. Index numbers will then be

(...continued)
error resulting from incorrect entries is considered acceptable.

computed on a model-year-specific basis. Assign an index number of 0 to the lowest KOEO failure rate computed for an individual station for each model year. Assign a index number of 100 to the highest KOEO failure rate computed for an individual station for the same model year. The index number for a particular station and model year will be computed as follows:

Index number = [(station KOEO failure rate – lowest failure rate)/
(highest failure rate – lowest failure rate] x 100

Model-year-specific index numbers for each station will then be averaged to compute an overall index number for the station. Model-year-specific index numbers that are missing due to a station testing fewer than 30 vehicles for those model years will be omitted from the computation.

30. **OBDII MIL Connection Failure Rate** – Compute station-specific and network-wide rates of OBDII non-connect failures (based on the pass/fail result recorded in each Initial test record) for each separate 1996 and newer model year included in test records for the analysis period. Exclude any model year with fewer than 10 records from the station-specific analysis, and do not compute a trigger rating for any station with fewer than 30 records total. Index numbers will then be computed on a model-year-specific basis. Assign an index number of 0 to the highest non-connect rate computed for an individual station for each model year. Assign an index number of 100 to a 0% non-connect rate. The index number for a particular station and model year will be computed as follows:

Index number = [(highest OBDII non-connect rate – station rate)/highest rate]
x 100

Model-year-specific index numbers for each station will then be averaged to compute an overall index number for the station. Model-year-specific index numbers that are missing due to a station testing fewer than 30 vehicles for those model years will be omitted from the computation.

31. **Overall OBDII Failure Rate** – Use the same computational approach as described under the *OBDII MIL Key-On Engine Off (KOEO) Failure Rate* trigger, except that the results will be based on the overall OBDII pass/fail result recorded in each Initial test record.

32. **VIN Mismatch Rate** – First calculate VIN mismatch rates for each station and the overall network. Identify all test records in which an OBDII-obtained VIN has been recorded and compare it to the standard VIN* recorded in the test record. Divide the number of "two-VIN" records in which the VINs are not identical by the total number of such records to compute the VIN mismatch rate for each station. Assign an index number of 0 to the highest mismatch rate that was recorded for an individual station.

*Depending on the program design and specific test result, this may be provided by the VID, scanned from the vehicle using a bar code reader, or manually entered by the inspector.

Assign an index number of 100 to a 0% mismatch rate. Compute the index number for a particular station as follows:

Index number = [(highest VIN mismatch rate – station rate)/(highest rate)] x 100

33. **OBDII PID Count/PCM Module ID Mismatch Rate** – To run this trigger, correct PID Count and PCM Module IDs for each applicable make/model combination must be developed and included in the existing VRT or a separate electronic lookup table incorporated into the triggers software. To determine the trigger results, first calculate the PID Count/PCM Module ID mismatch rate for each station and the overall network. Identify all test records in which PID Count and PCM Module IDs have been recorded and compare these data to the parameter values contained in the lookup table. Divide the number of "PID/PCM Module ID" records in which these parameters are not identical to those shown for the same make and model in the lookup table by the total number of such records to compute the mismatch rate for each station. Assign an index number of 0 to the highest mismatch rate that was recorded for an individual station. Assign an index number of 100 to a 0% mismatch rate. Compute the index number for a particular station as follows:

    Index number = [(highest mismatch rate – station rate)/(highest rate)] x 100

34. **Lockout Rate** – Assign an index number of 0 to the highest number of applicable lockouts that were recorded for an individual station during the analysis period. As noted in Section 3, the exact lockouts to be used in this calculation will be very program-specific. In general, security and other station-initiated lockouts (i.e., those caused by an action of the station) that are set on the test system should be counted in this calculation. Non-station-initiated lockouts (e.g., calibration failures and file corruption errors) should not be included since they are not a good measure of station/inspector performance. Assign an index number of 100 to zero lockouts. Compute the index number for a particular station as follows:

    Index number = [(highest number of lockouts – station lockouts)/
    (highest number of lockouts)] x 100

35. **Waiver Rate** – First calculate waiver rates for each station and the overall network by dividing the number of issued waivers by the number of passing emissions tests. Assign an index number of 0 to the highest waiver rate calculated for an individual station. Assign an index number of 100 to a 0% waiver rate. The index number for a particular station will be computed as follows:

