United States
Environmental Protection
Agency

Air and Radiation

EPA420-P-01-003
August 2001

Draft Guidance on Use of In-Program Data for Evaluation of I/M Program Performance

EPA420-P-01-003
August 2001

Draft Guidance on Use of In-Program Data for Evaluation of I/M Program Performance

Certification and Compliance Division
Office of Transportation and Air Quality
U.S. Environmental Protection Agency

Technical Guidance
Jim Lindner
(352)332-5444
lindner.jim@epa.gov

NOTICE

This technical report does not necessarily represent final EPA decisions or positions. It is intended to present technical analysis of issues using data thatC are currently available. The purpose in the release of such reports is to facilitate the exchange of technical information and to inform the public of technical developments which may form the basis for a final EPA decision, position, or regulatory action.

1. Introduction

This document is intended to provide guidance for performing I/M program evaluations using operating program data. The next section is a background of EPA regulation of state I/M programs and a history of methods used to evaluate these programs*. Section 3 describes general approaches to I/M program evaluation. Section 4 focuses on Process-Based measurements and how they relate to I/M program effectiveness and evaluation studies, while Section 5 deals with Results-Based program evaluation analyses.

Equipment specifications, Quality Control and Quality Assurance procedures, test procedures, vehicle pre-conditioning and other details specific to performing emission measurements in a centralized or decentralized network can be found in the EPA guidance documents "IM240 and Evap Technical Guidance[1]" and "ASM Technical Guidance[2]. The importance of proper vehicle pre-conditioning should not be overlooked and both of the guidance documents cited provide information on this topic. It should be noted that is pre-conditioning is not addressed, it is likely that the estimation of program benefits will be underestimated as the resulting emissions measurements will be higher.

It is strongly recommended that any state considering the use of in-program data for program evaluation purposes work closely with their respective regional EPA office and the Office of Transportation and Air Quality (OTAQ) to ensure the most up-to-date practices are incorporated into the evaluation. Methods other than those outlined in this guidance document may be acceptable; however, close coordination with the appropriate EPA regional office and OTAQ will be even more critical if a state intends to develop program evaluation protocols and analyses not discussed in this document.

It should also be recognized given the difficulties associated with I/M program evaluations, that an evaluation based on <u>both</u> out-of-program data (e.g. RSD or roadside pullover) and in-program data will provide a more accurate estimate of overall program performance than simply relying on one method alone. For instance, at this time there is no proposed method of estimating the air quality benefit of pre-test repair using in-program data; however, analyses of RSD may provide information on this important element of an I/M program.

2. Background History of I/M

The Environmental Protection Agency (EPA) has had oversight and policy development responsibility for vehicle inspection and maintenance (I/M) programs since the passage of the Clean Air Act (CAA) in 1970[3], which included I/M as an option for improving air quality. The first I/M program was implemented in New Jersey in 1974 and consisted of an annual idle test of 1968 and newer light-duty gasoline-powered vehicles conducted at a centralized facility. No tampering checks were performed and no repair waivers were allowed.

---

* This section is identical to Section 2 of "Guidance on Use of Remote Sensing for Evaluation of I/M Program Performance July 2001 DRAFT". It is included in this document because it provides a short history of I/M program development that many may find useful.

I/M was first mandated for areas with long term air quality problems beginning with the Clean Air Act Amendments of 1977[4]. EPA issued its first guidance for such programs in 1978[5]; this guidance addressed State Implementation Plan (SIP) elements such as minimum emission reduction requirements, administrative requirements, and implementation schedules. This original I/M guidance was quite broad and difficult to enforce, given EPA's lack of legal authority to establish minimum, Federal, I/M implementation. This lack of regulatory authority -- and the state-to-state inconsistency with regard to I/M program design that resulted from it -- was cited in audits of EPA's oversight of the I/M requirement conducted by both the Agency's own Inspector General, as well as the General Accounting Office.

In response to the above-cited deficiencies, the 1990 Amendments to the Clean Air Act (CAAA)[6] were much more prescriptive with regard to I/M requirements while also expanding I/M's role as an attainment strategy. The CAAA required EPA to develop Federally enforceable guidance for two levels of I/M program: "basic" I/M for areas designated as moderate non-attainment, and "enhanced " I/M for serious and worse non-attainment areas, as well as for areas within an Ozone Transport Region (OTR), regardless of attainment status. This guidance was to include minimum performance standards for basic and enhanced I/M programs and was also to address a range of program implementation issues, such as network design, test procedures, oversight and enforcement requirements, waivers, funding, etc. The CAAA further mandated that enhanced I/M programs were to be: annual (unless biennial was proven to be equally effective), centralized (unless decentralized was shown to be equally effective), and enforced through registration denial (unless a pre-existing enforcement mechanism was shown to be <u>more</u> effective).

In response to the CAAA, EPA published its I/M rule on November 5, 1992[7], which established the minimum procedural and administrative requirements to be met by basic and enhanced I/M programs. This rule also included a performance standard for basic I/M based upon the original New Jersey I/M program and a separate performance standard for enhanced I/M, based on the following program elements:

- Centralized, annual testing of MY 1968 and newer light-duty vehicles (LDVs) and light-duty trucks (LDTs) rated up to 8,500 pounds GVWR.

- Tailpipe test: MY1968-1980 - idle; MY1981-1985 - two-speed idle; MY1986 and newer - IM240.

- Evaporative system test: MY1983 and newer - pressure; MY1986 and newer - purge test.

- Visual inspection: MY1984 and newer - catalyst and fuel inlet restrictor.

Note that the phrase “performance standard” used above was initially used in the CAA and is misleading in that it more accurately describes program design. Adhering to the “performance standard” does not guarantee an I/M program will meet a specific level of emissions reductions. Therefore, the performance standard is not what is required to be implemented, it is the bar against which a program is to be compared.

At the time the I/M rule was published in 1992, the enhanced I/M performance standard was projected to achieve a 28% reduction in volatile organic compounds (VOCs), a 31% reduction in carbon monoxide (CO), and a 9% reduction in oxides of nitrogen (NOx) by the year 2000 from a No-I/M fleet as projected by the MOBILE model. The basic I/M performance standard, in turn, was projected to yield a 5% reduction in VOCs and 16% reduction in CO. These projections were made based upon computer simulations run using 1992 national default assumptions for vehicle age distributions, mileage accumulation, fuel composition, etc., and were performed using the most current emission factor model then available for mobile sources, MOBILE 4.1. That version of the MOBILE model was the first to include a roughly 50% credit discount for decentralized I/M programs, based upon EPA's experience with the high degree of improper testing found in such programs. This discount was incorporated into the 1992 rule, and served to address the CAAA's implicit requirement that EPA distinguish between the relative effectiveness of centralized versus decentralized programs.

The CAAA also required that enhanced I/M programs include the use of on-road testing and that they conduct evaluations of program effectiveness biennially (though no explicit connection was made between these two requirements). In establishing guidelines for the program evaluation requirement, the 1992 I/M rule specified that enhanced I/M programs were to perform separate, state-administered or observed IM240's on a random sample of 0.1% of the subject fleet in support of the biennial evaluation. Unfortunately, the program evaluation procedure for analyzing the 0.1% sample was never developed with sufficient detail to actually be used by the states. In defining the on-road testing requirement, the 1992 rule required that an additional 0.5% of the fleet be tested using either remote sensing devices (RSD) or road-side pullovers. Furthermore, the role that this additional testing was to play -- i.e., whether it was to be used to achieve emission reductions over and above those ordinarily achieved by the program, or whether it could be used to aid in program evaluation -- was never adequately addressed.

At the time the 1992 I/M rule was being promulgated, EPA was criticized for not considering alternatives to the IM240. California in particular argued in favor of the Acceleration Simulation Mode (ASM) test, a steady-state, dynamometer-based test developed by California, Sierra Research, and Southwest Research Institute. In fact, this test had been considered by EPA while the I/M rule was under development, but the combination of IM240, purge, and pressure testing was deemed sufficiently superior to the ASM that EPA dismissed ASM as a credible option for enhanced I/M programs. Nevertheless, EPA continued to evaluate the ASM test in conjunction with the State of California and by early 1995, sufficient data had been generated to support EPA's recognizing ASM as an acceptable program element for meeting the enhanced performance standard (even though the ASM itself was still deemed marginally inferior to the IM240, in terms of its emission reduction potential).

In early 1995, when the ASM test was first deemed an acceptable alternative to IM240, the presumptive, 50% discount for decentralized programs was still in place. Even at that time, however, the practical importance of the discount was waning, in large part due to program flexibilities introduced by EPA aimed at allowing enhanced I/M areas to use their preferred decentralized program designs. This flexibility was created by replacing the single, enhanced I/M performance standard with a total of three enhanced performance standards:

* High Enhanced: Essentially the same as the enhanced I/M performance standard originally promulgated in 1992.

* Low Enhanced: Essentially the basic I/M performance standard, but with light trucks and visual inspections added. This standard was intended to apply to those areas that could meet their other clean air requirements (i.e., 15%, post-1996 ROP, attainment) without needing all the emission reduction credit generated by a high enhanced I/M program.

* OTR Low Enhanced: Sub-basic. Intended to provide relief to those areas located inside the OTR which -- if located anywhere else in the country -- *would not have to do I/M <u>at</u> <u>all</u>*.

Despite the additional flexibility afforded enhanced I/M areas by the new standards outlined above, in November 1995 Congress passed and the President signed the National Highway Systems Designation Act (NHSDA)[8] which included a provision that allowed decentralized I/M programs to claim 100% of the State Implementation Plan (SIP) credit that would be allowed for an otherwise comparable centralized I/M program. These credit claims were to be based upon a "good faith estimate" of program effectiveness, and were to be substantiated with actual program data 18 months after approval. The evaluation methodology to be used for this 18-month demonstration was developed by the Environmental Counsel of States (ECOS), though the criteria used are primarily qualitative, as opposed to quantitative. As a result, the ECOS criteria developed for the 18-month NHSDA evaluations were not deemed an adequate replacement for the CAAA and I/M rule required biennial program effectiveness evaluation.

In January 1998, EPA revised the I/M rule's original provisions for program evaluation by removing the requirement that the evaluation be based on IM240 or some equivalent, mass-emission transient test (METT) and replacing this with the more flexible requirement that the program evaluation methodology simply be "sound"[9]. In October 1998, EPA published a guidance memorandum that outlined what the Agency considered to be acceptable, "sound," alternative program evaluation methods[10]. All the methods approved in the October 1998 guidance were based on tailpipe testing and required comparison to Arizona's enhanced I/M program as a benchmark using a methodology developed by Sierra Research under contract to EPA. Even though EPA recognized that an RSD-based program evaluation method may be possible, a court-ordered deadline of October 30, 1998 for release of the guidance prevented EPA from approving an RSD-based approach at that time.

The focus of this document is to provide methods states may use to estimate I/M program benefits using program data. A separate guidance document is devoted to program evaluations using RSD. As its operating premise, EPA recognizes that every program evaluation method will have its limitations, regardless of whether it is based upon an RSD approach or more traditional, tailpipe-based measurements. Therefore, no particular program evaluation methodology is viewed as a "golden standard." Ideally, each evaluation method would yield similar conclusions regarding program effectiveness, provided they were performed correctly. Unfortunately, it is unlikely we will see such agreement among methods in actual practice, due to the likelihood that different evaluation procedures will be biased toward different segments of the in-use fleet. Therefore, it is conceivable that the most accurate assessment of I/M program effectiveness will result from evaluations which combine multiple program evaluation methods.

3. General Approaches to I/M Program Evaluation

3.1 Defining Program Evaluation

Aside from the technical challenges involved in gathering I/M program evaluation data, there are also subtleties regarding what data is necessary that must be understood. The evaluation of Basic I/M programs is strictly qualitative as per standard SIP policy protocols used to evaluate stationary source emission reductions. Historically, these type of qualitative evaluations have included verification of such parameters as waiver rates, compliance rates, and quality assurance/ quality control procedures, but they have not involved quantitative estimates of emission reductions using in-program or out-of-program data.

The evaluation of Enhanced I/M programs is not as clearly defined and is left to the discretion of the Regional EPA based on the data available. In some instances, it may be possible to estimate the cumulative emission reductions, that is the current fleet emissions are compared to what that same fleet's emissions would be if no I/M program were in existence. However, directly measuring the fleet's emissions to determine the No-I/M baseline is not possible in an area that has implemented an I/M program. Therefore, in order to determine quantitatively whether the level of SIP credit being claimed is being achieved in practice, it becomes necessary to rely on modeling projections to estimate the No-I/M fleet emissions or measure the emissions of a surrogate fleet that is representative of the I/M fleet. Obtaining emission estimates from a No-I/M test fleet based on in-program data would obviously require a traditional tailpipe test be performed on a fleet of No-I/M vehicles; however, it is recognized that this may not be possible to do in all cases due to time, resource or operational constraints.

Two other analyses are also possible that can provide useful information regarding program performance. The first method may be thought of as "one-cycle" since it compares the current I/M fleet emissions to the same I/M fleet's emissions from a previous year or cycle. An analysis such as this would yield information with regard to how the program is improving or declining from year to year. The other method should be considered "incremental" in that it compares the current I/M fleet's emissions to that same fleet's emissions while being subjected to a different I/M program, for instance, comparing a fleet's emissions in an area that has just implemented an IM240 program to that same fleet's emissions the previous year when a Basic Program was in operation. It should be noted, that there is a window of opportunity prior to and during the start-up of any I/M program, or program change, to actually analyze the fleet emissions that would provide empirical data on the No-I/M fleet emissions. If resources and time permit, it is recommended that these baseline data be analyzed in order to reduce I/M program evaluation dependency on modeling projections and provide the most accurate measure of I/M program performance.

3.2 Process vs. Results Based Analysis

Analysis of I/M program performance can be thought of in two distinct ways: Results-Based or Process-Based. A Results-Based analysis is more commonly used for looking at the performance of I/M programs, including comparisons of emissions reductions, pass/fail/waiver rates, and other uses of the data collected within the program. Out-of-program data may also be used, such as remote sensing or roadside tests to determine the emission levels of vehicles between and independent of regular I/M tests.

In a Process-Based analysis of I/M program effectiveness, each of the major steps in the I/M process is evaluated separately:

- achievement of proper fleet coverage
- performance and documentation of accurate emissions inspections
- documentation of repair operations on failing vehicles

The underlying concept of a Process-Based analysis is that if one step in the process is ineffective, then the I/M program is ineffective. A single ineffective process can become the bottleneck of the entire program. On the other hand, even if all processes in an I/M program are operating as designed, the overall effectiveness is not guaranteed; the program is just more likely to be effective.

For example, greater fleet coverage means more vehicles are receiving tests and possible repairs. Similarly, factors such as the test method used, instrument calibration and operation, choice of cutpoints, absence of inspection station fraud, and the effectiveness of vehicle repairs contribute to the effectiveness of an I/M program. Results-Based analysis may show significant fleet emissions reductions resulting from the program, but if the tests were done with uncalibrated instruments, the repairs last only for a short time, or only a small portion of the fleet is actually being tested, then the I/M program may not be effective.

When Process-Based analysis is used in combination with Results-Based analysis, a much more thorough understanding of the effectiveness of an I/M program may be achieved. If a Results-Based analysis indicates that an I/M program is ineffective, a state can have difficulty in determining the cause. In this situation, a Process-Based analysis can help identify where the loss of program effectiveness occurs.

For Process-Based measures to be used to evaluate an I/M program, some methods or standards for evaluation are needed. Unfortunately, EPA is not in a position to provide these standards as the standards should be based on actual operating data, although EPA may provide broad guidelines and/or standard calculation procedures for performing these Process-Based analyses as needed. Nonetheless, EPA recognizes that in many instances, judging the Process-Based performance of an I/M program may be performed by states operating similar programs exchanging results from their analyzer, dynamometer and OBDII Tester audits, as well as repair data relating to number and type of repair, etc. This sharing of knowledge is occurring informally in many forums such as IM Solutions, Clean Air Conference, monthly status calls between states and routine phone calls and Emails. It is not clear at this time if the IM community would support routinely providing this information to an agreed upon clearing house to facilitate the exchange of this information, or if the program information is felt to be too sensitive to permit its free distribution.

In the following sections, methods are described and examples presented for both Process-Based (Section 4) and Results-Based (Section 5) analyses. Many of the examples presented use actual

I/M program data taken from several regions; however, the locations will be identified simply as State 1, State 2, etc.

4. Process Based Measures of Effectiveness

4.1. Participation Rate

The fleet for an I/M program area may be defined either as the set of vehicles registered in the area, or as the set of vehicles driven in the area. Results from various RSD programs have shown that the two fleets are often quite different. Figure 4-1 is a diagram of a typical mix of vehicles for an I/M Program area. The vehicles driven in the area may be registered in the area, or may originate outside the area. Some of the vehicles registered in the I/M program area may no longer be driven there, if the vehicle owner moves or the vehicle is sold. The set of vehicles that participate in the I/M program may include most of the vehicles that are both registered in the area and are located (driven) there. The greatest emissions reduction benefit would be achieved if the set of vehicles that are driven in the program area all participated in the I/M program. This goal is more difficult to achieve in some areas than others; for example, the Kansas City metropolitan region is partly in the state of Missouri and partly in the state of Kansas, so many of the vehicles driven in Kansas City, Kansas are registered in Kansas City, Missouri, and vice-versa.

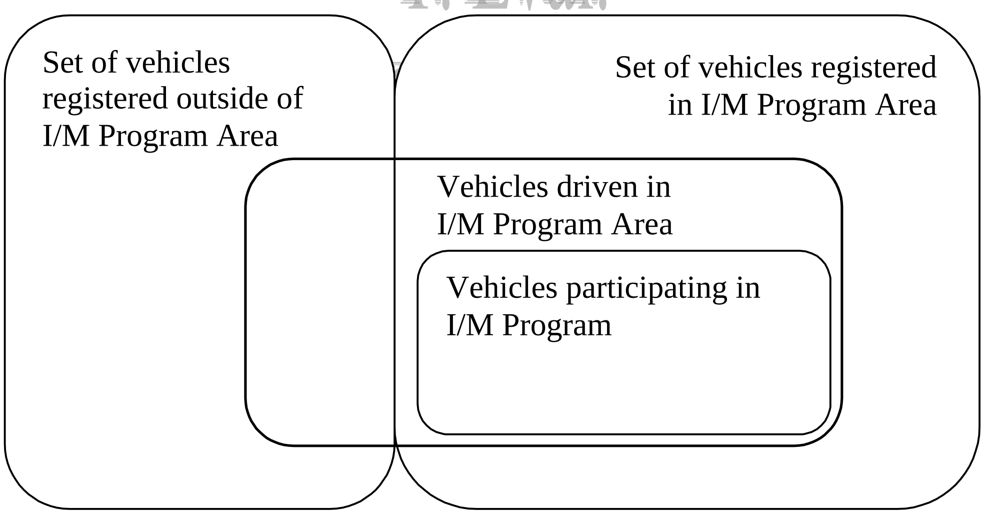


Figure 4-1. Mix of Vehicles within an I/M Program Area

To evaluate the performance of an I/M program, a first basic step is to define the participation rate of vehicles eligible for the program. Even the most carefully administered I/M program may be undermined if a significant portion of the fleet avoids the tests. The goal here is to compare a set of vehicles participating in the I/M program to both the registered fleet and the driven fleet. Emphasis should be placed on comparison to the registered fleet, since location of registration is almost always used to define the program area; however, even greater emissions reductions could be achieved in any area by expanding the program to include all vehicles driven in the area.

The most basic measure of fleet coverage is to compare counts of the number of vehicles in the registered fleet, the driven fleet, and the I/M fleet. Although these rough estimates will contain errors, given the minimal effort required to obtain these estimates, they should be performed and recorded. For example, the registered fleet (usually taken from a state registration database) often includes large numbers of vehicles that have been sold or moved out of area. A registration database that is not consistently updated as vehicles migrate makes the I/M program participation rate appear to be lower than it is, and makes it difficult to identify vehicles that really are located in the area but not participating in the program. License plate readers, such as those used by RSD and pneumatic vehicle counting devices can be used to estimate the driven fleet. However, such readers and counters can have sampling errors depending on the locations for the readers. Because newer vehicles are usually driven more than older vehicles, the RSD data may actually catch more of the “travel fraction” than the “registration fraction” in an area.

The analysis described in this and the following sections is based on data from states’ Vehicle Inspection Databases, registration databases, and repair databases. Datasets may have several million records and require multiple gigabytes of computer memory to process. The EPA contractor (Eastern Research Group) who performed these analyses used a Digital Alpha DS20 Unix system with 100 GB of hard drive space and 1 GB of RAM with SAS statistical analysis software.

4.1.1 Comparing Vehicle Age Distributions

One method of assessing the participation rate is to compare the vehicle age distribution of the registered fleet, the I/M fleet, and the driven fleet. Distributions are used in place of counts due to the large differences in the fleets. In the absence of a fully updated registration database, distributions may still be compared to determine whether the registered and tested fleets are qualitatively the same. This type of comparison is shown in Figure 4-2 using data from State 2. From this figure it may be seen that the set of registered vehicles has a larger proportion of early 1980’s vehicles than does the I/M set, which might indicate that owners of older vehicles are avoiding inspections. The driven fleet that was observed on the roads by RSD contains even fewer vehicles from the oldest model years than the I/M set, indicating that some of the older vehicles that are registered but not participating in the I/M program may not be driven often. The registration fleet has a mean age of 9.4 years, the I/M fleet, 8.2 years, and the RSD fleet, 7.0 years.

4.1.2 Matching Registration Records with I/M Records

Comparisons between the registered fleet and the I/M fleet could be done directly by attempting to match each registration record with an I/M record. However, the registration database may not be updated each time a vehicle is sold outside the area, leading to overstatement of the difference between the two fleets. Figures like 4-2 include the implicit assumption that these sales are evenly distributed over the model years; if this is not the case, then bias may be introduced.

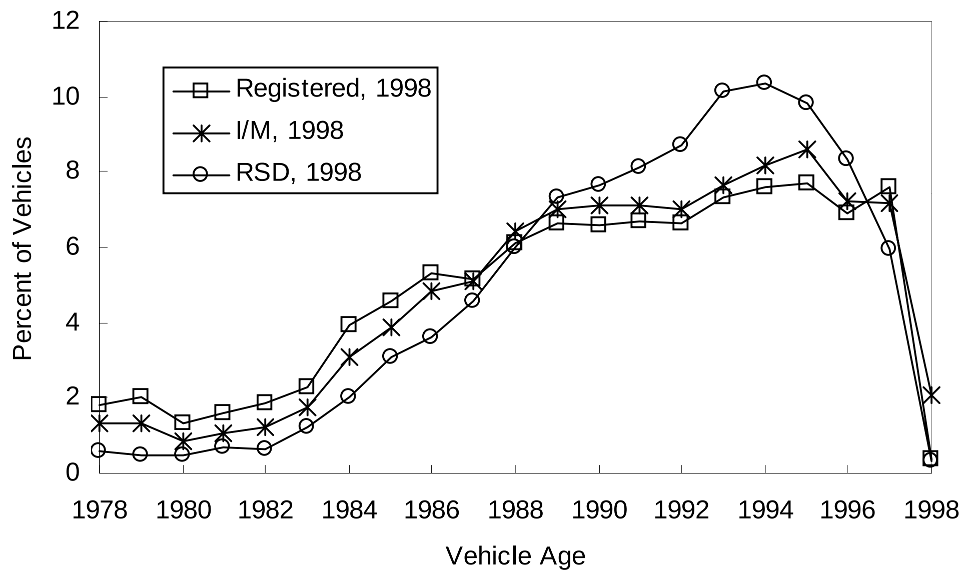


Figure 4-2. Distribution of Vehicles in I/M Program, Registration Database, and Observed through Remote Sensing

4.1.3 Using Year-to-Year Trends

Year-to-year trends in the age distribution of the I/M fleet may also be informative even though there can be many reasons for shifts. For example, if a fleet had a larger portion of new vehicles each year, it might be concluded that an improving economy was helping encourage the replacement of old vehicles with new ones. This doesn't seem to be the case for State 3, shown in Figure 4-3. The average vehicle age increases from 7.3 years in the first program year shown to 8.0 by the fourth program year.

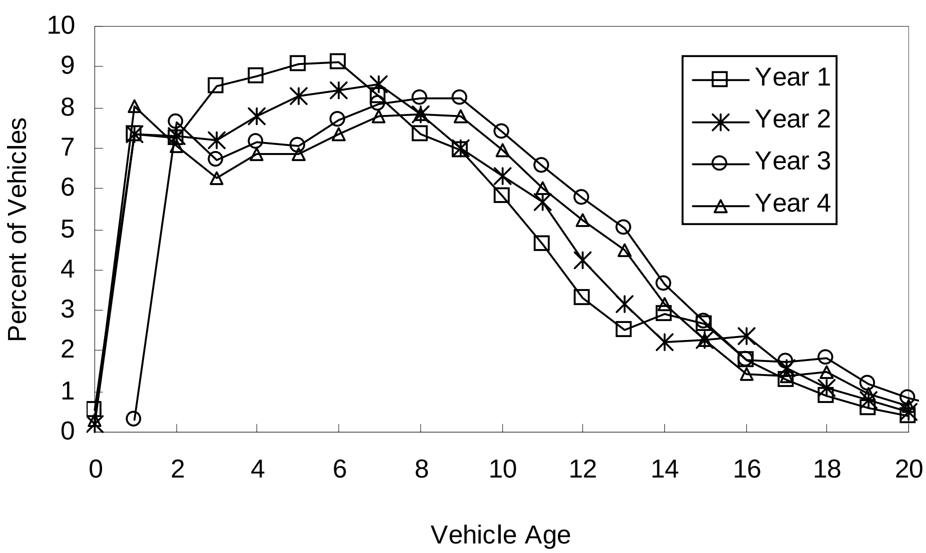


Figure 4-3. Vehicle Age Distribution over Four Years of I/M Tests

4.1.4 Using Multi-Year Trends

Multiple years of I/M program data may also be used to find the rate at which vehicles leave the program between test cycles. Vehicles that leave the program may have been sold and removed from the fleet, or they may remain in the area without participating in the I/M program. For State 3, vehicles were tracked over the four years of data being used. It was found that almost 80% of the vehicles tested each year returned for testing the next year, as shown in Figure 4-4. From the data available, it is not possible to determine whether the other 20% of vehicles were sold outside the program area or simply dropped out. Figure 4-4 shows that the percentage of vehicles returning the next year decreases significantly for vehicles aged 10 years or greater. These vehicles are also the most likely to fail the I/M test, possibly leading the owners to avoid further testing. In Figure 4-5, the percentage of vehicles that return the year following a failed I/M test is presented. Since the return rate is considerably lower than the overall average shown in Figure 4-4, it seems reasonable to conclude that fear of failing the test has led some vehicle's owners to drop out of the program.

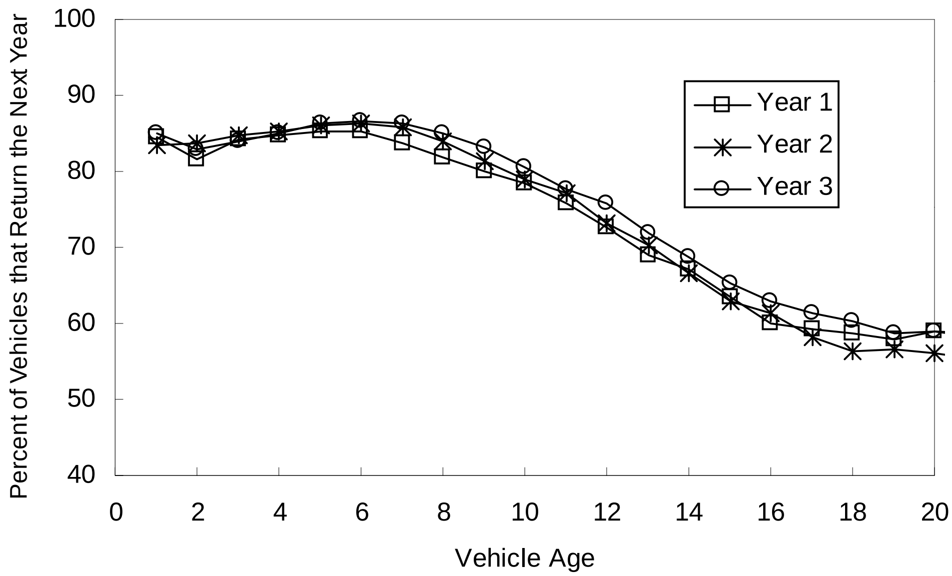


Figure 4-4. Percentage of Tested Vehicles That Return for Testing the Following Year

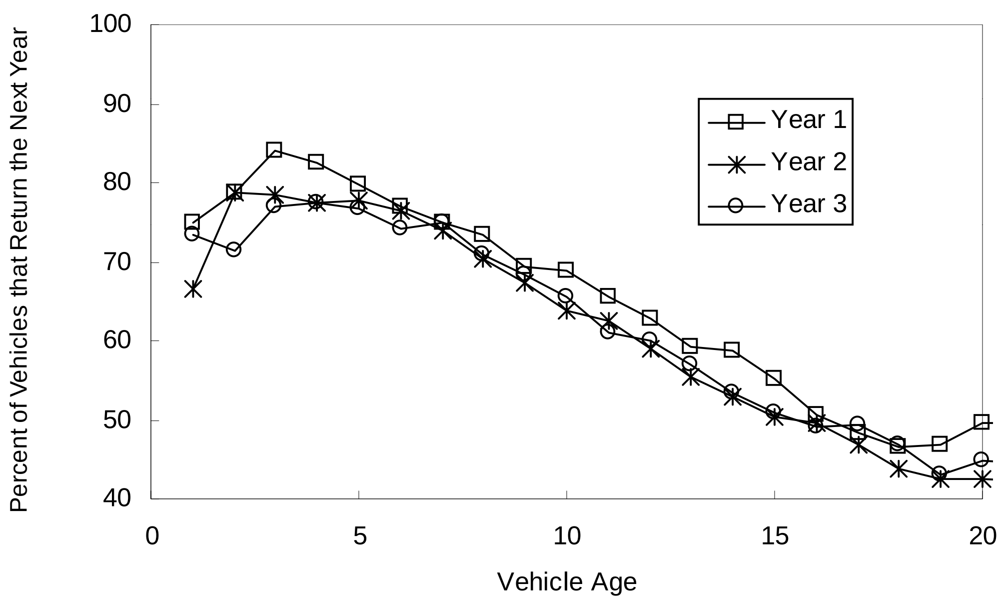


Figure 4-5. Percentage of Failing Vehicles that Return for Testing the Following Year

4.1.5 Parking Lot Sticker Surveys

Data from parking lot sticker surveys have been used by states as a cost-effective method to estimate I/M program compliance rates[11, 12]. Care must be taken to ensure that the surveys capture a representative sample which will require appropriate geographic coverage. Also, procedures must be documented and in place to minimize the opportunity for fraudulent stickers to be obtained by those motorists seeking to avoid the program.

4.1.6 Recommended Best Practice

One of the five methods described above should be used to verify compliance rate estimates used in the SIP, as well as for estimating average emission reductions when used with failure rate and emission data. The primary goal is to diligently update and maintain the accuracy of the vehicle registration database, so that direct comparison between the sets of vehicles registered and participating in the I/M program may be made. License plate reading equipment like that used in RSD studies may be used to confirm the accuracy with which the vehicle registration database represents the fleet. Until a high level of confidence in the accuracy of the registration database is developed, comparisons of distributions such as those shown in Figures 4-2 and 4-3 should be used to qualitatively compare the set of vehicles that undergoes I/M testing to the registration database. Figures like 4-4 and 4-5 should be used to estimate the rate at which vehicles drop out of the I/M program. Parking lot surveys have been used by many states as a cost-effective way to estimate compliance rates also.

4.2. I/M Effectiveness

4.2.1 QA/QC

The effectiveness of the inspection process itself may be influenced by many factors. The inspection is primarily based on the measurement of vehicle emissions. Any factors that degrade the accuracy of the emissions measurement contribute to the degradation of the I/M program. Such factors might include improper analyzer calibrations, analyzers that require maintenance, inaccurate data entry of vehicle information, emissions cutpoints that are too loose or too stringent, emissions tests with excessively large measurement errors, and inspection station fraud.

The following sub-sections provide a discussion and examples of ideas for techniques that can be used to evaluate many of the factors that contribute to ineffective I/M programs. Passing grades on all factors does not necessarily guarantee a successful I/M program. On the other hand, a poor grade on one factor can act as a bottleneck preventing an I/M program from being effective. Beyond merely using these techniques to demonstrate I/M program effectiveness, a state can use these techniques to identify for itself areas of inspection effectiveness that are good and areas where improvements need to be made.

This analysis of in-program I/M data should also be performed prior to any analysis of emissions reductions so that emissions reduction calculations will be based on the data of known quality.

4.2.1.1 Instrument Calibrations

Records of I/M program analyzer calibrations can be used to measure the drift of analyzers between calibrations. If many analyzers in a state's I/M program drift substantially, the results of measurements are suspect. Ideally, all analyzers should drift no more than the specification of the analyzers.

For example, in State 3 analyzers must be calibrated at least every 72 hours. Before calibration, each analyzer is checked for drift by measuring the calibration gas mixture, whose concentration is known within a specified precision. If the analyzer has not drifted since the last calibration, its readings for the calibration gas will be close to the bottle label value, and little calibration adjustment will be necessary. The difference between this pre-calibration analyzer reading and the label concentration in the gas mixture is a direct measure of instrument drift. Analyzers that consistently drift little from calibration to calibration can be expected to produce more accurate measures of vehicle emissions than those that drift greatly.

Six months of instrument pre-calibration data containing 90,781 calibrations from 2,324 instruments was examined. We examined the analyzers' drift characteristics on readings for HC, CO, CO_2, and O_2 for zero, mid-span, and high-span gases. For this example, the CO high-span gas is analyzed, which had a label value of 4.0%. The BAR90 analyzers, which were used in this I/M program, have an accuracy specification of ±0.15 % for a 4% CO gas. Accordingly, it is expected that most of the 90,781 pre-calibrations should fall within about ±0.15 % of 4.00%. Any pre-calibrations that fall greatly outside this range would cause concern.

Figure 4-6 shows a histogram of the 90,781 pre-calibrations for all instruments in the state during this period. About 86% of the values are within ±0.15 % of 4.00%. However, 3.7% of the values are zero, and 0.5% of the values are between 0.1% and 3.5%. These unexpected values raise concern and should be investigated. Several explanations may exist for these unexpected values. In any case, states that have tighter distributions of pre-calibration values and have a system in place for addressing out-of-spec values have a better chance of having an effective I/M program.

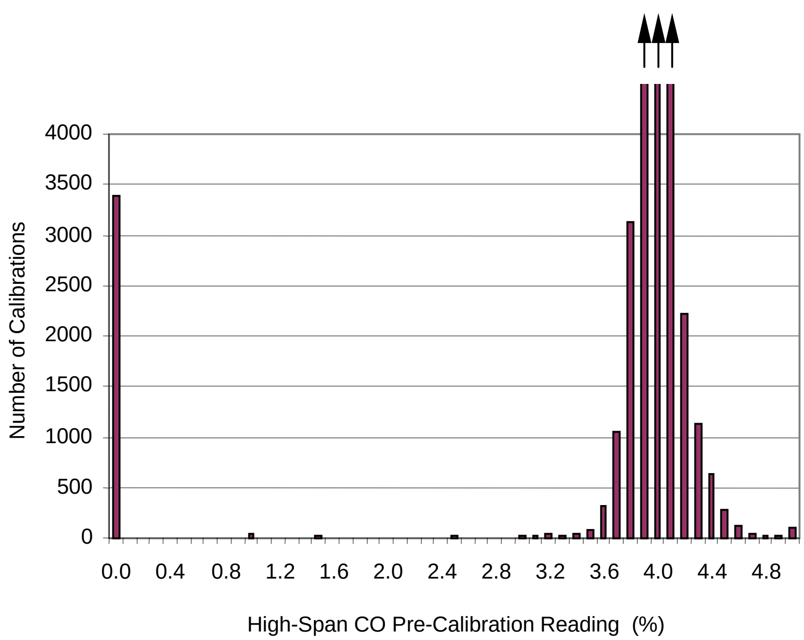


Figure 4-6. Distribution of Values for High-Span CO Pre-Calibrations

Instrument Calibrations Recommended Best Practice
Instrument calibration data, especially the pre-calibration readings, are a good indicator of instrument drift and should be tracked regularly. Instruments that consistently drift more than the instrument specifications should be repaired.

4.2.1.2 Instrument Audits

Independent instrument audits of I/M program emissions analyzers with certified bottled gas can also be used to evaluate analyzer accuracy. This additional instrument check is valuable because instruments can experience periods when they are out of calibration even if the pre-calibration data shows that the instrument has little drift. One possible cause is problems with the line leading from the tailpipe probe to the instrument. Instrument calibrations introduce gas at the instrument; instrument audits and vehicle tests introduce gas at the tailpipe probe. Obstructions, leaks, or contamination might cause audits (and emissions measurements) to be out of calibration.

For example, I/M analyzers were calibrated as normal using a station's normal supply of calibration gas. Nothing abnormal was seen in the calibration data recorded in the VID. The instruments were routinely challenged using a supply of bottled gas separate from the station's calibration gas. Most instruments passed the audits for zero, low-span, and high-span gases, as is shown for CO_2 in Figure 4-7. However, one instrument showed varying behavior from day to day with values biased low by about 30% on several days, as is shown in Figure 4-8.

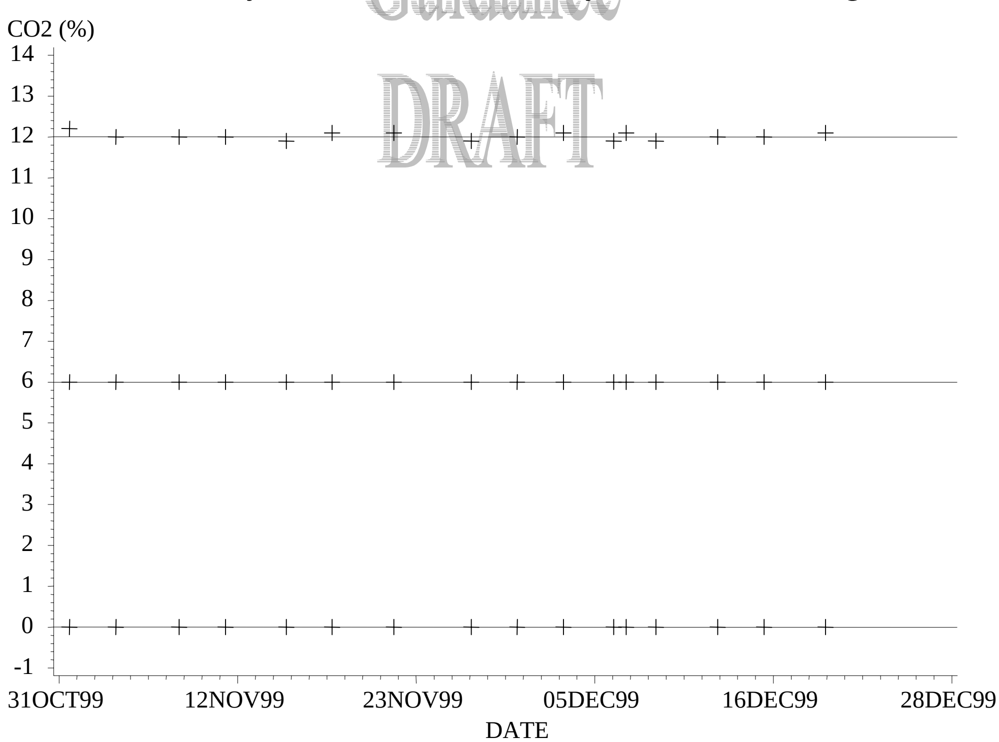


Figure 4-7. Good Analyzer Results for 3 Audit Gases

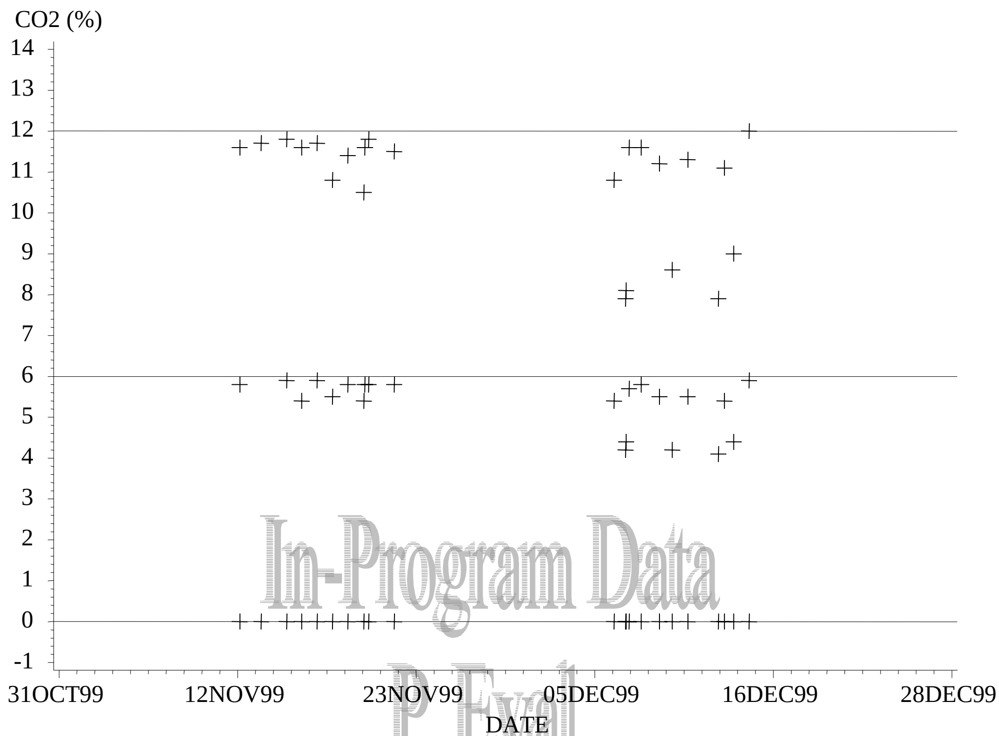


Figure 4-8. Poor Analyzer Results for 3 Audit Gases

These audits indicated a recurring instrument problem that was not caught by the station staff or the VID data. The problem was serious even though the measured quantity (CO_2) is not a pollutant of interest (HC, CO, or NO_x), since CO_2 is used to correct for exhaust dilution; inaccurate corrections would be made with an erroneous CO_2 value. The result would be inaccurate determinations of dilution-corrected HC, CO, and NO_x.