    Index number = [(highest waiver rate - station waiver rate)/highest rate] x 100

36. **Diesel Vehicle Inspection-Related Triggers** – As noted in Section 3, states that include emissions inspections of Diesel vehicles in their programs may also choose to run separate triggers aimed at the performance of these tests. Many of the above triggers (e.g., *Offline Test Rate*, *Emissions Test Failure Rate*, *Safety-Only Test Rate*, etc.) are directly applicable to Diesel emissions tests. In such cases, formulas

identical to those shown above can be combined with Diesel test data to produce these trigger results.

To compute separate trigger results for average light-duty and heavy-duty Diesel opacity scores, the following approach should be used. Compute station-specific and network-wide average opacity scores on Initial Diesel emissions tests for each separate model year included in test records for the analysis period. Exclude any model year with fewer than 10 records from the station-specific analysis, and do not compute a trigger rating for any station with fewer than 30 records total. Index numbers will then be computed on a model-year-specific basis. Assign an index number of 0 to the lowest average opacity score on the Initial test computed for an individual station for each model year. Assign an index number of 100 to the highest average opacity score on the Initial test computed for an individual station for the same model year. Compute the index number for a particular station and model year as follows:

Index number = [(station average opacity   lowest average opacity)/
(highest average opacity – lowest average opacity)] x 100

Model-year-specific index numbers for each station will then be averaged to compute an overall index number for the station. Model-year-specific index numbers that are missing due to a station testing fewer than 10 vehicles for those model years will be omitted from the calculation.

Equipment Triggers

1. **Analyzer Calibration Failures** – As discussed in Section 4, the 2-point calibration method incorporated into some brands of analyzer software used in many programs will mask any bench calibration problems. This trigger can be run in such programs; however, its results will be less meaningful.* The calibration procedure for these analyzer brands therefore needs to be modified to one in which an independent check of the bench calibration, using the low gas, does occur. This change would produce accurate calibration results and remove the bias from this trigger.

   Assign an index number of 0 to the highest rate of overall analyzer bench calibration failures recorded for an individual analyzer. An index number of 100 should be assigned to analyzers that have no calibration failures. The index number for a particular analyzer will be computed as follows:

   Index number = [(highest bench calibration failure rate – analyzer failure rate)/
   highest failure rate] x 100

*Although an analyzer that incorporates the 2-point calibration method cannot fail the "low cal check," it could still fail a bench calibration due to an excessive response time if this calibration functionality is active. As a result, this trigger would still have some value in programs that choose to continue to allow the use of the 2-point calibration method.

2. **Leak Check Failures** – Assign an index number of 0 to the highest rate of leak check failures for an individual analyzer. An index number of 100 should be assigned to analyzers that have no leak check failures. The index number for a particular analyzer will be computed as follows:

   Index number = [(highest leak check failure rate – analyzer failure rate)/ highest frequency] x 100

3. **Low Calibration Gas Drift** – This trigger tracks average calibration drift on a pollutant-specific basis (e.g., for HC, CO or NO). For the pollutant of interest, first calculate the average difference between the low calibration gas value and the actual reading over the analysis period. Data from both passing and failing calibrations should be used in the calculation. Assign an index number of 0 to the greatest average difference between the cylinder value and the measured reading for the low gas. Assign an index number of 100 to zero average difference between the two sets of values. The index number for a particular analyzer will be computed as follows:

   Index number = [(greatest average difference in drift – analyzer difference)/ greatest difference] x 100

   If a state wishes to use this trigger to track analyzer performance for more than one pollutant, it is recommended that CO, and NO drift be tracked. HC drift does not need to be tracked separately since HC measurement performance should generally track CO performance.[*]

4. **Analyzer Response Time** – Reference T_{90} values are stored in memory for each of the channels when the analyzer bench is first manufactured or installed in the test system (in the case of a bench replacement). Actual T_{90} values are then measured during each calibration and compared to the reference values. This trigger tracks the average difference between these actual and reference T_{90} values. Possible response time triggers include the following:

   a. A CO response time trigger will generally track both CO and HC analyzer performance;

   b. An NO response time trigger will track NO cell performance in a loaded mode program; and