States that have some sort of instrument audit program in their I/M program would potentially be able to identify instruments that are out of calibration by analyzing the data as described above.

Instrument Audits Recommended Best Practice

A standardized method of instrument audit program provides an added level of confidence that instruments are accurate. We have found cases where instruments calibrated well and showed no drift between calibrations, but provided inaccurate results when challenged with a separate source of gas.

4.2.1.3 DCF Check

The measurement of exhaust emissions concentrations can be confounded by the dilution of the exhaust gas by non-optimal probe placement, leaking exhaust systems, cylinder misfires, and excess oxygen from air pumps. Some I/M program emissions analyzers use measured CO and CO_2 concentrations to calculate a dilution correction factor to correct raw exhaust emissions concentration values for this dilution to arrive at emissions values on an undiluted basis.

Assuming stoichiometric combustion of gasoline, an exhaust dilution correction factor (DCF) can be estimated using a carbon mass-balance and the measurements of CO and CO_2. These constituents are measured in the non-dispersive infrared bench of the analyzer. The equations are based on the average composition of gasoline.

First, define the variable x:

$$x = \frac{CO_2}{CO_2 + CO}$$

where the CO_2 and CO values are in percent.

Then the dilution factor, dcf_{CO/CO_2}, is as follows:

$$dcf_{CO/CO_2} = 100\frac{x/(4.64 + 1.88x)}{CO_2}$$

If a fuel other than standard gasoline is used, the 4.64 constant will be different. For example, the constants for methane (CNG), propane (LNG), methanol (M-100), and ethanol (E-100) are 6.64, 5.39, 4.76, and 4.76, respectively. The constants for reformulated gasoline and oxygenated gasoline will depend on gasoline composition, but are generally not far from 4.64.

In addition, many emissions analyzers also measure exhaust gas oxygen concentration with an electrochemical cell. Assuming an ambient air oxygen concentration of 20.9%, the exhaust oxygen measurement can also be used to estimate dilution in the exhaust. A dilution correction factor based on the measured oxygen concentration O_2 is:

$$dcf_{O_2} = \frac{20.9}{20.9 - O_2}$$

This relationship assumes that the tailpipe oxygen concentration for stoichiometric combustion and no air in-leakage is 0.0% O_2. Field measurements indicate that new vehicles with no exhaust system leaks and operating at stoichiometric air/fuel ratio have 0.0% tailpipe oxygen concentrations.

If CO, CO_2, and O_2 are measured correctly, the independent DCFs (CO/CO_2 and O_2) for each vehicle inspection should agree well with each other. Emissions results for two-speed idle tests in State 3 were examined and the DCFs were calculated for each test on each vehicle. Figure 4-9 shows a plot of the high-speed idle DCF based on CO/CO_2 versus the high-speed idle DCF based on O_2 for each emissions test. The plot shows that many of the points fall near the 1:1 line as expected; however, many also fall far off the 1:1 line. Those points that fall off the line represent analyzer sensors for CO, CO_2, or O_2 that are broken or out of calibration, data entry errors, or tests on vehicles that use fuels far different from gasoline. Ideally, all points would fall near the 1:1 line.

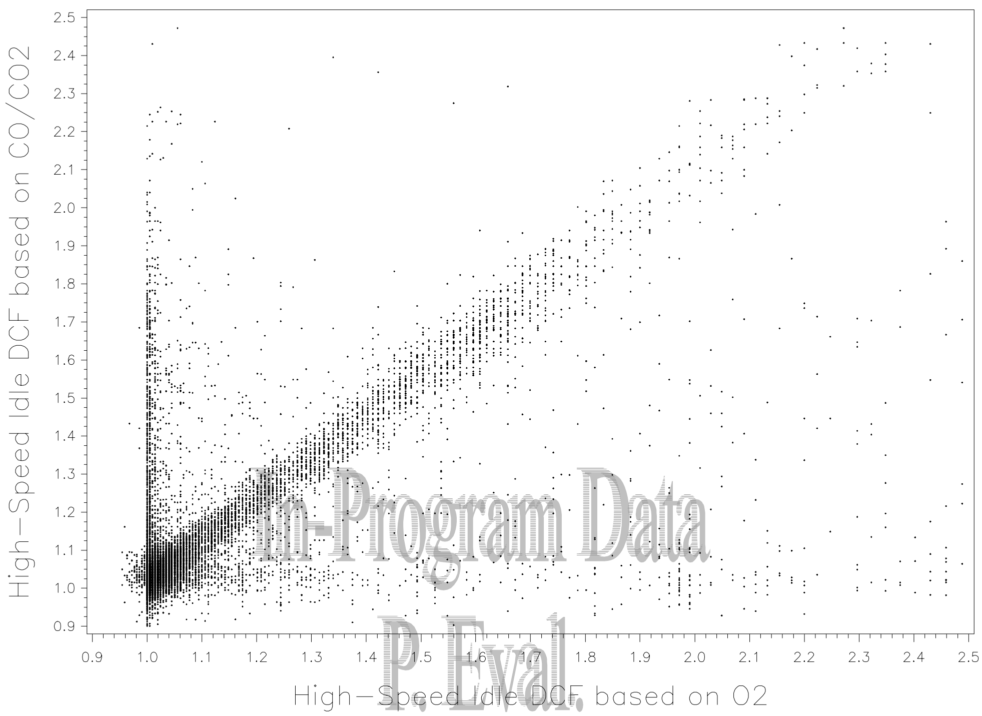


Figure 4-9. Comparison of High-Speed Idle DCFs in State 3.

Each state could use this evaluation of CO, CO_2, and O_2 data for every emissions inspection to demonstrate the fraction of inspections that meet a minimum requirement. Tolerances for agreement between the two types of DCFs can be determined from the I/M analyzer accuracy specifications for CO, CO_2, and O_2 and the local gasoline composition. The plot with this data indicates that the difference between the two DCFs should be no larger than about ±0.14. The dilution correction factor relationships are a consequence of gasoline combustion stoichiometry. Therefore, it also follows that a relatively constant relationship exists among the undiluted exhaust gas concentrations of O_2, CO, and CO_2 from a gasoline-fueled engine even if the engine produces significant concentrations of HC, CO, and NO_x. Analyzer manufacturers could use this relationship to provide a check of each emissions test as it was being performed. If the relationship was not satisfied, the analyzer operator would see a flag to indicate that analyzer maintenance should be performed.

DCF Check Recommended Best Practice

The raw (before any corrections) concentration measurements of all emissions tests should indicate that combustion of gasoline (or of whatever fuel is used) is the source of emissions. One way to check this is to compare calculated dilution correction factors based on CO/CO_2 against those based on O_2. For every emissions test they should agree within about ±0.14. If they do not, the emissions test may be inaccurate. DCF checks can be made on records in the VID, but it may be best to incorporate them in the analyzers so that inspection stations can address the problem immediately.

4.2.1.4 Inspection-Repair Sequence

An analysis of vehicle inspection/repair records from the VID can be used to evaluate the accuracy and completeness of data in the VID system. States that have better VID systems have more reliable inspection and repair data and therefore can better support their claims of effective I/M programs. The following is an example of an analysis of inspection/repair sequences for State 3. In this state no repair records were kept, and so the example cannot make use of repair information. One of the key items of Section 4.3 is the strong recommendation for states to maintain good records for vehicle repairs performed as part of an I/M program.

Each vehicle was tested at an I/M station on one or more occasions. On each occasion, the VID contains a variable that gives the type of test (Initial or Re-test) and a variable that gives the result of the emissions test (Pass or Fail). The Test Type variable has special rules for designating whether a test is an Initial or a Re-test. For Initial tests, customers are charged for the inspection. If they fail the inspection but return after repairs within 5 days, then the second test is designated a Re-test, and the customer is not charged for the Re-test. If more than 5 days have elapsed, then the second test is designated an Initial test and the customer is charged again. Consequently, a test that is designated Initial may actually be a follow-up test in an effort to get the vehicle to meet I/M requirements. In any case, four combinations of these two variables are possible for each occasion. For analysis purposes, the four combinations were given designators as shown in Table 4-1.

Table 4-1. Designators for Test Type and Result

Designator	**Test Type**	**Emission Test Result**
IF	Initial	Fail
IP	Initial	Pass
RF	Re-test	Fail
RP	Re-test	Pass

Then, for each unique VIN, the designators were concatenated in chronological order to create a sequence number that describes the testing sequence that each vehicle experienced during I/M testing. For example, for a vehicle that initially failed and then passed on a re-test, the test sequence would be IF, RP. The frequency distribution of the resulting test sequences is shown in Table 4-2.

The distribution shows that the top ten most frequently found sequences accounted for 99.64% of the vehicles tested. Although it is recognized that some of the vehicles may have incomplete test cycles because the test cycle was begun in the last few days of the data set period, some of these sequences raise questions. Why are 1.37% of the vehicles tested a second time after they pass? Why do 0.82% of the vehicles undergo no further testing when they failed initially? An important part of an analysis of inspection/repair sequences is to document the explanation for these apparent anomalies.

Table 4-2. Frequency Distribution of Test Sequences in State 3

Test Sequence	**Vehicle Frequency**	**% of Vehicles**

IP	3,413,802	94.24
IF, RP	66,987	1.85
IP, IP	49,771	1.37
IF	29,682	0.82
RP	21,509	0.59
IF, IP	9,183	0.25
IF, RF, RP	7,037	0.19
IF, RF	4,790	0.13
IP, RP	4,365	0.12
IF, RF, IP	2,192	0.06
450 Other Test Sequences	13,093	0.36
Total	3,622,411	100.00

Approximately, 450 less frequently used sequences accounted for the remaining 0.36% of the tested fleet. Many of these remaining sequences seem to be unlikely. For example, what could be the reason for 21 vehicles having the sequence IP, IP, IP, IP, IP, IP, IP? It is suspected that these sequences represent database data entry problems instead of real situations. Better inspection database systems should be able to reduce the occurrence of these unlikely test sequences.

Inspection Repair Sequence Recommended Best Practices

When a good inspection data set is combined with a good repair data set, the sequences of inspections and repairs should make sense. Cross-checking between these data sets can identify many errors in VID data sets. The sequence of each vehicle should tell a simple story. If it does not, data entry problems probably exist.

4.2.1.5 VID Check

Since the in-program data is the primary basis of the I/M program evaluation, a series of basic data checks should be used to demonstrate the accuracy and completeness of the data in the database. The following list may serve as a starting point for basic validation checks in future I/M program evaluations.

1) The beginning and ending dates of the VID data under consideration should be specified.

2) A frequency distribution of almost all database variables should be provided to demonstrate the accuracy and completeness of data entry. Missing and nonsensical values should be included in the distribution to show the frequency of improper entry.

3) A distribution of the emissions measurements is a special case of the above. Ideally, no observations with missing values should be present. Also, all observations should have a CO_2 concentration between about 6% and 17%, since a combustion process must be present.

4) The fraction of observations with both the license plate and the VIN missing should be determined.

5) The validity of each VIN should be checked in some manner. In the simplest method, the check digit in 1981+ VINs can be checked. More extensive VIN checking efforts could involve comparison of the recorded vehicle description information with the corresponding information from a VIN decoder.

6) Each license plate should be associated with only a single VIN.

7) Within a single I/M cycle, each vehicle should have a recognizable and reasonable test and repair sequence. For example, a vehicle with a “fail, repair, fail, repair, pass” sequence is reasonable, but one with a “fail, repair, pass, pass, pass, repair, fail, fail” sequence is not. Data entry problems by test stations and repair stations can produce unreasonable sequences. Accordingly, a frequency distribution of sequences can be an indicator of the extent of data entry problems.

VID Check Recommended Best Practices

These checks are probably the most fundamental VID data checks. They involve sanity checks on every field in the VID. Distributions of numeric variables, frequency distributions of categorical fields, x*y plots, and range checks can all be used to find how data is improperly entered in the database.

4.2.2 Test Data

A discussion of the effectiveness of emissions inspections is necessary to evaluate their contribution to the overall I/M program of a state. If a state’s I/M program covers the fleet well and has great repair stations, but the emissions inspection stations cannot properly identify high emitting vehicles, the overall effectiveness of the I/M program will suffer.

Perhaps the most fundamental part of the discussion of emissions measurement is a definition of the inspection flow sequence. The inspection sequence would first define the vehicles that are subject to I/M testing. For example, this might be 1975 to 1995 light-duty, gasoline-fueled vehicles. Then, perhaps all-wheel-drive vehicles get a two-speed idle test, and all remaining vehicles get an ASM test. All of the steps in the inspection flow would be defined, including station type (e.g. test only, test and repair, centralized, decentralized), test type (e.g. IM240, ASM, gas cap check) and associated cutpoints, model year group selections, waiver thresholds, and exemption criteria. This inspection sequence should be presented as a flow diagram.

Next, the flow diagram should be annotated to show the number of vehicles and inspections that occurred in the state for the evaluation period. This would allow a between-state comparison to be made of corresponding parts of the emission inspection sequence. For example, one state might have a waiver threshold of $200 with 2% of vehicles waived, while another state has a waiver threshold of $500 with only 0.3% of vehicles waived.

Next, the important characteristics of the emissions tests used should be defined. This would include emissions test type and emissions pass/fail criteria (i.e. cutpoints).

Correlations can be built to use short emissions test results (e.g. ASM, two-speed idle, IM240) to predict reference emissions test results (e.g. IM240, FTP). The importance of vehicle pre-conditioning in any correlation study or program evaluation effort must not be overlooked as inconsistent pre-conditioning will have an adverse impact on the test program. The IM240 test can be the reference test, or it can be a short test when the FTP is the reference test. Studies that apply these correlations indicate that the greatest source of error for a vehicle receiving an incorrect pass/fail designation by the short test is the difference in the responses of vehicles to the short and reference tests[13, 14]. These studies indicate that measurement errors of the short test and of the reference test are small contributors to incorrect pass/fail designations. Therefore, states should report the variance of the deviations between their short test (if they use one) and a reference test. A state could measure this variance by performing out-of-program reference tests on a sample of program-eligible vehicles. Alternatively, a state could simply quote the variance measured by other states. However, states that can demonstrate a smaller variance will tend to have the better inspection effectiveness.

4.2.2.1 Measurement Error

The measurement error of an emissions test is an estimate of the uncertainty in the reported emissions of a single measurement. Tests that have large measurement errors will cause the pass/fail status of some vehicles to be improperly designated; however, studies have shown that such tests can still provide emission reduction benefits for the fleet as a whole (14 above). For each emissions test type, the measurement error (as determined by replicate testing of vehicles) should be reported. States may choose to report measurement error calculated from data taken in other states, or they may choose to calculate measurement error based on their own data of repeat emission measurements.

This measurement error for an emissions test can be calculated from repeat emissions measurements on a sample of vehicles. A state could obtain repeat measurements by performing them on vehicles that are being inspected as part of the normal I/M program. The vehicles that receive repeat measurements should be selected to cover the range of emissions levels represented in the fleet. In general, a stratified sampling technique will provide the most useful information from the fewest measurements. The measurement error is calculated by pooling the variance of each repeated vehicle's measurements. However, the variance for each vehicle must be calculated after transforming all emissions measurements to a space where measurement error is relatively constant for all emission levels. We have found that the natural log transformation provides this attribute for most emissions tests. An example of the calculation of measure error is provided in Reference 13 above and is briefly outlined in Appendix A.

4.2.2.2 Cutpoints

The cutpoints applied to emissions measurements to designate a vehicle as a pass or fail also have an important influence on the correctness of the designation and thereby on the overall measurement effectiveness. An analysis of cutpoint effectiveness could be performed on in-program data. States should already have an understanding of the role that cutpoint selection plays in identifying vehicles that need repair versus vehicles that are sent to repair. The following conceptual discussion is meant to reinforce that understanding, and it will lead to suggestions for evaluating and optimizing cutpoint selection. With regard to optimizing cutpoints there are those who believe there should be methods to get information on the emissions and repair rates of vehicles <u>below</u> current cutpoints. The rational for this approach is that without this information, state I/M program administrators would only be able to look to higher cutpoints to search for an optimum.

Figure 4-10 qualitatively shows the emissions distributions of vehicles in a state's I/M program fleet subject to a common cutpoint. All vehicles that have a properly functioning emission control systems are in the lower emitting distribution (shown by the thin line); these vehicles are non-repairable since they have no problems. All vehicles that have problems with their emission control systems are in the higher emitting distribution (shown by the thick line); these vehicles could be repaired if the I/M program could identify them. The two distributions have a significant overlap in emissions. This overlap is a consequence of the emissions characteristics of specific non-repairable and repairable vehicles. For vehicles of the same age and technology, some broken vehicles will have emissions lower than some properly operating vehicles.

Wherever the cutpoint is chosen (shown by the dashed vertical line in the figure), some vehicles will be properly designated and some vehicles will be improperly designated as pass or fail. Improper designations include two types: non-repairable vehicles called a fail, and repairable vehicles called a pass. Where should the state set its cutpoint? If a state sets a high (loose) emissions cutpoint, most failures will be repairable, few failures will be non-repairable, but only a small fraction of all repairable vehicles will be sent for repairs. The state's airshed incurs an environmental cost from these false passes. If a state sets a low (stringent) emissions cutpoint, a larger fraction of all repairable vehicles will be sent for repairs, but many non-repairable vehicles will also be sent for repairs. In this case, vehicle owners incur an expense for taking their vehicle to get a repair for a problem that does not exist.

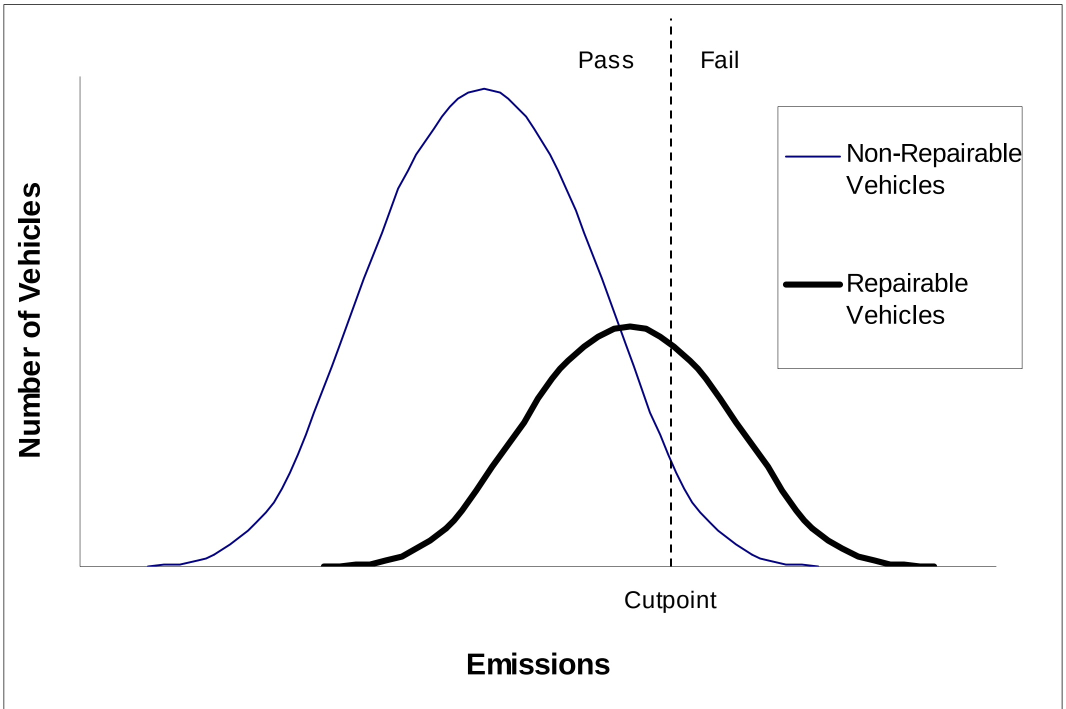


Figure 4-10. Conceptual Emissions Distributions of Repairable and Non-Repairable Vehicles

With estimates of the cost of false passes, the cost of false fails, and the distributions of repairable and non-repairable vehicles (as shown conceptually in Figure 4-10), a state could optimize the selection of the cutpoint by minimizing the total cost, which is the sum of false pass and false fail costs.

Although the actual accounting in this situation will likely be difficult, a state could estimate the costs of false passes and false fails by an economic analysis. False pass costs would be driven by the influences of excess emissions released and would include health costs, cost of further stationary source limits, and costs of not achieving SIP goals. False fail costs would be driven by inconvenience costs including time and repair costs lost by owners taking vehicles to repair shops for problems that do not exist. Excess emissions identified estimates would have to be obtained from paired testing using a state's I/M test and a suitable reference test, and in many instances, such as with idle, IM240 and ASM tests, data sets exist that could aid in this effort.

For a state to optimize the location of the cutpoint, knowledge of the shape of the non-repairable and repairable distributions at emissions above and below the cutpoint is also required. An analysis of the distributions above the current cutpoint should be performed first since the I/M program will already have the data. The parts of the distributions for emissions above the cutpoint value can be determined by an analysis of in-program emissions and repair data. The state should analyze these distribution shapes and report them; they are an indication of the ability of the emissions measurements to resolve (i.e. separate) the repairable from the non-repairable vehicles. Some emissions tests may be better able to resolve repairable and non-repairable vehicles than other tests. Later, the discovery of repairable and non-repairable distributions below the current cutpoint using typical in-program I/M data could be made, but it is more difficult. There are two potential problems: fast-pass and fast-fail emissions measurements and the unknown repair needs of vehicles with emissions below the cutpoint. Without accurate emissions and repair data for at least a sample of vehicles with emissions below the current cutpoint, the search for a more cost-effective cutpoint below the current cutpoint cannot be made with in-program data.