[*] As discussed previously, HC and CO performance should track fairly closely for SPX-brand analyzers since the same Horiba analytical bench is used for both constituents. The Sensors analytical benches that are used by ESP, Snap-On, and WorldWide have separate sample tubes for HC and CO (as well as a third one for CO_2); however, the same raw exhaust gas is flowed through them. It is therefore expected that degradation would occur in both HC and CO bench performance if the cause is "filming over" of the optical windows due to dirty exhaust gases. Other bench-specific problems (e.g., with its electronics or optical source) would affect the performance of only that bench. However, these are less common and would be expected to result in total bench failure (and subsequent repair).

c. An O_2 response time trigger will oxygen sensor performance. As noted in Section 4, this is not considered very important except in programs involving VMAS™ flow measurement, since O_2 is not a pass/fail pollutant.

To run any of the response time triggers, both the reference and actual T_{90} values applicable to the pollutant of interest must be recorded in the calibration file. Assign an index number of 0 to the greatest average difference between the actual and reference values for a single analyzer. Assign an index number of 100 to zero average difference between the two sets of values. The index number for a particular analyzer will be computed as follows:

Index number = [(greatest average difference – analyzer average difference)/ greatest average difference] x 100

5. **NO Serial Number** – To run this trigger, the NO cell serial number must be recorded in the calibration record or other file (e.g., the FSR log). Assign an index number of 0 to analyzers whose calibration records show the same NO cell serial number for the last 12 months. Assign an index number of 100 to analyzers whose calibration records show the same NO cell serial number for the last 3 months or less. Index numbers for all other analyzers will be calculated as follows:

   Index number = [1.33 * (12 – number of months since analyzer NO serial number was changed)/12] x 100

6. **Calibrations per Gas Cylinder** – In programs involving multiple brands of test systems, the first step is to normalize the data to eliminate bias due to manufacturer type. This is done by computing the average amount of gas used per cylinder by each brand on a network-wide basis. The resulting averages are then used to normalize the data from the individual analyzers, using the following equation:

   Analyzer gas usage = (analyzer calibrations per cylinder/network average calibrations per cylinder for that brand) x 100

   Thus, an analyzer that has conducted 30 calibrations using one gas bottle compared to the network average of 40 would be assigned a normalized usage of (30/40) x 100, or 75. This calculation should be done separately for the high and low calibration gases.

   Next assign an index number of 0 to the greatest number of normalized calibrations recorded with the previous gas cylinder lot number for an individual analyzer. Based on the above equation, an index number of 100 will be equal to the average normalized number of calibrations per gas cylinder. If the number of normalized calibrations for a particular analyzer is less than the average, the index number should also be set to 100. The index number for all analyzers with a higher than average number of calibrations will be computed as follows:

   Index number = [(highest number of calibrations – analyzer number of calibrations)/(highest number of calibrations – 100)] x 100

As noted in Section 4, this trigger should be run separately for the low and high calibration gases.

7. **Dynamometer Coast Down Calibration Failures** – Assign an index number of 0 to the highest rate of coast down calibration failures recorded for an individual test system. An index number of 100 should be assigned to test systems that have no coast down failures. The index number for a particular test system will be computed as follows:

   Index number = [(highest coast down failure rate – test system failure rate)/ highest failure rate] x 100

8. **Changes in Dynamometer Parasitic Values** – This trigger tracks the occurrence of what are considered excessive changes in calculated parasitic losses, as recorded in the calibration file. Assign an index number of 0 to a test system if the Calculated Parasitic Loss at any of the speed points recorded in the calibration record changes by 10% or more between successive parasitic determinations. While 10% is recommended as the initial value to be used in determining when the change in parasitic values is excessive, it should be user-configurable in case future adjustment is found to be needed. Assign an index number of 100 to a 0% frequency. The index number for a particular test system will be computed as follows:

   Index number = [(10% – largest percent change in successive parasitic values at any speed point)/10%] x 100

9. **Gas Cap Tester Calibration Failures** – Assign an index number of 0 to the highest rate of gas cap tester calibration failures recorded for an individual test system. Assign an index number of 100 to test systems that have no gas cap calibration failures. The index number for a particular test system will be computed as follows:

   Index number = [(highest gas cap tester calibration failure rate – test system failure rate)/highest failure rate x 100

10. **Pressure Tester Calibration Failures** – Assign an index number of 0 to the highest rate of vehicle evaporative system pressure tester calibration failures recorded for an individual test system. Assign an index number of 100 to test systems that have no pressure tester calibration failures. The index number for a particular test system will be computed as follows:

    Index number = [(highest pressure tester calibration failure rate – test system failure rate)/highest failure rate x 100

11. **Opacity Meter Calibration Failures** – Assign an index number of 0 to the highest rate of opacity meter calibration failures recorded for an individual test system.

Assign an index number of 100 to test systems that have no opacity meter calibration failures. The index number for a particular test system will be computed as follows:

Index number = [(highest opacity meter calibration failure rate – test system failure rate)/highest failure rate x 100

12. **BAR97 O_2 Analyzer Response Time** – This trigger was previously mentioned under the *Analyzer Response Time* trigger. To run the trigger, both reference and actual T_{90} values for O_2 must be recorded in the calibration file. Assign an index number of 0 to the greatest average difference between the actual and reference values for a single analyzer. Assign an index number of 100 to zero average difference between the two sets of values. The index number for a particular analyzer will be computed as follows:

    Index number = [(greatest average difference – analyzer average difference)/ greatest average difference] x 100

13. **Excessive Differences in VMAS™ Flow** – Assign an index number of 0 to the highest rate of occurrences in which the reference hose-off flow value (determined for each VMAS™ unit at the time of manufacture) differs from actual hose-off flow measurements that are performed during in-use calibrations by more than 10%. While 10% is recommended as the initial value to be used in determining when the difference in measured versus reference flow values is considered excessive, it should be user-configurable in case future adjustment is found to be needed (e.g., based on actual VMAS™ flow data collected by the program). Assign an index number of 100 to a 0% rate. The index number for a particular test system will be computed as follows:

    Index number = [(highest rate of excessive differences – test system rate)/ highest rate] x 100

Trigger Weightings

As discussed previously, it is recommended that trigger weightings be developed as part of an overall triggers system, and used to combine selected trigger results into a smaller set of composite station and inspector results.* This objective of this design element is to reduce the volume of trigger results that program staff will need to sift through to identify likely problem performers. To Sierra's knowledge, no programs that are currently running triggers systems are using such an approach, most likely for the following reasons:

1. The existing triggers systems are relatively new and thus do not represent what might be considered a "mature" use of this type of analytical system.

*Equipment triggers will not be weighted.

2. Most states are using relatively few triggers, each of which is targeted at a particular performance concern specific to their programs. The results from each trigger are therefore important in tracking these individual performance concerns.

3. The method being used by most states to perform their triggers analyses is relatively simple and does not involve much computational complexity outside of that needed to calculate the individual trigger results. Attempting to add such a composite weighting scheme is likely seen as more trouble than it is worth.

Despite the fact that no states are currently generating composite trigger results, it is recommended that it be considered as part of developing a comprehensive triggers system. The following recommendations are therefore provided to aid those states that may be interested in this functionality.

To implement this element, the station/inspector triggers that have been included in the triggers system should be evaluated to determine which ones can be logically grouped together. Once the logical groupings have been determined, each trigger within the individual groups will be assigned a weighting. These weights will then be used to translate the individual triggers results into composite ratings for each station or inspector.

Table 11 illustrates how trigger weightings and composite ratings could be used in a program. In this example, 12 total triggers have been grouped into the following four composite groups:

- Clean piping;
- Fraudulent testing;
- Evaporative emissions; and
- Sticker fraud.

Each of the triggers in the four groups are aimed at tracking performance related to the subject category. Various weightings have been assigned to each of the triggers that reflect their expected relative effectiveness in identifying problem performance related to the category. For example, the repeat emissions trigger is heavily weighted in the clean piping group due to its expected strength in identifying such behavior.

The table also contains example index numbers and composite ratings for a hypothetical station. In this example, it can be seen that the *Repeat Emissions* trigger has identified the station as a potential clean piper, even though the other clean piping related triggers do not show as clear a picture of this. The station is rated very well by most of the fraudulent testing triggers, which could be simply because it has found clean piping to be an easier way to cheat. However, the after-hours test trigger also shows poor performance. Considered in combination, it appears that the station (or a single inspector employed by the station) may be conducting after-hours clean piping.