The use of fast-pass and fast-fail emissions measurements increases throughput at I/M stations but impedes determination of the emissions distributions. Whenever an emissions test is cut short by invoking fast-pass or fast-fail criteria, the emissions level of the full test is obviously lost. In some I/M programs, whether a test result is from a fast test or a full test may not be recorded. Use of fast-pass algorithms contaminates emissions measurements below the cutpoint; fast-fails contaminate measurements above the cutpoint. If in-program data is to be used for optimizing cutpoints, fast-fail algorithms should be used only above some high emissions value, where cutpoints would never be considered, and fast-pass algorithms should be used only if the instantaneous emissions measurement of a vehicle is at a fraction (e.g. 50%) of the standard cutpoint value. This would allow an analysis of the full cycle emissions data for all inspections in a window, for example, between 50% of the cutpoint and 400% of the cutpoint.

The second area of information required to optimize the cutpoint is the distribution of the repairable and non-repairable vehicles below the current cutpoint; however, there is no unobtrusive, cost-effective method to obtain such data. Normally, no vehicles that are designated pass are sent to repair, and therefore, the fractions that are repairable and non-repairable are not known. Therefore, the only way to find these fractions is to try to repair or to diagnose a sample of the passing vehicles. This could be done with a random sample of the fleet that passed by offering the vehicle owner an incentive to participate. The cost of the incentive would be paid for by the increased cost-effectiveness of the I/M program after cutpoints are adjusted. Given the anticipated difficulties of such a study, it may be best left for a joint study between EPA and interested states to perform a pilot study that would provide insight into this question. But it does seem clear that states that had access to such cutpoint optimization procedures would tend to have better I/M programs than states that did not, and their I/M programs would benefit from the optimization.

4.2.2.3 Recommended Best Practices

A state should provide a process flow diagram of the flow of vehicles through its I/M program. The diagram should show vehicle counts at all points. The emissions tests used should be defined and evaluated in terms of measurement error and vehicle-vehicle response differences with respect to a reference test (FTP or IM240). A definition and effectiveness evaluation of cutpoints should be made. Effectiveness should be evaluated in terms of false fails and false passes based on the repairs performed whenever possible.

4.2.3 Out-of-Program Comparison Data

States also may be able to use out-of-program comparison data to demonstrate inspection effectiveness[*]. Only in-program data can be used to demonstrate the I/M program data quality of a state's particular program as discussed in Section 4.2. However, the quality of the emissions inspections themselves may be judged using out-of-program comparison data . Two techniques

[*] The term out-of-program comparison data is used here to distinguish from the term out-of-program data that is typically used to refer to RSD or road side pullover data.

for doing so are discussed below. States may be able to suggest other techniques to help put the inspection effectiveness of a state's I/M program in perspective.

A round robin is a technique commonly used by laboratories to cross check analytical methods among a group of laboratories. For example, diesel fuel samples taken from a single bulk quantity are sent to different labs for analysis of aromatics. The labs may analyze the aromatics by their method of choice (e.g. FIA Hydrocarbon, HNMR, CNMR, GC-MS, Aniline Point, etc.) or by all the methods each lab has available. Analysis of the round robin results from all labs reveals which labs reported results that were significantly different from the participants in the round robin. Those "outlying" labs can then investigate the details of their analytical methods. If several different types of samples are sent to each lab, the results can also be used to look for biases among the analytical methods. The same round robin technique may be applied to emissions inspections as well and is commonly used by auto manufacturers and regulatory laboratories.

4.2.3.1 Vehicle Round Robin Testing

The first technique might be to send test vehicles to different I/M stations for testing. Shipment could be done using vehicle transporters so that the emissions characteristics are not changed greatly as a result of mileage accumulation; some states already do this. The vehicles would be selected to cover a range of technologies, model years, and emissions levels. The emissions of these vehicles could be tested at different I/M stations within the state. Analysis of results would indicate the variability among I/M stations in the state. If repeat tests were performed on the vehicles at each station, the variability of emissions testing at participating stations could be determined.

A slight variation of this application might be even more useful. Vehicles could be transported for testing at I/M stations in neighboring states. Where large populations are near a state border, private vehicle owners could be paid an incentive to participate in a state-to-state I/M program comparison effort. Since neighboring states may use different emissions measurement methods (e.g. IM240, ASM, two-speed idle, pressure, purge and pressure, gas cap check, etc.), these results would provide data to evaluate emissions measurement effectiveness of the different techniques and to establish relationships among the different methods. If the transport of vehicles is not possible, at a minimum, gas bottles of known concentration could be measured at the respective test facilities within a give state or among neighboring states to assess analyzer accuracy and judge the relative effectiveness of the slight differences that will invariably exist between analyzer QA/QC procedures.

4.2.3.2 Test Crew Round Robin Testing

In a second technique, instead of transporting vehicles, I/M instruments and test crews could be sent to neighboring states. The crews would set up at neighboring state I/M stations and inspect some of that state's vehicles. Vehicles would be inspected by their state's crew and then would be offered an incentive to undergo I/M testing by the out-of-state crew. Reciprocal agreements among neighboring states would provide for reciprocal testing visits and sharing of data. This technique would provide a much large sample of vehicles tested by two emissions measurement methods than the first technique.

4.2.3.3 Recommended Best Practices

The quality of the emissions inspections themselves can be judged using out-of-program comparison data. Round robins of vehicles or I/M analyzers with crews sent to I/M stations of adjacent states can be sources of data for comparisons. Emissions measurements of vehicles or gas bottles of known concentration analyzed by two different I/M programs will reveal measurement bias between the programs. If resources permit, the information provided by such efforts is believed to be worth pursuing.

<u>4.3. Effectiveness of Repairs</u>

4.3.1 Number and Type

State 3 requires all state-certified repair stations to record in the Vehicle Information Database the repairs that were made to each vehicle. For each repair event, the repair station records all repair actions that were made to the vehicle from a list of 34 repair types. Supporting information is also entered for station identification, vehicle identification, repair cost, repair date and time, etc.

Tables 4-3 and 4-4 show the frequency of repair station actions taken for each repair type for passenger cars in two different model year groups. Table 4-3 shows results for 2,486 repair events on 1976-1980 model year vehicles, and Table 4-4 shows results for 2,593 repair events on 1991-1995 model year vehicles. These model year groups were chosen to show the differences in repair types and frequencies for vehicles of different technologies and ages. The 34 repair types are described in the first column of each table. The last column of each table gives the percent of repair events that involved the item indicated.

In general these tables show the level of repairs that were made to these vehicles. Such data documents that repairs are being made and therefore, on the simplest level, the I/M program is causing repairs to be made to vehicles in the fleet. A state that has a larger fraction of its vehicles undergoing repairs in comparison to another state can, all other things being equal, be expected to have a more effective I/M program. Obviously, stations that perform repairs where none are needed will decrease effectiveness. Additionally, whether these repairs are effective at reducing emissions must also be demonstrated. This is the subject of the next sub-section.

4.3.2 Emission Reductions

A state can demonstrate the effectiveness of its I/M program by performing an analysis of in-program emissions measurements before and after repairs. At the simplest level, this can be demonstrated by the average emissions of repaired vehicles before and after repair and the average emissions change. State 3 used the ASM2525 test for its I/M program. Table 4-5 shows the averages for the repaired vehicles in the two chosen model year groups.

Initially, such a table seems to indicate that the I/M program is producing real emissions reductions. However, because of the "regression toward the mean" effect, any emissions reductions based on the same measurements used to declare vehicles as emission failures are biased. Thus, even if no repairs were made to failing vehicles, the average change of measured

Table 4-3. Repair Station Actions for 1976-1980 Cars

			Reconnected	Defective and Not Repaired	Not Applicable	Item is OK	Replaced	Repaired, Cleaned, or Adjusted		% Replaced, Repaired, Cleaned, or Adjusted
1	SPLUGS	spark plugs	0	7	309	1693	443	34		19.2
2	IGWIRE	ignition wires	5	15	320	1987	159	0		6.6
3	DISTR	distributor	0	7	327	1909	203	40		9.8
4	SPAADV	spark advance	9	37	258	1908	23	251		11.4
5	SPATIM	spark timing	0	5	180	1321	1	979		39.4
6	VACLEA	vacuum leaks	34	76	7	1739	16	614		26.7
7	IDLMIX	idle mixture	0	23	93	287	5	2078		83.8
8	IDLSPE	idle speed	0	8	122	478	3	1875		75.5
9	CARINJ	other carburetor or fuel 8injection work	1	328	287	1321	98	451		22.1
10	AIRFIL	air filter	0	9	219	1736	477	45		21.0
11	CHOKE	choke	1	32	442	1861	17	133		6.1
12	TAC	thermostatic air cleaner	28	156	495	1708	18	81		5.1
13	PCV	positive crankcase ventilation	5	16	376	1841	144	104		10.2
14	AIRINJ	air injection	22	222	1301	874	19	48		3.6
15	EGR	exhaust gas recirculation	81	213	785	1124	159	124		14.6
16	EVAP	evaporative control	14	81	572	1795	8	16		1.5
17	GASCAP	gas cap	0	9	869	1604	4	0		0.2
18	CATCON	catalytic converter	3	555	846	996	86	0		3.6
19	FFR	fuel filler restrictor	0	57	1133	1294	1	1		0.1
20	O2SENS	oxygen sensor	0	4	2405	66	10	1		0.4
21	TPS	throttle position switch	0	1	2369	107	1	8		0.4
22	WOT	wide open throttle sensor	0	0	2402	81	1	2		0.1
23	MAP	manifold absolute pressure sensor	1	0	2430	54	1	0		0.1
24	MAF	mass air flow sensor	0	13	2367	87	3	16		0.8
25	CTS	coolant temperature sensor	0	3	2166	311	4	2		0.2
26	TVS	thermal vacuum switch	9	57	1471	909	33	7		2.0
27	OTHSEN	other sensors	1	24	1933	522	5	1		0.3
28	PROM	engine management computer	0	8	2335	141	2	0		0.1
29	ENGINE	engine management computer	0	401	1049	1011	0	25		1.0
30	PVALVE	carburetor power valve	0	128	918	1271	66	103		6.8
31	CFLOAT	carburetor float	0	96	809	1365	66	150		8.7
32	EGRPAS	egr passages	1	179	973	1205	2	126		5.2
33	EGRCTL	egr controls	36	143	1009	1213	19	66		4.9
34	OTHER	other repair items	0	53	1299	1045	59	30		3.6

Table 4-2. Repair Station Actions for 1991-1995 Cars

			Reconnected	Defective and Not Repaired	Not Applicable	Item is OK	Replaced	Repaired, Cleaned, or Adjusted		% Replaced, Repaired, Cleaned, or Adjusted
1	SPLUGS	spark plugs	1	10	214	1614	697	55		29.1
2	IGWIRE	ignition wires	2	15	274	2086	213	1		8.3
3	DISTR	distributor	0	8	883	1479	187	34		8.5
4	SPAADV	spark advance	0	1	398	2152	3	37		1.5
5	SPATIM	spark timing	0	2	401	1741	2	445		17.3
6	VACLEA	vacuum leaks	13	3	10	2384	6	175		7.5
7	IDLMIX	idle mixture	0	9	681	1508	0	393		15.2
8	IDLSPE	idle speed	0	5	531	1690	0	365		14.1
9	CARINJ	other carburetor or fuel 8injection work	1	35	535	1670	27	323		13.5
10	AIRFIL	air filter	0	6	312	1798	425	50		18.3
11	CHOKE	choke	0	0	1875	704	6	6		0.5
12	TAC	thermostatic air cleaner	1	5	1585	989	1	10		0.5
13	PCV	positive crankcase ventilation	5	1	454	1900	169	62		9.1
14	AIRINJ	air injection	2	4	1685	883	6	11		0.7
15	EGR	exhaust gas recirculation	9	14	982	1370	99	117		8.7
16	EVAP	evaporative control	1	1	517	2060	4	8		0.5
17	GASCAP	gas cap	0	0	793	1795	3	0		0.1
18	CATCON	catalytic converter	1	342	154	1512	582	0		22.5
19	FFR	fuel filler restrictor	0	1	795	1786	8	1		0.3
20	O2SENS	oxygen sensor	9	48	170	1445	872	47		35.8
21	TPS	throttle position switch	0	2	486	2027	16	60		2.9
22	WOT	wide open throttle sensor	0	0	1282	1305	1	3		0.2
23	MAP	manifold absolute pressure sensor	2	2	1033	1549	3	2		0.3
24	MAF	mass air flow sensor	0	5	1355	1207	9	15		0.9
25	CTS	coolant temperature sensor	1	1	452	2103	22	12		1.4
26	TVS	thermal vacuum switch	1	0	1137	1450	1	2		0.2
27	OTHSEN	other sensors	1	5	654	1910	16	5		0.8
28	PROM	engine management computer	0	21	537	1994	31	8		1.5
29	ENGINE	engine management computer	0	143	762	1544	1	141		5.5
30	PVALVE	carburetor power valve	0	2	2449	133	1	6		0.3
31	CFLOAT	carburetor float	0	1	2469	117	2	2		0.2
32	EGRPAS	egr passages	0	10	1170	1309	1	101		3.9
33	EGRCTL	egr controls	5	12	1134	1359	26	55		3.3
34	OTHER	other repair items	0	17	1156	1223	49	146		7.5

emissions for the fleet would show a decrease. The reason for this is that vehicles that are declared failures tend to have measurements with positive emissions measurement errors. Therefore, states need to use a technique for producing the data for a table such as Table 4-5 in a manner that corrects for regression toward the mean. Section 4.3.4 describes such a method.

Table 4-5. Observed Average Emissions Before and After Repairs

	N	Average ASM2525 Concentration Before Repair			Average ASM2525 Concentration After Repair			Average ASM2525 Concentration Change		
		HC	CO	NO_x	HC	CO	NO_x	HC	CO	NO_x
		(ppm)	(%)	(ppm)	(ppm)	(%)	(ppm)	(ppm)	(%)	(ppm)
1976-1980 Cars	2486	187	1.58	1143	106	1.05	870	-81	-0.52	-273
1991-1995 Cars	2591	87	0.84	902	35	0.14	511	-52	-0.70	-391

By combining the repair data with the emissions data, an analysis will reveal the emissions effects of different combinations of repair types. For example, Table 4-6 shows the most frequent combinations of repair types for the two chosen model year groups. The 15 most frequent repair combinations for the 1976-1980 cars accounted for 33% of the repair events for this vehicle group. For the 1991-1995 car group, the 8 most frequent repair combinations accounted for 33% of the repair events.

An examination of individual repair combinations, their associated average emissions before repair, and the emissions changes that the repairs produced shows expected effects of repairs on emissions. For example, for the 1991-1995 car group, Repair Slate D5 (EGR) was applied to vehicles with very low HC, very low CO, and very high NO_x emissions and resulted in small changes in HC and CO, but large decreases in NO_x. On the other hand, Repair Slate D3 (Catalytic Converter and O_2 Sensor) was applied to vehicles with moderately high HC, CO, and NO_x emissions and resulted in relatively large decreases in HC, CO, and NO_x emissions. For the 1976-1980 car group, Repair Slate A10 (major carburetor work) was applied to vehicles with the highest average HC and CO and just about the lowest NO_x and resulted in large decreases in HC and CO and large increases in NO_x.

Each state is encourage to collect repair data in a similar way, then comparison of results such as those shown in Table 4-6 could be part of a repair program evaluation. For example, it would be expected that repair stations perform the same repair slates on corresponding technology vehicles in different states, although the frequency distribution will vary with test type and cutpoints. In addition, the average before-repair emissions and emissions changes for those repair slates should be similar among different states with comparable repair programs. If one state's repair stations applied repair slates more indiscriminately than another state's, the differences among before-repair emissions averages would be smaller and emission decreases would be smaller.

Table 4-6. Emission Reductions Associated with Combinations of Repairs

		Type of Repair														**Average ASM2525 Concentration Before Repair**			**Average ASM2525 Concentration Change After Repair**		
Repair Slate	**N**	**Spark Plugs**	**Spark Advance**	**Spark Timing**	**Vacuum Leaks**	**Idle Mixture**	**Idle Speed**	**Carb or Finj Work**	**Air Filter**	**EGR**	**Catalytic Convertor**	**O2 Sensor**	**Carb Power Valve**	**Carb Float**	**Other**	**HC (ppm)**	**CO (%)**	**NO_x (ppm)**	**HC (ppm)**	**CO (%)**	**NO_x (ppm)**
1976-1980 Cars																					
A1	258					X	X									140	1.44	812	-37	-0.38	56
A2	121			X		X	X									142	1.32	1129	-46	-0.27	-237
A3	75					X										123	1.53	928	-10	-0.31	11
A4	63			X	X	X	X									122	0.99	1114	-50	-0.29	-279
A5	51					X	X	X								140	2.41	749	-10	-1.19	165
A6	40				X	X	X									150	1.05	900	-56	-0.25	10
A7	35									X						115	0.16	2780	-38	-0.02	-1720
A8	32					X	X		X							132	1.43	973	-7	-0.20	26
A9	28		X	X		X	X									114	1.30	1320	-25	-0.11	-217

A10	22					X	X	X					X	X		219	3.78	598	-79	-1.73	333
A11	19					X	X	X	X							165	3.50	589	-74	-2.11	483
A12	18					X	X			X						89	0.42	2265	-1	0.18	-1278
A13	18			X		X	X	X								118	2.76	757	-41	-1.04	-165
A14	17	X		X		X	X									216	1.01	1246	-99	-0.07	-67
A15	16		X			X	X									98	0.58	989	-30	0.16	-351
1991-1995 Cars																					
D1	301											X				101	1.75	581	-74	-1.66	-169
D2	237										X					69	0.25	1317	-48	-0.20	-812
D3	80										X	X				85	0.60	1159	-70	-0.55	-728
D4	58	X														85	0.34	783	-40	-0.23	-273
D5	55									X						25	0.11	1589	7	0.04	-719
D6	49														X	45	0.19	570	-19	-0.13	-225
D7	44			X												80	0.63	1176	-41	-0.36	-446
D8	38	X										X				120	1.79	559	-89	-1.80	-194

4.3.3 Repair Lifetimes

Once a state has shown that its I/M program is causing repairs to be made and the repairs are causing emissions reductions, the final effect the state should quantify is the lifetime of the repairs. If repairs last only a short time, the emissions benefits may only last a short time. If the repairs last many years, then it is at least possible that the emissions benefits may last many years. In addition, long lasting repairs help reduce the number of repairs that will be expected in future years. In other words, one reason the number of repairs is low in a given year may not be because of a failure of the vehicle inspections to identify them. Instead, it may be because repairs made in previous years are durable.

The duration of repairs can be evaluated by analyzing a good repair database. For this example, the repair data from State 3 was analyzed. For this state the I/M program repair data for five consecutive years was available and subset of the vehicles that had any repair performed in the first year was selected. The number of days between that first repair and the next repair of any kind was calculated. If the vehicle did not get a second repair in the five-year data set, then the duration was set to 1825 days for plotting purposes. Figure 4-11 shows the result of that distribution.

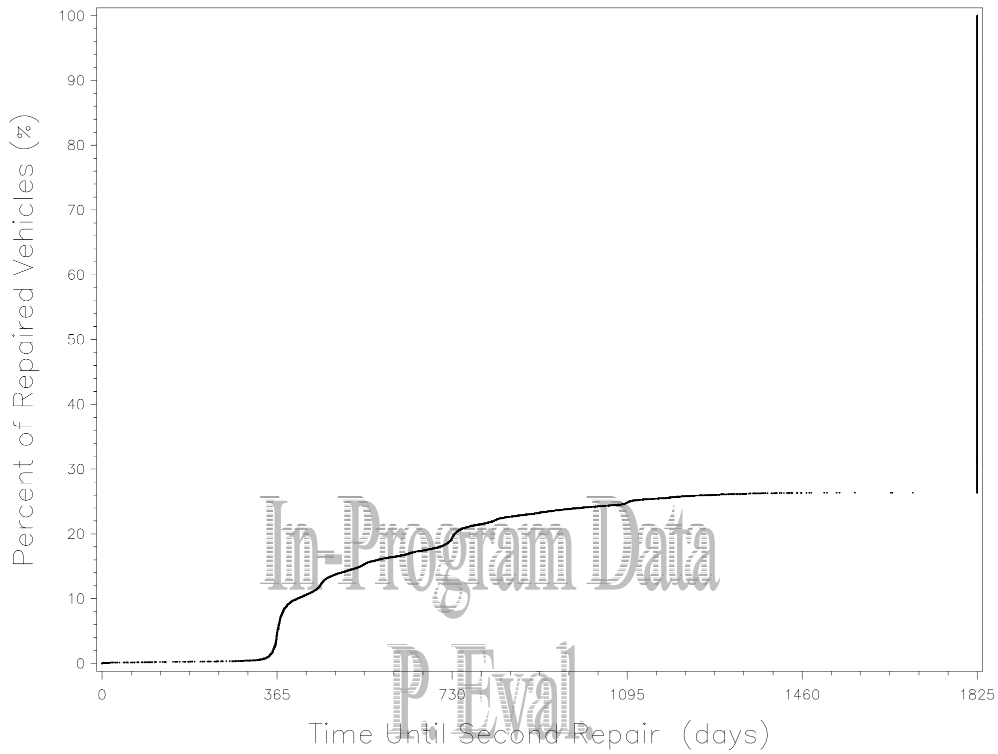


Figure 4-11. Distribution of Intervals from Any Year 1 Repair to Any Next Repair

The cumulative distribution shows that about 25% of the vehicles that had a repair in the first year had at least a second repair by the fifth year. This leaves 75% of the vehicles that had a repair in the first year and did not receive a repair (or at least did not have a second repair recorded in the database) for the next four years. Perhaps more importantly, the plot also indicates that by the end of the second year about 20% of the vehicles already had a second repair. This rapid rise in subsequent repair intervals suggests that some vehicles require frequent repairs.

The programmatic implications will depend on an analysis by repair type. Some repairs may be routine adjustments that are not really the result of serious degradation. Examples are idle speed and idle mixture adjustments on carbureted vehicles. This contrasts with catalytic converter replacements, which should not be performed routinely on any vehicle.

A more detailed analysis of this repair data by vehicle age, vehicle technology, and repair type should illustrate the situations where repair durability is strong and where it needs to be increased. Such an analysis could help a state improve its repair stations' performance. From an I/M program evaluation perspective, an analysis of overall repair duration for the repaired vehicles and a targeted analysis for different repair types would demonstrate that, beyond simply making repairs, the repair stations are making repairs with a quantifiable durability.

4.3.4 Other Measures

In an effective I/M program the vast majority of vehicles that initially fail the emissions inspection will require only a single repair event to pass the emissions inspection. In less effective I/M programs, some vehicles will make repeated trips between inspection and repair in an effort to meet annual I/M emissions requirements. The cause for such "ping-ponging" may be emissions measurement error, faulty repair diagnosis, or poor repair quality. Whatever the cause, the vehicle owner will be frustrated. Emissions measurements and repair events with date and time stamps are required to evaluate "ping-ponging" events.

I/M program inspection and repair databases also reveal that some owners of failing vehicles will go from inspection station to inspection station to try to find a station that will pass their vehicle. This so-called "shopping around" is distinguishable from "ping-ponging" because for "shopping around" consecutive inspections do not have repair events between them.

In this example, State 3 apparently recorded all repair types, even if they occurred at different repair events, as a single repair event. Accordingly, separate "ping-ponging" from "shopping around" cannot be filtered out. Table 4-7 shows the distribution of repeated fails for State 3 as an example of the type of result that could be expected from an analysis of "ping-ponging."

Table 4-7. Distribution of Repeated Adjacent Inspection Failures Prior to a Pass

Fail Sequence	**Number of Vehicles**
F	78,878
FF	16,871
FFF	2,711
FFFF	587
FFFFF	130
FFFFFF	25
FFFFFFF	13
FFFFFFFF	2
FFFFFFFFFF	1

Another measure of repair effectiveness is a comparison of the cost of all repairs to the reduction of all emissions. Cost-effectiveness values ($/ton) could be calculated for the I/M program overall and for individual repair slates. The calculations would require the logging of the repair bill for each repair event. Calculated cost-effectiveness values can then be compared with reference values from sources such as U.S. EPA and California BAR or other states.

4.3.4 In-Program Studies to Measure Repair Effectiveness

Various methods can be used to quantify repair effectiveness using modifications to normal in-program procedures. The effectiveness of repairs can be examined by comparing the change in emissions of failing vehicles when they are repaired with changes in emissions of vehicles that are not repaired but just tested again. This comparison must be performed to avoid so-called regression toward the mean, which would cause repair emission benefits to be over-estimated. An example is described below. Other such methods for measuring repair effectiveness may be devised.

A subset of vehicles failing the initial emissions test (Test A) would be assigned to the Evaluation Group or the Control Group. These vehicles would be selected from the set of all failing vehicles using a stratified random sampling method. Vehicles in the Evaluation Group would immediately receive a second emissions test (Test B) and would then be sent for repairs based on their result on Test A (i.e., even if they passed Test B). When these vehicles returned from repair, they would be given the repair follow-up emissions test (Test C). Following Test A, vehicles in the Control Group would immediately receive a second emissions test (Test D) and would immediately be given a third emissions test (Test E). Then, these vehicles would be sent to repair and would return for their repair follow-up emissions test.

The actual emissions benefit of repairs is (C-B) - (E-D). This is the change in emissions before and after repairs of the Evaluation Group vehicles less the change in emissions of vehicles in the Control Group that did not receive repairs. It is critical that Test A results not be used to calculate repair benefits. Doing so would introduce a bias in the calculated benefits. It is also critical that all vehicles that fail Test A and pass Test B be sent for repairs. To not do so would also introduce a bias in the calculated benefits.

4.3.5 Repair Data Collection

Repair data needs to be collected by I/M programs for analyses of repair effectiveness to be made. Development of data collection requirements can begin with the approaches used in states that are currently collecting such data. Then, improvements to the approaches can be made as states gain experience collecting and analyzing the data.

Repair stations should enter vehicle, emissions, and station information for each repair they make:

- station identification,
- vehicle identification,
- repair date and time,
- repair cost, and
- repair codes for standard repairs such as those in Table 4-1a.

This repair information should be entered each time a vehicle enters a repair station for work. In most states, most repair work is done in repair stations that are not connected to the VID, or repairs are done by the vehicle owner. Therefore, to allow the VID to achieve completeness and accuracy targets for repair data, techniques need to be developed for acquiring repair data.

4.3.6 Recommended Best Practices

Section 4.3 discussed methods for evaluating the effectiveness of repairs in an I/M program. Unfortunately, most current I/M programs place greater emphasis on accurately measuring vehicle emissions and designating vehicles as pass or fail than on ensuring or even monitoring the quality of vehicle repairs. This natural emphasis is probably a consequence of the more quantifiable aspect of emission measurement over vehicle maintenance. As the discussions in Section 4.3 demonstrated, acquiring a database of vehicle repairs would provide information and opportunities that are not currently available in most I/M programs. Therefore, one the most important recommendations is for states to develop database systems which are capable of monitoring vehicle repairs so that the beneficial aspects of the analysis of those databases can be realized. The list below summarizes the key aspects in this regard:

Repair data collection. States need to make a concerted effort to collect repair information on all vehicles participating in the I/M program. The data should be collected in a manner such that it can be matched to emissions data for each vehicle. Each visit of a vehicle to a repair station should generate a record in the database. The record would include vehicle identification, codes for the types of repairs performed, and the cost of the repairs. Strategies must be developed to ensure that all repairs performed would be recorded in the database. One possibility worth consideration is for states to certify repair stations.

Number and type of repairs. Once the database is created, simple counts of the number and type of repairs demonstrate that repairs are being performed. Analysis of the data would show what types of repairs are common for different types of vehicle technologies.

Repair lifetimes. Analysis of the repair data set could be used to quantify the duration of repairs. While some repairs are routinely performed as vehicles go out of adjustment, others reflect the lifetime of repair components and the general competence of repair stations. Repair lifetimes should be compared among different states to determine the typical repair lifetimes in different I/M programs.

Emissions reductions for repairs. By combining the emissions database with the repair database, it would be possible to demonstrate that repairs are actually reducing emissions. More specifically, an analysis would quantify how emissions are being reduced for each type of repair. Such analyses from different states sould be compared to arrive at a consensus estimate of the reductions that can be achieved by certain types of repairs. As a side benefit, the fingerprint of emissions on vehicles that have failed the inspection could be associated with the types of repairs that successfully caused the vehicle to pass the follow-up emissions test. Such relationships could be used to develop diagnostic guidance for repair stations to use.

Measures of customer inconvenience and repair cost. The combined repair and emissions databases could be used to determine the extent of customer inconvenience produced by repeated visits between inspection and repair stations at the time of the annual or biennial inspection. Such so-called ping-ponging can be produced by excessively stringent cutpoints, inspection emissions test measurement error, faulty repair diagnosis, or poor repair quality. When repair costs are included in the repair database, the total customer repair dollars can be determined. Also, the repair costs for each type of repair can be determined with respect to the emissions reductions that are achieved.

In-program studies to measure repair effectiveness. Slight modifications to the inspection sequence for a subset of vehicles in the I/M program can produce data that will provide an estimate of the effectiveness of the I/M program. The modifications are used to eliminate biases produced by the so-called regression toward the mean effect.

5. Results Based Measures of Effectiveness

This section will outline procedures for analyzing the data in I/M vehicle inspection records. Previous methods developed by stakeholders, contractors and EPA for this analysis will be reviewed in Section 5.1 and 5.2. Section 5.3. contains descriptions of a new set of analysis procedures as well as a brief discussion of the use of out-of-program data. Section 5.5 discusses the testing of evaporative emissions. None of the procedures use MOBILE modeling; comparisons are made between different years of test data and between different programs, but projections to no-I/M levels are not attempted. The significance of any results obtained through analysis of the I/M test records must be weighted by the findings from the procedures in Sections

3 and 4. Additionally, the data validation methods described in Section 4 must be applied prior to analysis. It is also important to realize that the model year results described in this section should be weighted by vehicle miles traveled or some other travel fraction weighting.

5.1 ECOS Method

The Environmental Council of States (ECOS) Group was formed in 1996 to develop an evaluation process for state I/M programs with test and repair networks[15]. The primary objective of the group was to develop common criteria to demonstrate equivalency to EPA's I/M program standard. Twelve criteria were developed for a short-term qualitative evaluation that was to be performed 6 months after program start-up. A successful completion of each criteria conferred a set number of points that counted toward a successful fulfillment of the ECOS program evaluation requirements. However, the focus of the criteria was on the comparison of test-and-repair I/M stations to test-only stations, so that other differences that might exist between programs, such as test type, data quality assurance, or cutpoint stringency, were not evaluated. A second longer-term quantitative evaluation was then to be performed 18 months after program start-up. One of the difficulties with the implementation of the ECOS method was that each state chose a set of criteria from the twelve options to apply to their program, so it was possible to choose analyses that provided favorable results, and ignore other analyses with unfavorable results. Use of the ECOS criteria was discontinued in 1999.

5.2 EPA Tailpipe I Method

This method was based primarily on work done for EPA by Sierra Research, Inc. in 1997[16]. The original study done by Sierra was focused on comparing designer I/M tests to known reference tests such as the IM240. However, in response to a court ordered deadline that required EPA to establish program evaluation protocols, this study was used and modified so that it could meet this need.

Under this method, a small sample of vehicles that has already met the I/M program requirements is recruited for an additional I/M test. Emissions data from these vehicles is compared to a baseline program that closely matches EPA's requirements for an "Enhanced I/M" program. Regions that use I/M tests other than the IM240 are required to develop and apply a correlation to relate emissions data from their program to equivalent IM240 results. The MOBILE5 model is used to correct for regional differences between the two programs, such as altitude, climate, or fuels. The specific steps that have been taken to apply the method for several I/M programs[17, 18] are listed in Table 5-1. The final result of the comparison between the program under evaluation and the benchmark program is a ratio of the effectiveness of the two programs.

The benefit of the Tailpipe I method is its capacity to condense comparisons between the I/M program and the benchmark program into a single ratio. Also, the concept of developing a correlation between the program test (TSI, ASM, etc.) and the IM240 test is a valuable tool for comparing in-program data from programs using different tests. However, the reliance on the MOBILE5 model to make the regional corrections and determine the no-I/M levels (see Table 5-1) may introduce error to the results. The method also requires the use of an I/M program compliance rate, which can be difficult to determine. Finally, while the use of a single comparison between the two programs is convenient, it may result in some loss of detail, and relevant information that might be found through a multi-faceted approach could be missed.

Table 5-1. Steps for Application of the EPA Tailpipe I Method for an I/M Program Using the Two-Speed Idle Test

1	A random, stratified sample of about 800 vehicles is selected for use in developing a relationship between the state's two-speed idle test results and IM240 test results.
2	Back-to-back IM240 and two-speed idle tests are conducted on the sample of vehicles. This dataset is used to develop a correlation between the results of the two speed idle test and IM240 emissions.
3	An estimated IM240 result was calculated for each I/M test record, using the correlation between two-speed idle test results and IM240 emissions that was developed according to Step 2.
4	The 2% random sample of complete IM240 tests that is collected annually by Phoenix, Arizona is obtained, representing data from a benchmark program.
5	Separately for each program (program under evaluation and benchmark program): An average IM240 emissions level is calculated by model year.
6	Separately for each program: Travel fractions based on registration distributions and MOBILE5 annual mileage accrual rates are used to calculate a single average emissions level.
7	The Arizona model year average emission levels are converted to match the program under evaluation by correcting for any differences in fuel, altitude, climate, and calendar year effects.
8	MOBILE5b is used to model Arizona's average emission levels with and without an I/M program in place. Inputs are based on local area parameters for the program under evaluation. The results of this modeling are used to calculate a percent reduction in emission levels, or benefit, achieved by the benchmark Arizona program.
9	Average IM240 emissions levels for Arizona were calculated in Step 5. The benefit of the Arizona program was calculated in Step 8. These two results are used to calculate the average IM240 emissions level for Arizona without an I/M program in place (No-I/M levels).
10	The No-I/M emission levels calculated in Step 9 are compared to the average estimated IM240 emission levels in program under evaluation that were calculated in Step 5. These results are used to calculate the percent reduction, or benefit, of the program under evaluation.
11	The benefits of the two programs are then compared.

5.3 Use of Data Trends

The use of data trends can be used to highlight differences between programs that may provide useful information if investigated further. Several different types of analysis using I/M program records are considered:

Fleet average emissions changes for a single I/M program year,
Fleet average emissions changes comparing multiple years of testing,
Emissions changes in individual vehicles over multiple years of testing, and
Comparisons with other I/M program results (different states or regions).

Fleet average emissions changes over a single year are computed in order to determine whether the I/M program results in emissions reductions over a single program cycle, after any failing vehicles are identified, repaired, and re-tested. Without an I/M program in place, the vehicle would deteriorate and emissions would increase over time. With an I/M program in place, deterioration should be identified and the vehicle repaired at each test cycle. Looking at vehicle emissions over multiple years where overall fleet emissions are being reduced as new vehicle emissions technologies are introduced into the fleet makes this problem of identifying I/M program effectiveness even more difficult. The goal in investigating fleet average emissions changes over a single year is to determine whether deterioration is actually being identified and reduced through repairs. A lack of emissions reductions in one year of program data would indicate that any long-term fleet average emissions reductions are attributable to fleet composition changes, rather than I/M program results. This type of analysis is demonstrated in Section 5.3.1.

If an I/M program benefit within a single year is shown, then the emissions averages of the fleet over time should be examined for long-term effects. Due to the problems associated with determination of no-I/M emissions levels (i.e., moving away from empirical data with MOBILE modeling, or attempting to project next years' emissions levels from this years' failed test results), analysis methods are presented herein that are based on year-to-year data. These year-to-year comparisons are included in Section 5.3.2. Section 5.3.3 contains a similar analysis, but fleet changes are eliminated by tracking individual vehicles that participated in the program over multiple years.

In Section 5.3.4, program data from three different areas is compared. The comparisons are made using two-speed idle data from two areas and IM240 data from a third. An additional discussion of using a correlation to predict IM240 emissions levels from TSI results, as proposed in the EPA Tailpipe I method, is included there. However, none of the analysis suggested requires use of a correlation to compare data from states that use different types of tests.

5.3.1 Fleet Average Emissions Analysis for a Single Program Year

The single-cycle effect of an I/M program on a fleet may be found by comparing average emissions levels at the beginning and the completion of the test cycles (a test cycle includes all tests and retests for a vehicle, until it completes or drops out of the program). In Figure 5-1, for State 1, the initial and final IM240 HC emissions of all passenger cars are presented. The State 1 program allows vehicles to fast-pass the IM240 test, so results for the shorter tests must be projected to full test results. Methods for projecting full test results from fast-pass data may be found in the literature[19, 20]; however, care must be taken to fully understand the implications of using such algorithms as they may bias the results of the program evaluation analysis. The data in Figure 5-1 is grouped by initial and final test result. It can be seen that the average emissions of the vehicles that initially failed but were eventually repaired and passed decreased significantly, almost to the level of the vehicles that passed on the first attempt. Vehicles that dropped out of the program before being repaired and passing an inspection show almost no reduction (the two lines are difficult to differentiate because they lie almost on top of each other).

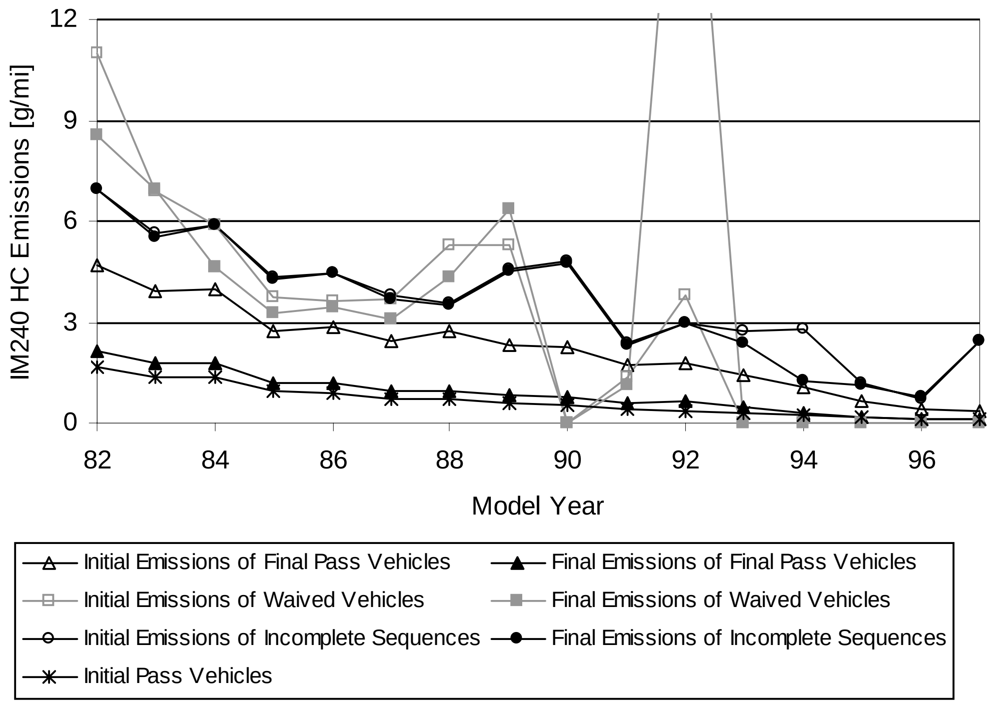


Figure 5-1. Initial and Final Emissions for All Passenger Cars, IM240 HC, State 1

While Figure 5-1 gives a good visual representation of the emissions reductions, it could be misleading on its own. For example, the figure shows extremely high emissions for 1992 vehicles that received waivers, but it doesn't show that this group includes only one vehicle. A minimum of 25 records per bin is often considered to be a cutoff below which averages are unreliable (as for the 1992 waived vehicles in Figure 5-1). Table 5-2 provides additional information about the data presented graphically in Figure 5-1, for the vehicles that initially or ultimately passed the I/M test. From the table, it may be seen that sample sizes vary greatly among the model years. It may also be seen that the standard deviation of the results is often as large or larger than the mean value; this large spread is not apparent from Figure 5-1.

Table 5-2. Initial and Final Emissions for All Passenger Cars, IM240 HC, State 1

	Initial Pass			Initial Fail (for Vehicles that Ultimately Pass)			Ultimate Pass (After Initially Failing)		
Model Year	Number of Vehicles	Mean HC [g/mile]	Std. Dev.	Number of Vehicles	Mean HC [g/mile]	Std. Dev.	Number of Vehicles	Mean HC [g/mile]	Std. Dev.
82	4831	1.66	0.97	889	4.67	5.38	889	2.15	1.18
83	12760	1.37	0.83	1689	3.91	4.03	1689	1.79	1.05
84	9885	1.37	0.86	1508	3.96	4.62	1508	1.77	1.07
85	23440	0.98	0.65	3910	2.75	3.12	3910	1.18	0.82
86	14504	0.87	0.62	1935	2.88	3.30	1935	1.16	0.82
87	32629	0.72	0.54	3028	2.42	2.94	3028	0.94	0.71
88	18189	0.73	0.56	1495	2.73	4.17	1495	0.98	0.75
89	41190	0.57	0.46	1707	2.30	3.79	1707	0.81	0.64
90	19388	0.54	0.45	812	2.27	3.67	812	0.78	0.68
91	45202	0.41	0.36	1351	1.73	2.91	1351	0.62	0.58
92	18782	0.36	0.34	533	1.80	2.63	533	0.64	0.57
93	44006	0.29	0.27	792	1.41	2.47	792	0.47	0.45

94	38857	0.21	0.22	393	1.09	2.32	393	0.31	0.39
95	22329	0.16	0.18	231	0.67	1.26	231	0.18	0.27
96	15457	0.12	0.12	159	0.43	1.37	159	0.14	0.19
97	9327	0.11	0.09	57	0.38	1.79	57	0.09	0.12

The other point of information not shown in either Figure 5-1 or Table 5-2 is that the emissions data that is averaged to generate each data point does not exhibit a normal (Gaussian) distribution. Figure 5-2 shows the distribution of values of IM240 HC in all records for the initial test on 1990 vehicles that ultimately passed. The data does not have a symmetric normal distribution: the vast majority of vehicles have emissions near zero, while the high emitting vehicles form a long "tail." When plotted on a logarithmic scale, the distribution is more nearly symmetric, as shown in Figure 5-3. Because the log-normal distribution includes only positive values, and because it condenses high values while spreading out the lower values, it is often used to transform emissions data. Averages of emissions should still be performed on the raw (linear space) values as for Table 5-2 since those averages represent the emissions of the model year average vehicle.