The composite evaporative emissions rating is also poor, thus indicating that the station is likely engaging in fraudulent performance on this portion of the test procedure as well. The station's sticker fraud composite rating is excellent; i.e., it does not appear to be cheating in this manner.

Although not shown on the table, an overall composite rating could also be computed by assigning relative weightings to each of the individual composite ratings. For example, assigning equal weights of 25% to each of the four individual composite groups would produce an overall composite rating of 61.0 for the example station. It is important to note that this overall rating gives relatively little indication of the very poor station performance seen with the *Repeat Emissions* or evaporative emissions-related triggers. This illustrates the importance of digging deeper into the trigger results for poor performers to better understand the underlying causes for such ratings.

Table 11
Example Trigger Weightings and Composite Ratings

Trigger No.	Trigger Description	Default Weight	Example Index Number
Clean Piping Composite Group			
3	Short periods between failing and subsequent passing tests	20%	85.7
11	Average HC emissions scores	15%	55.2
12	Average NO emissions scores	15%	61.6
14	Repeat emissions	50%	25.7
		Composite Rating	45.7
Fraudulent Testing Composite Group			
2	Offline test rate	40%	89.8
4	Untestable on dynamometer rate	20%	93.4
17	After-hours test volume	20%	33.5
18	VID data modification rate	20%	97.8
		Composite Rating	80.9
Evaporative Emissions Composite Group			
8	Functional gas cap check failure rate	50%	22.3
9	Functional vehicle pressure test failure rate	50%	18.9
		Composite Rating	20.6
Sticker Fraud Composite Group			
15	Unused sticker rate	50%	95.6
16	Sticker date override rate	50%	97.8
		Composite Rating	96.7

This example demonstrates the type of information that is available through the triggers system and how it can be evaluated in combination to provide insight into station performance. The triggers system should allow users to accept the defaults shown in the table, change any of the values (total weightings for each composite rating must remain 100%) for a single report cycle, or change the values and save them as updated defaults. In the latter case, the original defaults will be saved as a separate table for possible future use. Users should also be able to specify 100% for a single weighting factor, to view the effect of only that particular performance trigger on the composite station ratings.

Based on the above computations and the weightings specified by the user, a single performance index score will be computed for each station and composite grouping. For stations that have an insufficient number of test records to allow computation of one or more trigger ratings, the weightings for the missing results will be ignored and the remaining weightings renormalized to a total of 100%, prior to the computation of the overall composite ratings.

###

8. PERFORMANCE FEEDBACK

As noted in Section 2, it is critical that the performance of the selected triggers in correctly identifying poor performers be evaluated through some type of active feedback loop. The number of recommended triggers is obviously well above the amount needed for an optimized triggers system. Any state that attempted to implement and run all of these triggers would likely be overwhelmed by the sheer volume of data generated by the system. Staff would spend all of their time reviewing and attempting to interpret the results rather than using the information generated by the system to effect meaningful improvements in inspection and equipment performance. This is just the opposite of the desired objective, which is to increase the effectiveness of overall QA/QC efforts in improving this performance.

It is not readily apparent which of the recommended triggers would be best for each existing decentralized program. Just as each program has been designed in a somewhat unique fashion to fit individual state needs and desires, so each corresponding trigger system would need to be custom designed to meet the state's objectives and policies. For this reason, it cannot be simply recommended that all states implement triggers "A through F"; instead each state is likely to need to assess triggers "A to Z" before deciding on which ones are the best fit for its program.

A key issue in making this assessment is determining which triggers are most effective in identifying problem stations, inspectors, and test systems in each particular program. While these findings are expected to be relatively consistent among programs, they will not be identical. Differences in program design, test procedures, data elements being recorded and transmitted to the VID, etc., will affect which triggers work best for each specific program.