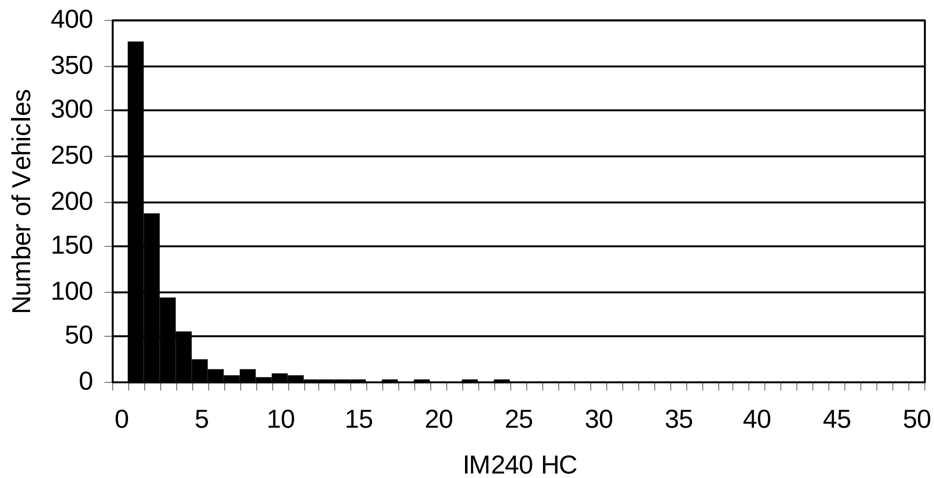


Figure 5-2. Distribution of Records for Single Data Point, IM240 HC, State 1

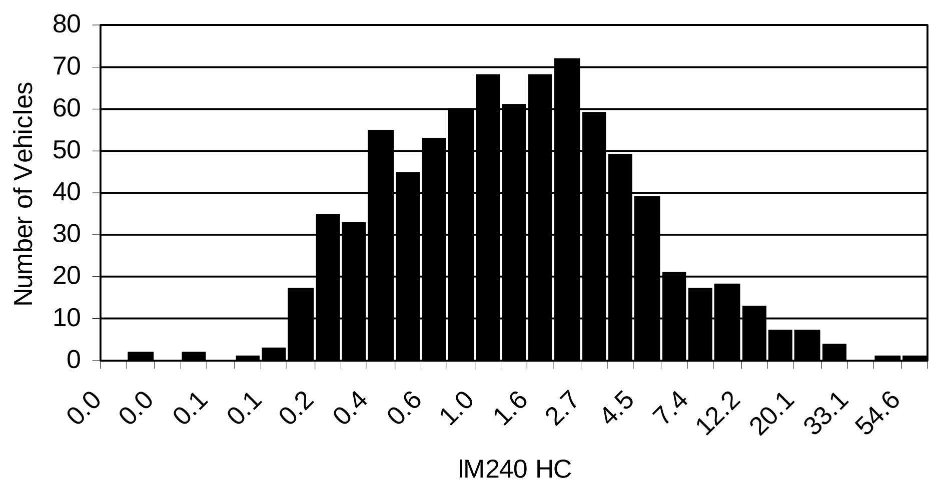


Figure 5-3. Log-Scale Distribution of Records for Single Data Point, IM240 HC, State 1

Figure 5-1 presented only the IM240 HC emissions levels for passenger cars. If the purpose of this report were to analyze the program effectiveness of State 1, additional figures would be given for light duty trucks (and heavy duty trucks, if covered by the program), and results for IM240 CO and NO_x would be presented as well. This level of detail is useful in identifying groups of vehicles with anomalous results, but larger trends may be easier to see in a more general presentation such as Figure 5-4. This figure presents the overall emissions reductions for passenger cars, as a percent decrease from initial to final test. Vehicles with only one test (i.e., initially passed or initially failed and dropped out of program) are included in the averages for both the initial and final tests. It is clear that the vast majority of emissions reductions result from the older vehicles.

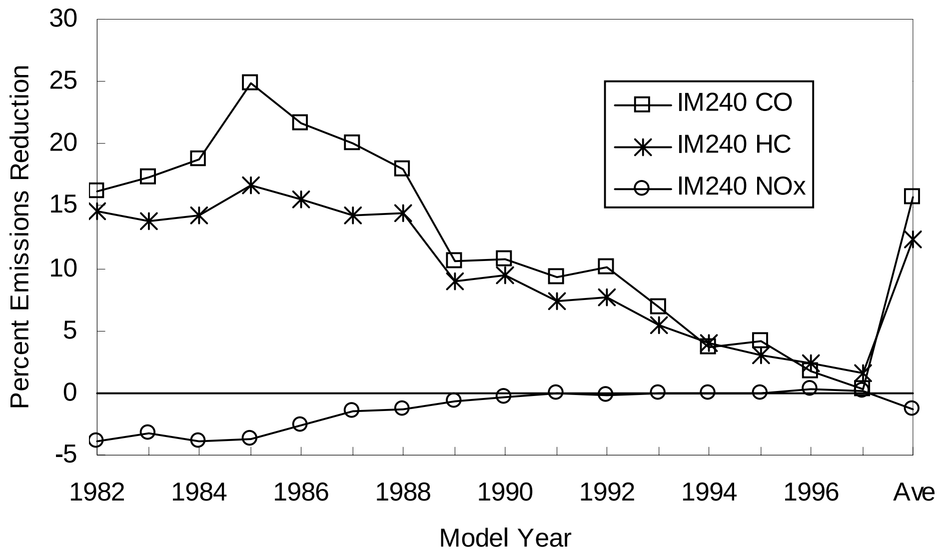


Figure 5-4. Overall Emissions Reductions for Passenger Cars, Incomplete Sequences Included, State 1

One assumption behind the data in Figure 5-4 was that the vehicles that left the program before passing a test (dropping out before completing their test sequence) remained in the area; data for their last inspection is included in the average. However, if these vehicles were sold or otherwise moved outside the program area, then they are no longer part of the fleet and the data for their last inspection should be removed from the final test average. This change was made for Figure 5-5, but resulted only in slightly greater average reductions.

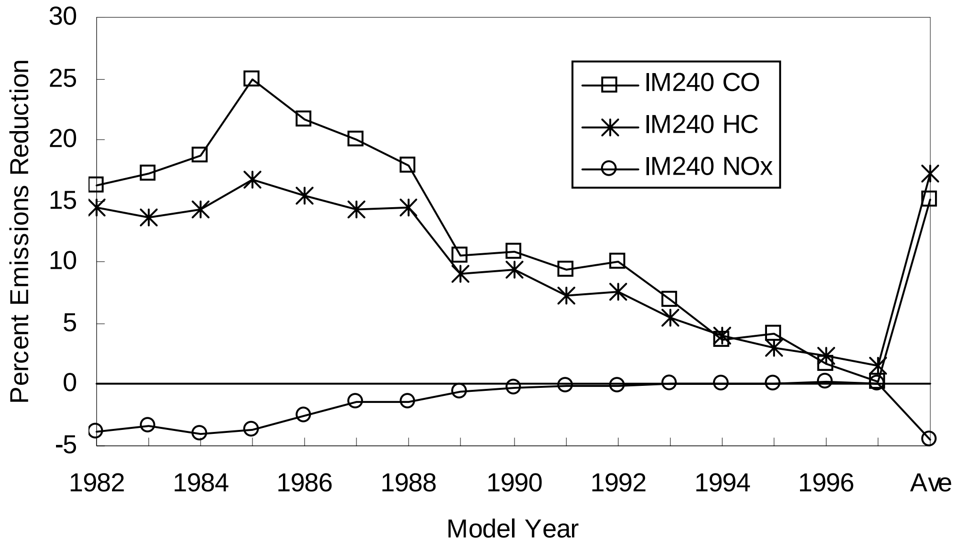


Figure 5-5. Overall Emissions Reductions for Passenger Cars, Incomplete Sequences Not Included, State 1

In addition to emissions reductions, the rate at which vehicles fail the inspection can be informative; for example, a very low fail rate may indicate that cutpoints are too high to identify some vehicles that would benefit from repair (see discussion of cutpoints in Section 4.2.2.2). The rate at which vehicles failed their initial test in State 1 is shown in Figure 5-6. The overall height of each bar indicates the total percentage of vehicles that failed their initial test; the different sections within the bars divide the vehicles by the result they finally achieved before leaving the program. Vehicles that receive waivers comprise a very small percentage of the total.

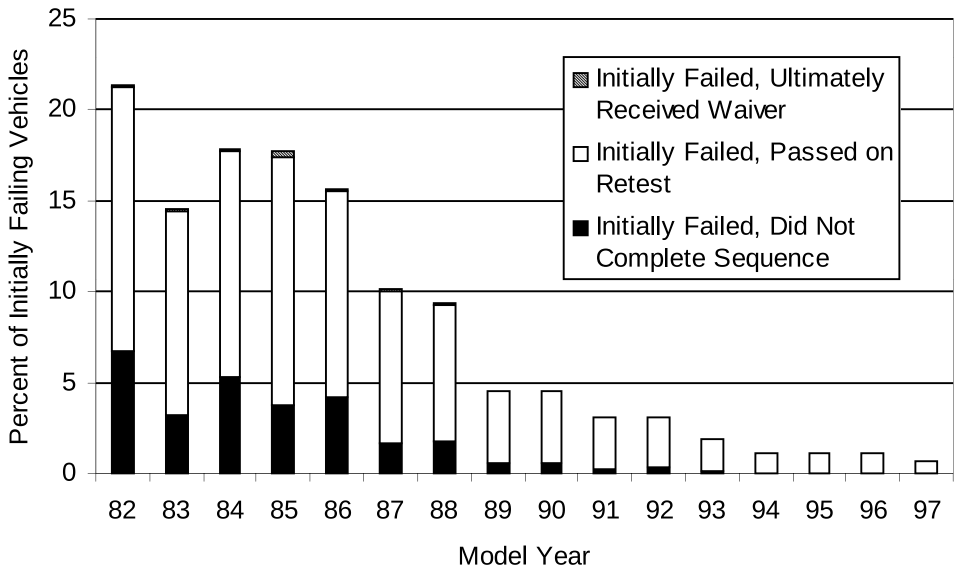


Figure 5-6. Fail Rate for Initial Test, IM240, State 1

Finally, the number of tests required for vehicles to complete the program is shown in Figure 5-7. Vehicles that passed their initial test are not included on this figure, since they each had exactly one test. This information is somewhat related to the repair information presented in Section 4.3, i.e. more effective repairs require fewer repeat tests before a vehicle passes. It is interesting to note that vehicles that eventually drop out of the program before passing tend to average almost as many repeat tests as vehicles that eventually pass. However, from Figure 5-1 it was seen that the emissions levels of these vehicles were almost unchanged from initial to final failed test. It is possible that these vehicles are not being repaired between tests, or that the owners leave the program in discouragement when repairs show no emissions benefit.

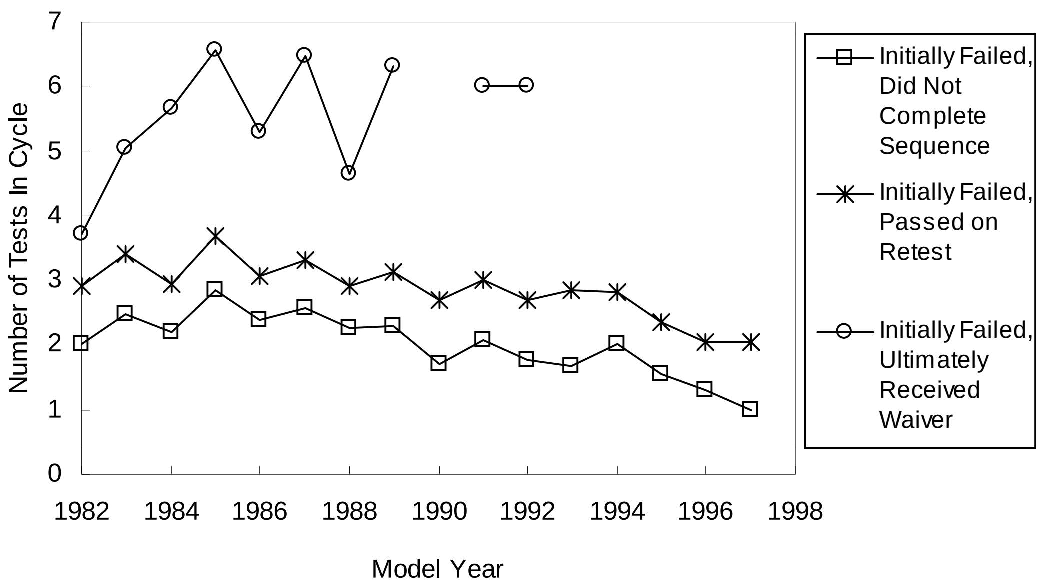


Figure 5-7. Average Number of Tests Required to Complete Program, State 1

5.3.1.1 Recommended Best Practices

It is recommended that analyses as illustrated in Figure 5-1 be used with vehicle miles traveled (VMT) data to obtain average emissions by model year and test sequence. Figures 5-4 and 5-5 demonstrate how in-program data may be used to estimate average emissions reductions by model year. Analyses such as those in Figure 5-6 should be used to track the rate at which waivers are issued; the rate at which vehicles are repaired, resulting in an air quality benefit; and the rate at which vehicles drop out the program, resulting in a lost air quality benefit, while Figure 5-7 type of analyses provide information to track the progression of vehicles through the program.

Each of these analyses uses only a single year of program data; this is the most basic level of emissions results analysis. Whenever possible, the use of analyses such as those depicted in Figures 5-1 and 5-4 through 5-7 should be combined with the multiple-year and multiple-state analyses described in 5.2.2 through 5.2.4.

5.3.2 Fleet Average Emissions Analysis for Multiple Program Years

The comparison of multiple years of I/M test records allows the observance of fleet emissions trends over time.

Trend analyses as portrayed in Figure 5-8 have been used by others[21] may be used to examine the changing emissions of the fleet over different program years. Each line shows the average emissions for the initial test of a different model year vehicle, plotted against the age of the vehicle at the time of test. Without an I/M program in place, the emissions of each model year would be expected to increase as the vehicles age. For the two-speed idle HC data of State 3, shown in the figure, the average emissions in the newest model years actually do show increasing emissions over time. However, the emission levels may still be so low that the vehicles are not yet affected by the I/M program. The significant increase in emissions levels between 1988 and 1987 illustrates the significance of cutpoints in fleet emissions as the 1987 HC cutpoints are almost twice as high as the 1988 cutpoints. For other fleets without a similarly large change in cutpoints, the gap in emissions between 1987 and 1988 is not seen, indicating that the gap on Figure 5-8 is not due to vehicle technology changes. For the model years older than 1987, the decrease in emissions as the vehicles age is clear, possibly indicating that the program is having an effect on this component of the fleet, or that high-emitting vehicles drop out of the program or are sold out of the program area to avoid further testing.

In summary, the primary purpose of Figure 5-8 is to look for potential problems, such as gaps between the model years that indicate inadequate cutpoints, or large increases in emissions within a model year as the vehicles age, indicating unchecked deterioration.

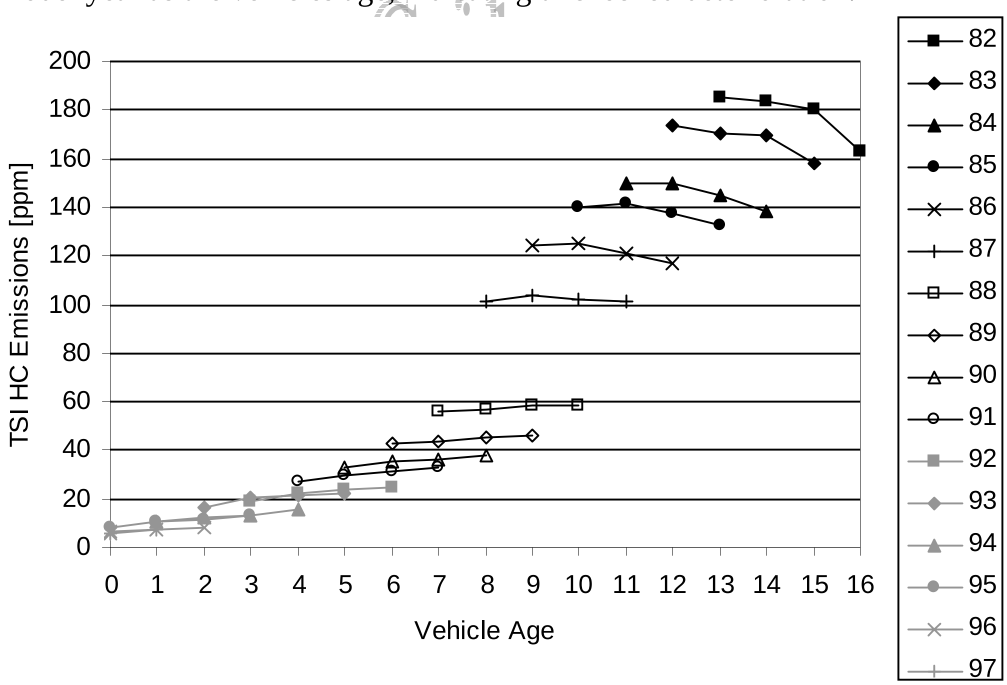


Figure 5-8. Emissions Averages at Different Vehicle Ages, TSI HC, State 3

Figure 5-9 shows the percent emissions reduction from initial to final test for State 3 over four years of I/M testing (similar to the single year of data shown in Figure 5-4). The x-axis of the figure is the vehicle age, so a vehicle with an age of five years in the first year of program data will be shown as six years old in the second year of program data and seven years old in the third year of data. This age-based presentation allows the emissions reductions over the different I/M program years to be compared based on the length of time the vehicle has had to deteriorate. Thus Figure 5-9 may be used to investigate whether the effectiveness of the I/M program changes over time. For example, an ideal case would be a fleet with no immigration of vehicles from outside the program area, covering a fleet of well-maintained vehicles. In this situation it would be possible for all vehicles to eventually be repaired to passing emissions levels, and no further emissions reductions would be achieved. While reductions didn't drop to zero for State 3, as shown in Figure 5-9, it does appear that the reductions decrease over the four years presented. This may indicate that many high-emitting vehicles have been repaired, fewer vehicles are failing the test, and the program is having a benefit. Conversely, it is possible that the emissions levels for the initial tests are increasing. Figure 5-8 is based on initial emissions, and indicates that while initial emissions for the 1988 and newer vehicles increase slightly from year to year as they age, the initial emissions of the older vehicles do not increase as they age from year to year.

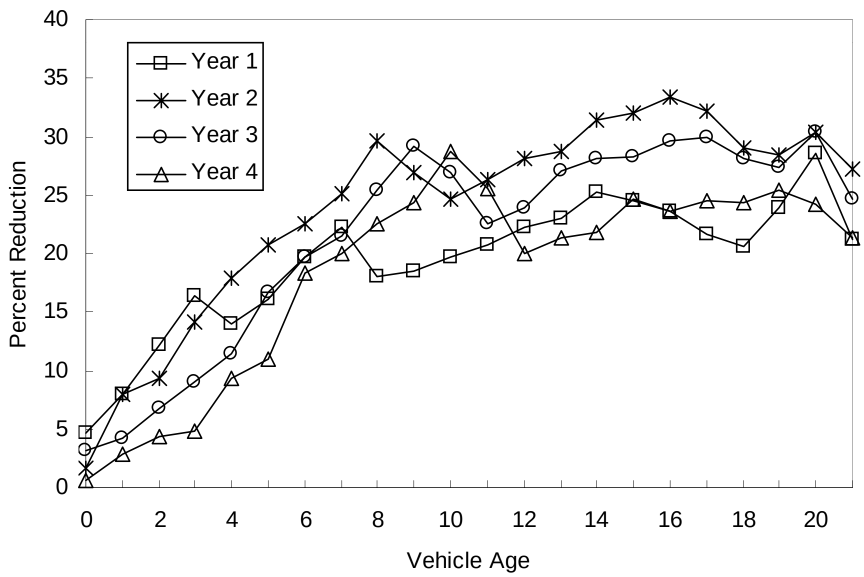


Figure 5-9. Percent TSI HC Reduction Over Four Years, State 3

The initial fail rate for the vehicles of State 3 is shown in Figure 5-10, for four years of program data. The trends correlate to what was seen in Figure 5-9 as the high fail rate for the older vehicles, which decreases over the four program years, fits well with the high reductions seen in Figure 5-9. Any inconsistencies between these two figures (i.e., a very low fail rate but high emissions reductions) might be an indication of a problem with the I/M program data.