There are also other important reasons for having some method for evaluating the efficacy of the individual triggers. One is to ensure that one or more of the selected triggers do not seriously misidentify stations and equipment as problem performers. Such an outcome would reduce the efficiency of this approach, increase the cost of required follow-up efforts, elicit strong negative reactions from the misidentified stations, and potentially lead to the suspension or abandonment of the triggers system due to concern over its accuracy. Another reason is that stations engaged in intentional fraud are expected to realize over time how they are being identified and modify their behavior in an attempt to find another loophole in the system. This concern has led at least one state to address this issue by changing its active triggers on a regular basis.[6] It is therefore likely that triggers implementation will in some regards be a constant process, in which new triggers are developed and evaluated on an on-going basis. As a result, it is essential that some type of feedback process be designed and incorporated into the triggers system to provide a means of such on-going evaluation.

One method for accomplishing this is to track the success rate of the identification of problem stations and test systems relative to the results found through subsequent investigations (e.g., station counseling sessions, overt and covert audits, FSR visits, etc.). Under this approach, the success of each trigger "identification" would be rated whenever a follow-up investigation is performed. A simple rating scheme (e.g., involving "yes" or "no" results) can be used to document the relative success of each trigger. Some potential problem stations and test systems will be identified by multiple triggers, in which case the success of each of the triggers can be documented through a follow-up investigation. The overall success rate (i.e., the total number of documented "yes" responses divided by the total number of "yes" and "no" responses) can be computed for each trigger and used to determine the relative success of all the triggers.

The above approach requires a great deal of follow-up; however, the state should be performing such follow-ups anyway to improve the performance of the identified stations and equipment. In addition, there is a second type of feedback method that would not require such follow-up activities. This approach involves conducting "post-mortem" reviews of trigger results for problem stations and test systems that are identified through other means, to determine which triggers would have been successful in identifying the stations/test systems. The primary issue with this method is the degree to which such problem performers would be identified independent of the triggers system. The most obvious mechanism would be through any random overt and covert audits conducted by the state. Other possible means include complaints from consumers or other stations, the results of referee actions* on vehicles previously tested by such stations, and any other existing investigatory approaches being used by the state.

It is likely that some combination of these two methods would be the best approach to evaluating the relative success of each trigger in identifying problem performers. Regardless of which approach is used, it is strongly recommended that performance feedback be continued on an on-going basis as part of any active triggers system.

###

*This refers to a system used in most decentralized programs in which motorists can bring vehicles in for a subsequent independent inspection at "referee" facilities operated by the state or a state contractor if they are dissatisfied with the results of the inspection performed on the vehicle at a licensed I/M station.

9. REFERENCES

1. Personal communication with Bob Benjaminson, California Bureau of Automotive Repair, April 2000.

2. Personal communication with Tom Fudali, Snap-On Diagnostics, April 2000.

3. Personal communication with Kent Rosner, MCI Worldcom, April 2000.

4. "IM240 & Evap Technical Guidance," Transportation and Regional Programs Division, OTAQ, U.S. Environmental Protection Agency, Report No. EPA420-R-00-007, April 2000.

5. Personal communication with Jay Gordon, Gordon-Darby Inc., May 2000.

6. Confidential survey of State I/M programs by Sierra Research, July 2001.

7. Personal communication from Rick Wilbur, EASE Diagnostics, March 2001.

8. State-confidential information obtained by Sierra Research, 1999-2001.

9. I/M Program Requirements Final Rule, 40 CFR 51.363(a)(3) and (c).

10. Personal communications from Mary Parker, Alaska Department of Environmental Conservation, August 2000.

11. Personal communication with Bob Benjaminson, California Bureau of Automotive Repair, April 2000.

12. "Acceleration Simulation Mode Test Procedures, Emission Standards, Quality Control Requirements, and Equipment Specifications," *Technical Guidance*, U.S. Environmental Protection Agency, EPA-AA-RSPD-IM-96-2, July 1996.

13. Peters, T.A., L. Sherwood and B. Carhart, "Evaluation of the Drift of Vehicle Inspection/Maintenance Emission Analyzers in Use -- A California Case Study," Society of Automotive Engineers, Paper No. SAE891119, presented at the 1989 SAE Government Affairs Meeting, Washington D.C., May 1989.

14. Personal communication with Martin Kelly, City Technologies, Inc., August 2001.

15. Personal communication with Carl Ensfield, Sensors, Inc., August 2001.

16. I/M Program Requirements Final Rule, 40 CFR 51.363(b).

17. Personal communication from Rick Wilbur, EASE Diagnostics, June 2001.