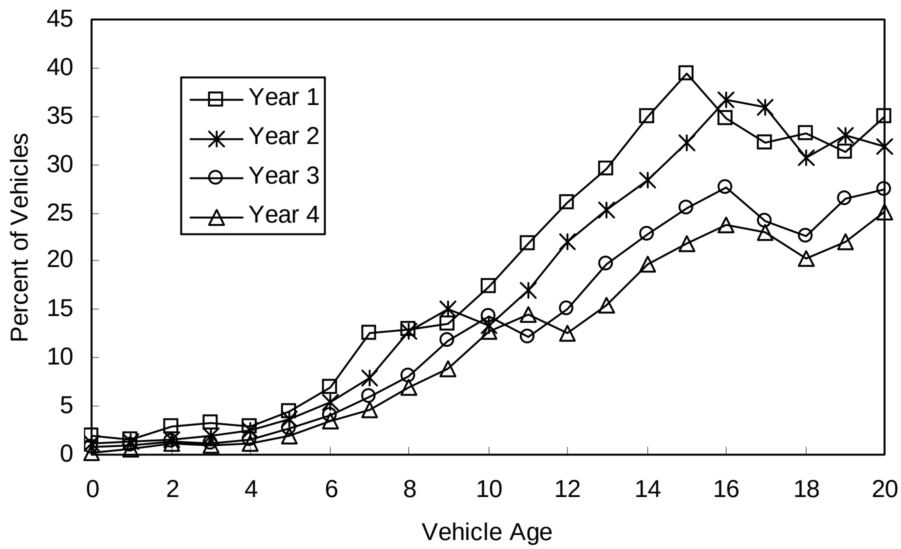


Figure 5-10. Initial Fail Rate, State 3

Figure 5-11 illustrates another type of analysis done by McClintock[22], looking at the skewness of State 3's TSI HC emissions. This figure shows the percent of total emissions that are contributed by the dirtiest 10% of vehicles in each model year, which is as high as 50% for State 3. Since they contribute such a large portion of the total emissions, repair of these vehicles provides a large portion of the emissions reductions an I/M program achieves. From Figure 5-11, it can be seen that the emissions contributed by the dirtiest 10% of the vehicles remains relatively constant over the program years. Again, there could be more than one explanation. For example, the overall fleet emissions may be decreasing, with emissions from the dirtiest 10% decreasing in approximately the same proportions, or the highest emitters result may be due to new vehicle equipment malfunctions or immigration of high-emitting vehicles.

Another way of looking at the skewness is to look at the emissions contributed by 10% of the dirtiest vehicles of the overall fleet, chosen without stratifying by model year. These vehicles are concentrated in the oldest model years, as shown in Figure 5-12, with very few newer vehicles in the group. The percent of the emissions that the overall 10% dirtiest vehicles in the fleet contribute to each model year is shown in Figure 5-13. For the oldest model years, this contribution is over 80%. This type of information could have several uses. For example, if a high-emitter identification program is being considered, Figures 5-12 and 5-13 could help identify model years with the greatest number of target vehicles. Also, changes in the distribution shown in Figure 5-13 from year to year of program data could identify cutpoint problems. For instance, if high emitters were increasingly concentrated at a certain age range, the cutpoints at that age might be too lax.

5.3.2.1 Recommended Best Practices

The analyses shown in Figures 5-8 through 5-13 should be used when multiple years of program data are available. Figure 5-8 may be used to look for potential problems, such as gaps between the model years that indicate inadequate cutpoints, or large increases in emissions within a model year as the vehicles age, indicating unchecked deterioration. The percent reductions over the program years shown in Figure 5-9 should be used to confirm that the program retains its effectiveness over time. The initial fail rates shown in Figure 5-10 should be analyzed in conjunction with Figure 5-9; high fail rates that are not coupled with high emissions reductions indicate problems with the program.

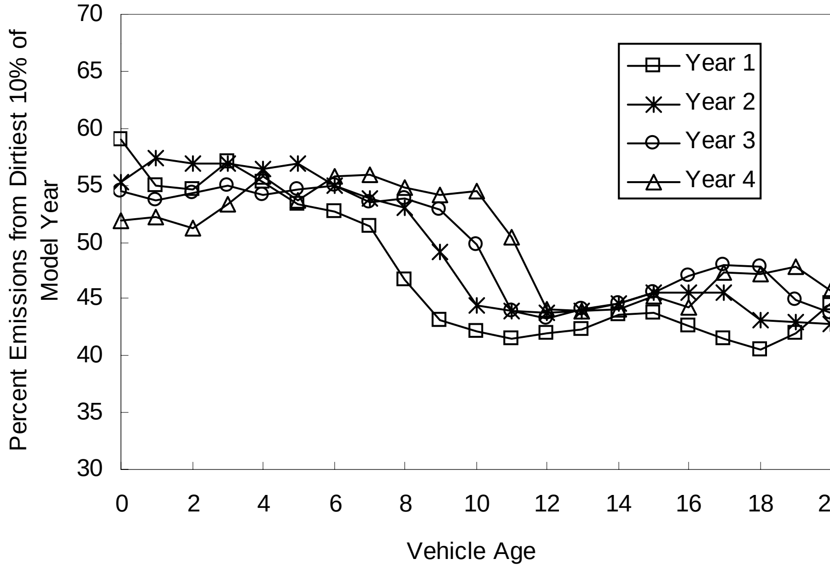


Figure 5-11. Percent TSI HC Emissions Contributed by Dirtiest 10% of Each Model Year, State 3

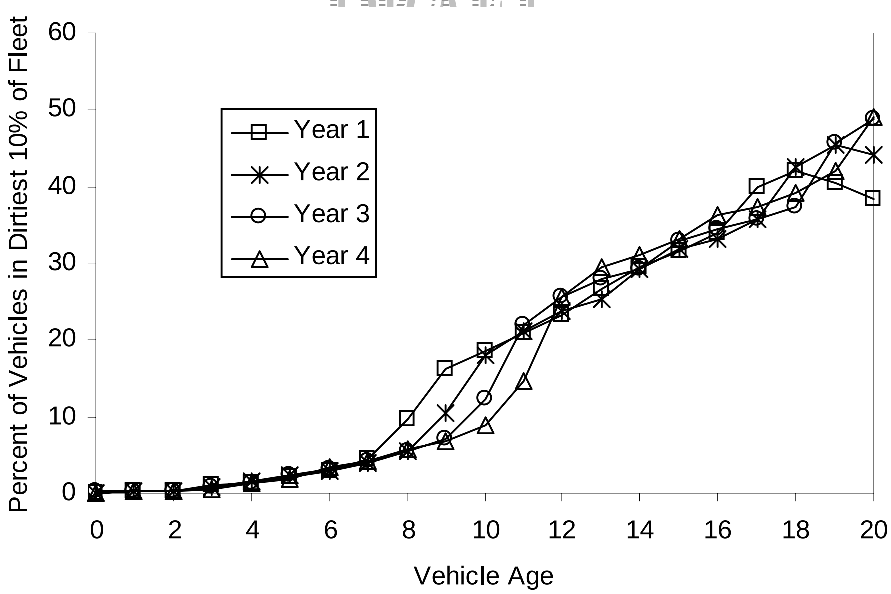


Figure 5-12. Distribution of the Dirtiest 10% of Vehicles in the Overall Fleet, State 3

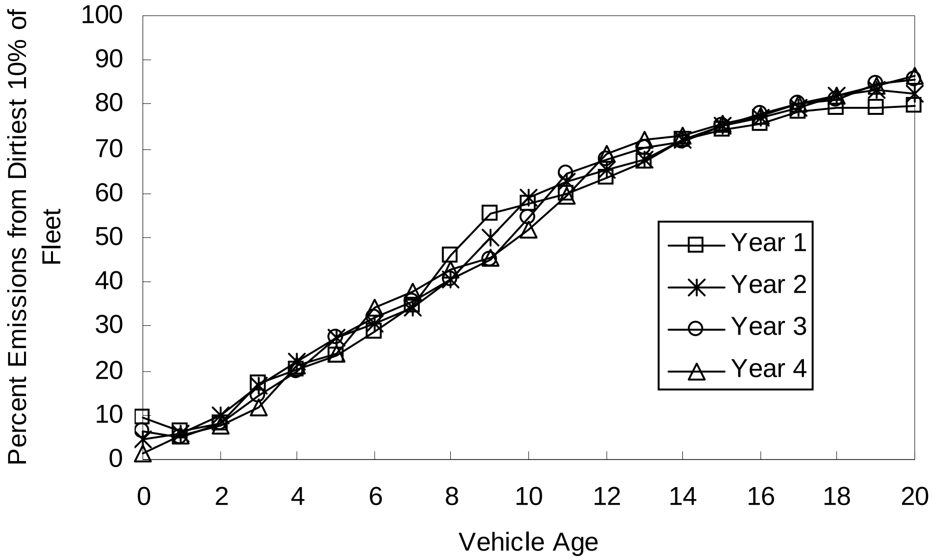


Figure 5-13. Percent of TSI HC Emissions Contributed by Dirtiest 10% of Overall Fleet

5.3.3 Emissions Changes in Individual Vehicles Over Multiple Program Years

Emissions changes over four years of testing were discussed in the previous section. However, possible changes in fleet composition (immigration and emigration of vehicles) lead to uncertainty about the meaning of some of the results. In this section, data from State 3 is used again, but only records for vehicles that were tested in all four years are used. Thus, any effects from changes in fleet composition are eliminated.

First, Figure 5-14 shows emissions levels as vehicles age (as was done for the entire fleet in Figure 5-8, in 5.3.2). The emissions levels for each of the newer model years, years that are not yet greatly affected by the program, are similar in the two figures. In the older model years, however, the emissions of the constant fleet in Figure 5-14 are lower than that of the entire fleet shown in Figure 5-8. This could indicate that an influx of dirtier vehicles is pushing up the average emissions of the fleet in Figure 5-8. The older model years in Figure 5-14 do not show a pronounced decrease in emissions levels as they age, as was seen in Figure 5-8. Since the decreases with age are not found in the constant fleet, the decreases seen in Figure 5-8 must have been due to the departure of high emitting vehicles from this portion of the I/M program fleet. However, the difference between Figures 5-8 and 5-14 cannot be used to estimate program benefit from attrition, since it is not known whether the departing vehicles moved out of the area or dropped out of the program without leaving the area. If more accurate vehicle tracking data were available, this type of analysis could provide some estimate with regard to program benefits resulting from vehicle attrition.

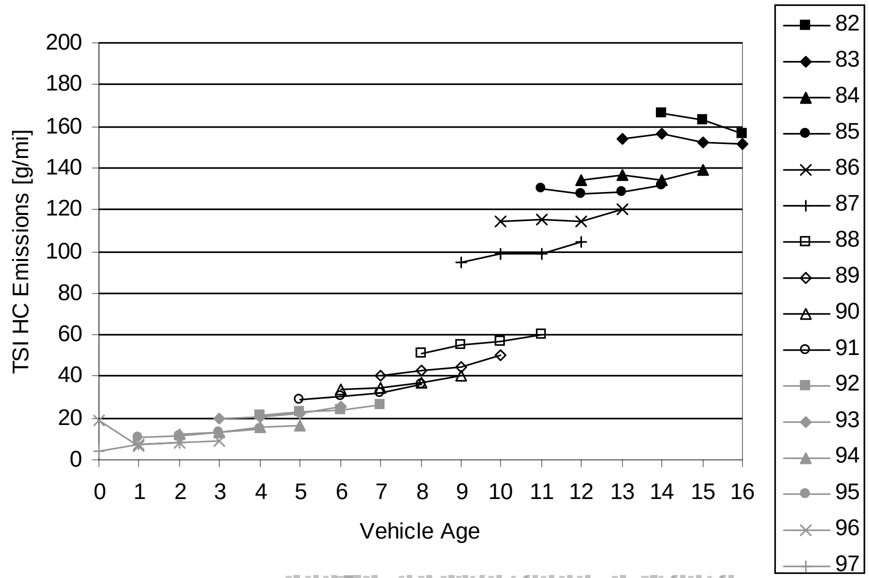


Figure 5-14. Emissions Averages at Different Vehicle Ages, Vehicles Tested All Four Years, TSI HC, State 3

The same type of plot is repeated again in Figure 5-15. However, this plot includes only vehicles in the year after they failed an I/M inspection initially and passed on a retest (i.e., if a vehicle failed and then passed on retest in Year 3, it is included in the Year 4 data here). The sample sizes are smaller so more scatter is evident in the data, but it is clear that the emissions are considerably higher than for the fleet as a whole; either these vehicles are poorly maintained and new problems arise each year, or repairs are not lasting a full year. The initial fail rate for these vehicles that previously failed and then passed is shown in Figure 5-16. Especially for the newest vehicles and years three and four of the program, the initial fail rate is significantly higher than the rate for the entire fleet, shown in Figure 5-10. This type of information might indicate that the program could achieve greater emissions reductions over the year if the test interval is shortened for vehicles that failed an earlier test. The collection and analysis of repair data described in Section 4.3 should be used to determine whether such changes could result in increased emissions reductions.

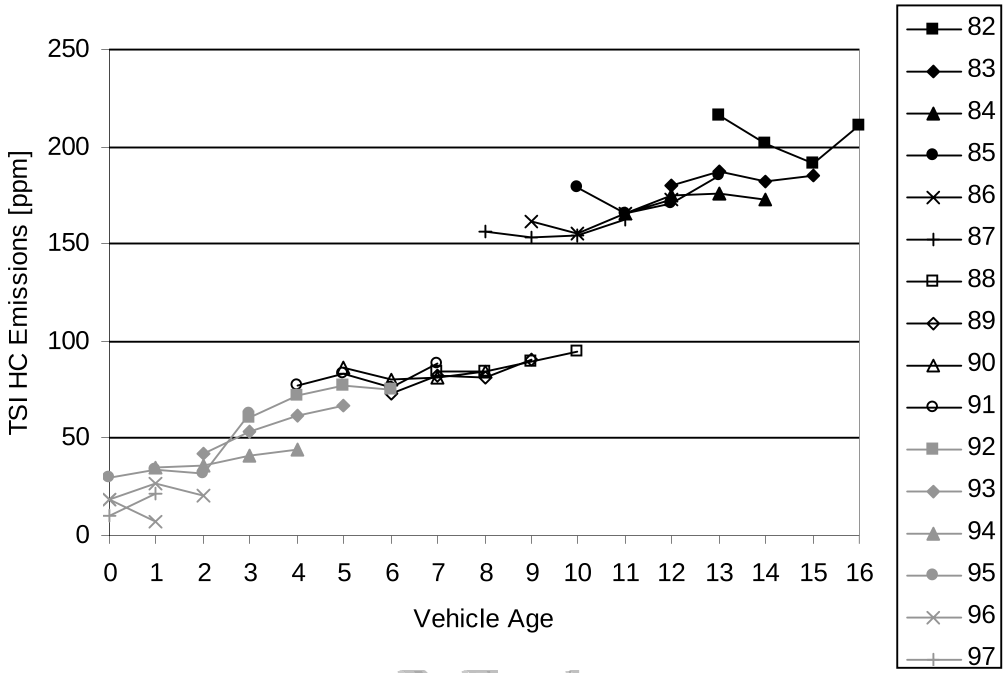


Figure 5-15. Emissions Averages at Different Vehicle Ages, Vehicles That Failed in Previous Year, TSI HC, State 3

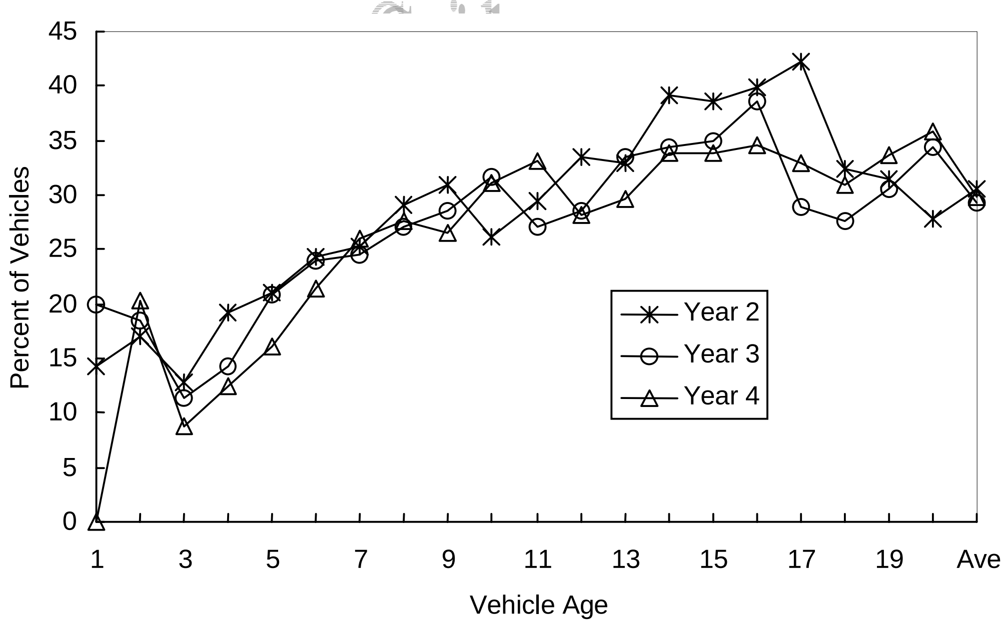


Figure 5-16. Initial Fail Rate, Vehicles That Failed in Previous Year, State 3

Figure 5-17, showing total reductions from initial to final test for the set of vehicles tested in each of the four years, correlates to Figure 5-9 for the whole fleet. The decreasing reductions seen in Figure 5-9 are seen again here, so it can be concluded that they were not caused by immigration or emigration of vehicles. Figures 5-17 and 5-9 are very similar overall.

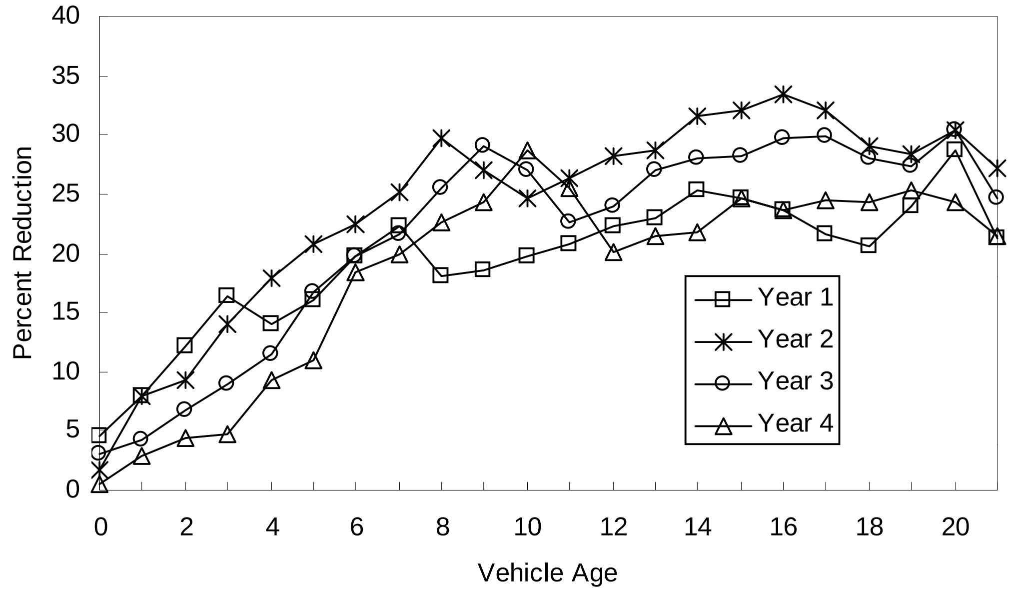


Figure 5-17. Percent TSI HC Reduction for Vehicles Tested All Four Years, State 3

If initial-to-final test emissions reductions are achieved by the same fleet year after year, a question arises as to whether the fleet is simply getting cleaner each year, or whether some of the gains made in one test cycle are lost by the start of the next test cycle. The change in emissions from final test one year to initial test the next year should be investigated as shown in Figure 5-18. Unlike Figure 5-17, this figure shows the percent increase. Vehicles that initially passed as well as those that initially failed and were repaired are included in the figure. Figure 5-18 shows that initial test scores are indeed higher each year than the previous year's final test scores, indicating that year-to-year vehicle deterioration does provide opportunities for I/M programs to achieve air quality improvements.

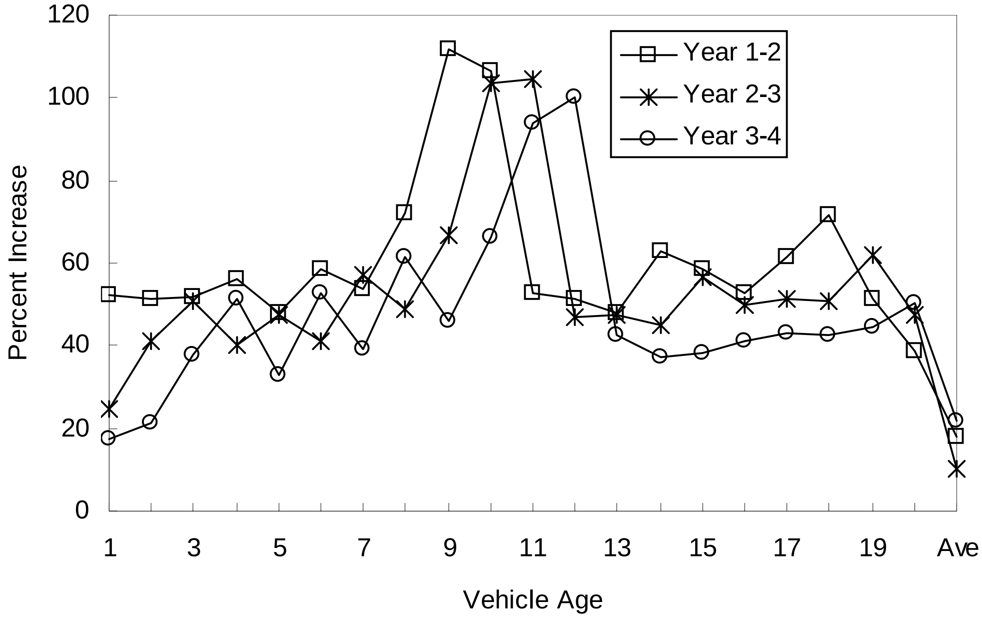


Figure 5-18. Increase in Emissions from Final Test in a Cycle to Initial Test in the Next Cycle, TSI HC, State 3

Comparison of Figure 5-15 and 5-8 indicated that emissions levels are somewhat lower for the vehicles in Figure 5-15 that participated in the I/M program for four years in a row compared to vehicles that were not tested in one or more of those years. The higher emissions of vehicles new to the program are shown in Figure 5-19; the bars indicate the ratio of the TSI HC emissions of vehicles new to the program to emissions for vehicles tested in at least one previous year. The new-to-program vehicles exhibit consistently higher emissions for Years 2, 3, and 4 of I/M program data examined. The lower emissions of the fleet that was tested yearly, as compared to the emissions of immigrating new vehicles, may indicate that the I/M program is providing a lasting benefit to the vehicles in the program, outweighing the effects shown in Figure 5-18.

The new-to-program vehicles comprise about 10% of the total tested vehicles for each model year, which might be a large enough sample to use them as a “No-I/M” fleet. However, since the origin of the vehicles is unknown (they may have just migrated from an I/M program in another area), it wouldn’t be certain that the new-to-program vehicles would really represent “No-I/M” vehicles. If the I/M program could determine the location of prior registration for these vehicles, then only those from non-I/M areas could be used to estimate the emissions of the No-I/M fleet while operating under the state’s local area parameters.

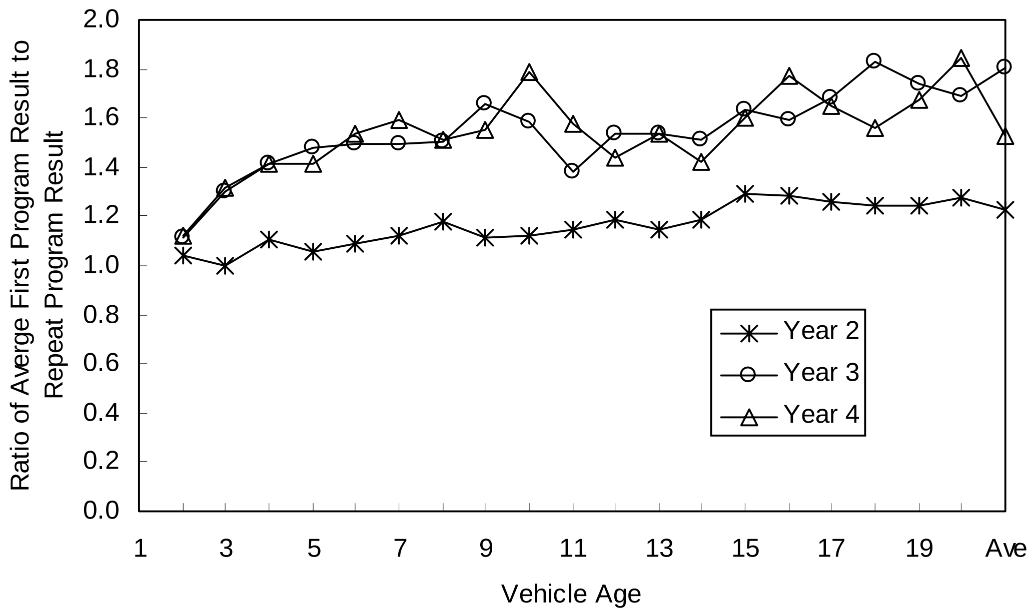


Figure 5-19. Ratio of Emissions of New Vehicles to Returning Vehicles, TSI HC, State 3

5.3.3.1 Recommended Best Practice

The value of the analyses described in this section is that they permit the examination of program trends without any effect of changes in fleet composition over the years. As seen by comparing Figure 5-14 to 5-8 the effects of vehicle immigration or emigration on emissions levels can be examined. Figures 5-15 and 5-16 should be used in conjunction with repair effectiveness analysis to determine repair durability between tests, and used to indicate if a change in testing frequency should be made. Figure 5-17 should be used to determine whether emissions reductions are achieved after several years of testing the same group of vehicles, while analyses depicted in Figure 5-18 should be used to determine how much of the reductions within a program year illustrated in Figure 5-17 are negated out by increases between program years.

5.3.4 Comparisons with Other Programs

The following comparisons between different I/M programs are qualitative only, due to the numerous differences in the programs being compared. Quantitative estimates may be possible, but would require the programs being compared be much more similar.

In this example, State 1 uses an IM240 test with a fast-pass component and a two-year test cycle; States 2 and 3 uses a TSI test with a yearly cycle and different cutpoints. Regional factors such as climate, altitude, and fuel are also different. However, these comparisons can be used to identify unusual trends that might not otherwise be noticed. For example, earlier Figure 5-8 for State 3 showed a large jump in TSI HC emissions between 1988 and 1987, when the cutpoints changed. In Figure 5-20 below, total percent reductions (similar to Figures 5-4 and 5-9) for three different I/M programs are presented. Total percent reductions are calculated from the change in average emissions from initial to final test. The data used is comprised of IM240 HC results for State 1, and TSI HC results for States 2 and 3. From the figure, it can be seen that the emissions reductions of State 3 excelled for the newest 10 model years, after which the cutpoints were increased to much higher levels, and the emissions reductions dropped well below the reductions achieved by the other states. The comparison to other programs, at first, suggests that State 3 might benefit from more stringent cutpoints for the older model year vehicles. However, upon further reflection, it could also mean that State 3 has achieved more significant reductions from more durable repairs in past years.

These two very different interpretations of Figure 5-20 demonstrate a key concept in I/M program evaluation, i.e. an I/M program must be evaluated using many different and complementary analysis tools to provide a balanced view. For example, to look only at the emissions reductions achieved during the inspection/repair cycle, but ignore emissions increases during the rest of the year, may lead to an inaccurate evaluation of the I/M program.

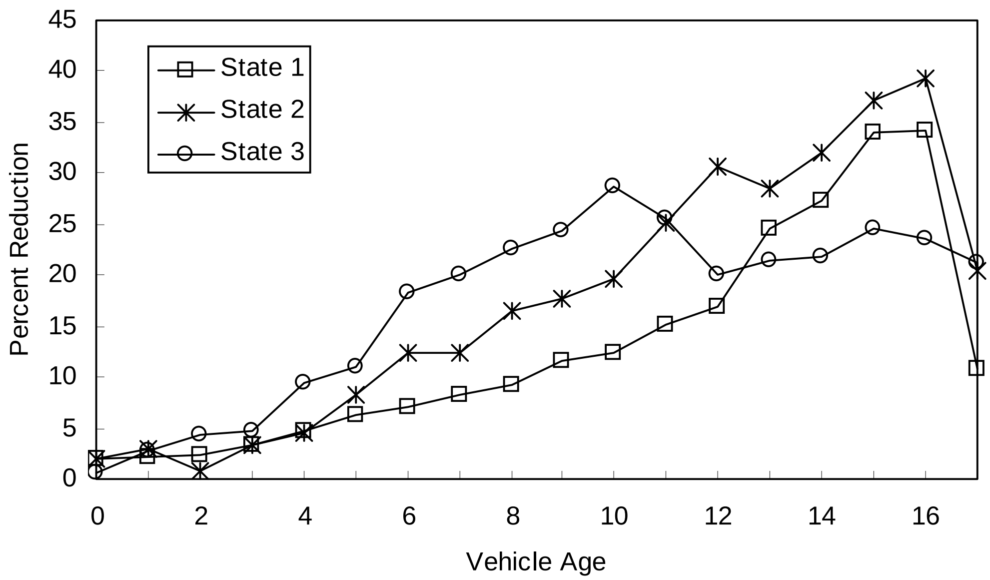


Figure 5-20. Comparison of Percent Reductions

Using the percent emissions reduction as the basis for comparison, as was done in Figure 5-20, eliminates the effect of the different units used by the TSI and IM240 tests. Thus comparisons may be made without having to convert the results of the two different tests to a common basis. However, the magnitude of the emissions reductions may differ for the different types of tests, so the information obtained from figures such as Figure 5-20 is only useful for identifying trends.

It should be noted that in general, comparing mass emission reduction estimates between programs is preferred to comparing percent reductions. Reporting reductions in units of mass would allow direct comparisons between programs to be made with less misunderstanding. For instance, an idle program study could report a 15% reduction in CO, while an IM240 program could report an 8% CO reduction and one may be led to believe that the idle program was twice as effective. However, this is not necessarily the case because the CO excess mass emissions for an idle test could be 25 g/mi, with I/M yielding a 3.75 g/mi reduction, while the IM240 area could have a CO excess mass emission of 80 g/mi, that would translate an 8% reduction into 6.4 g/mi.

Appendix A outlines procedures used for predicting IM240 mass emissions from TSI data in State 4. Input parameters for the correlation included TSI test result data, vehicle type, age, and engine size, as well as information about the emissions equipment of the vehicle. The correlation was applied to statewide TSI data. In Figures 5-21 and 5-22, the percent emissions reductions are shown when calculated using measured TSI data as compared to predicted IM240 data. It can be seen from the figures that the reductions are smaller when calculated using the IM240 data. Thus, comparison of TSI reductions in one state to IM240 reductions in another state may overstate the relative benefit of the TSI program. This is why it is preferred to report and compare emission reductions on a mass basis.

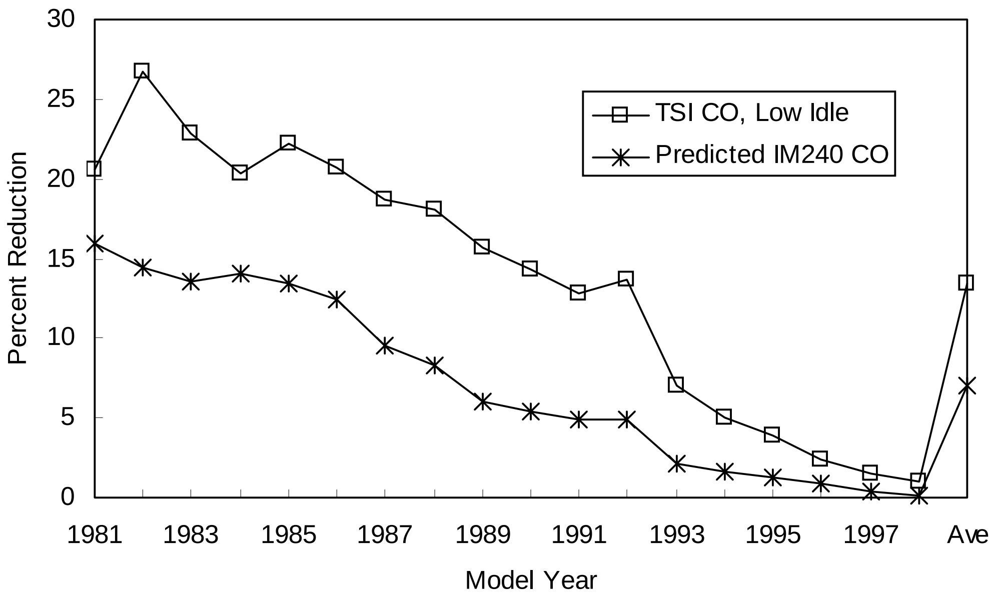


Figure 5-21. Percent Reduction of Measured TSI CO and Predicted IM240 CO

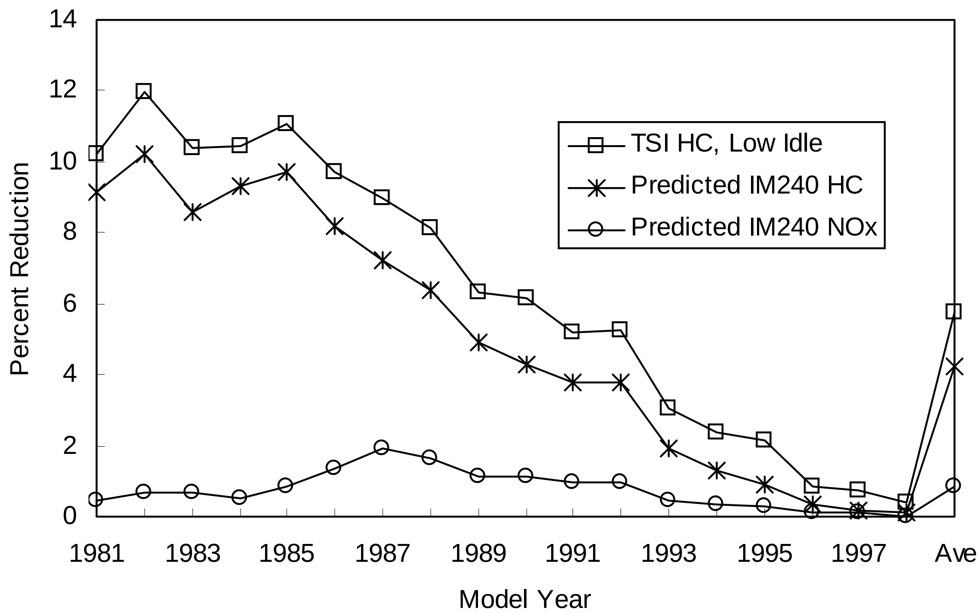


Figure 5-22. Percent Reduction of Measured TSI HC and Predicted IM240 HC and NO_x

Figures 5-23 and 5-24 show the initial fail rate in each of the three states, and the average number of tests from the initial failed test to the final passed test. From Figure 5-23, it is apparent that State 2 has a high failrate when compared to the other two states. However, Figure 5-24 shows that the average number of tests required to progress from the first failed test to the final test is comparatively low in State 2. Possible explanations might be that repairs made in State 2 are not holding between tests and must be repeated each year, or that motorists are learning to “beat the test” after they have failed once. Whatever the reason, the combination of information given by Figures 5-23 and 5-24 should be used by a state to highlight areas for further investigation of an I/M program.

5.3.4.1 Recommended Best Practice

The analyses presented in this section are very qualitative, since differences between the programs under comparison are not accounted for (i.e., climate, fuel, altitude, test type). A correlation to convert all tests to an equivalent basis could be used, but without additional corrections for program differences, results will still be qualitative. The use of a model like MOBILE would be required to completely bring the results of the three areas to an equivalent basis. However, Figure 5-20, 5-23, and 5-24 should be used as a tool to identify discrepancies in emissions reductions trends between different states. Differences in trends may indicate a weakness in one of the programs that would not appear without comparison to another program.

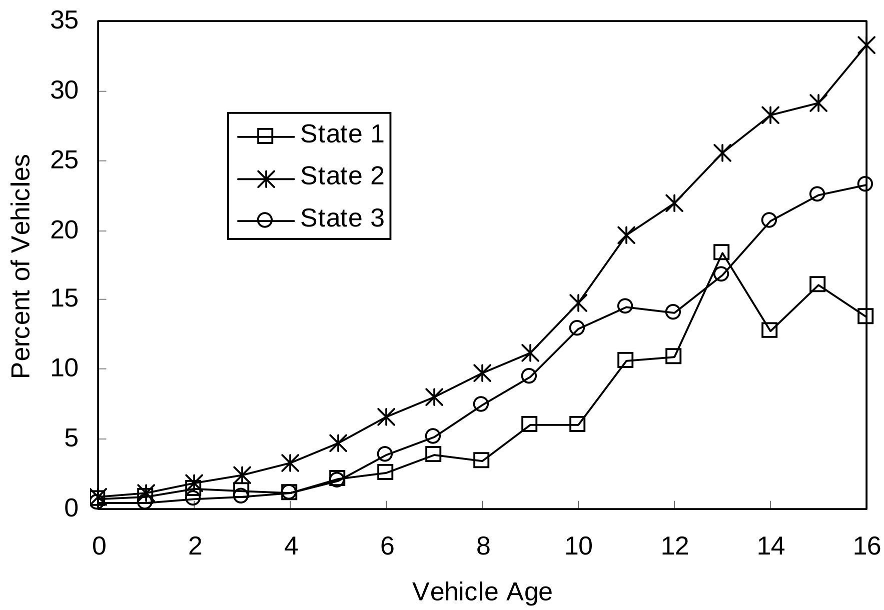


Figure 5-23. Initial Fail Rate

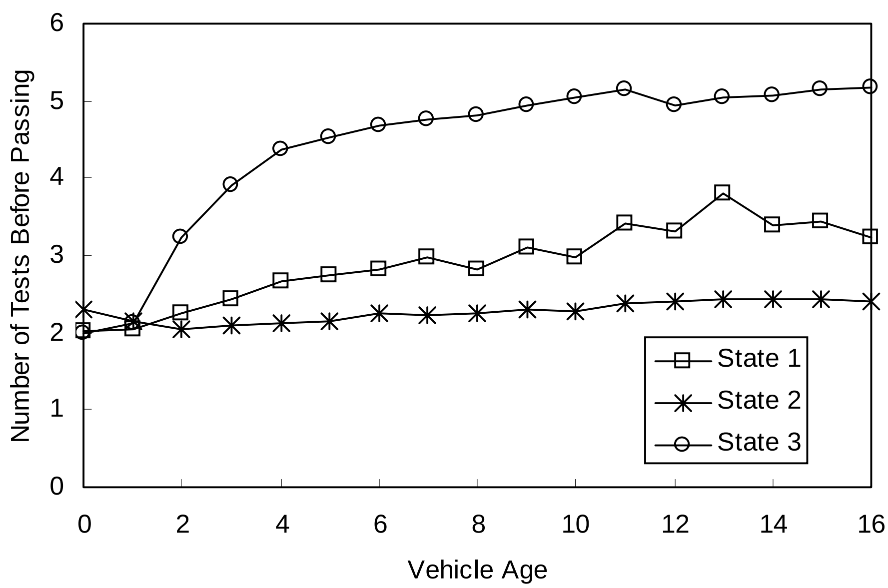


Figure 5-24. Average Number of Tests To Pass

5.3.5 Tracer Vehicles

The data analysis methods described in Section 5.3 include useful tools for understanding the effects an I/M program is having on one fleet, and some basic methods for comparing fleet-average results to different fleets are provided. A different approach for comparing fleets would be to select “tracer vehicles.” Tracer vehicles are make/model/engine combinations chosen because of their prevalence in most areas. Emissions comparisons from fleet-to-fleet based on these tracer vehicles would be used to highlight differences between the fleets. Since all tracer vehicles of a given make/model/engine combination should have had the same emissions levels when they were new, differences as they age may be attributable to the I/M program. Comparing the I/M program effects on tracer vehicles instead of on the entire fleet eliminates the effects of different fleet composition and allows a more direct comparison.

This comparison is made below, using data from States 3 and 5, both of which administer a yearly IM240 test. The IM240 HC emissions distributions of three late model year make/model/engine combinations from each state are presented in Figure 5-25. Model year 1994 vehicles are used, since that is the newest model year that is fully represented in both of the state data sets. The three make/model/engine combinations, which are the same for both states, were chosen as the three that are most heavily represented in both of the fleets. These distributions are intended to represent the emissions of vehicles in the two states when they are new. Similarly, IM240 HC emissions distributions for three 1984 make/model/engine combinations for both states are shown in Figure 5-26. These represent vehicles that have been affected by the I/M program as they have been operated within the state over many years. The make/model/engine combinations and sample sizes are listed in Table 5-3. The sample sizes are reasonably large, ranging from 400 to 1500 vehicles per combination.

Figure 5-25 does appear to show some differentiation between the two fleets. The curve for vehicles of Combination 4, State 3 is shifted to the right of the curve for Combination 4 vehicles from State 5; the same is true for Combinations 5 and 6. This may indicate that regional effects such as altitude, fuel, and climate are causing the emissions distributions of the two states to differ, since these vehicles are nearly new and should not be affected by deterioration. However, the emissions levels for these new vehicles are very low, and in State 3 a sharp peak is seen instead of a smooth distribution. Since the values are so clustered, it seems possible that this peak is located at some minimum measurable concentration. Both State 3 and State 5 allow a fast-pass and then use the results to project full test scores and many of the newest model year vehicles achieve a fast-pass at the earliest allowable second of the test, making it difficult to accurately project full-test emissions. Difficulties associated with projecting full test scores from fast-pass results were mentioned earlier in Section 5.3.1. As a result, it is not entirely clear whether the difference between the two distributions is real or is an artifact of data collection and processing methodology. The emissions distributions for the 1984 vehicles in Figure 5-26 do not show this effect. The distribution traces for each combination are now very similar for either state. If the new-vehicle differences shown in Figure 5-25 did represent real emissions differences between the two areas, the lack of difference in Figure 5-26 would indicate that greater deterioration is occurring in State 5 than in State 3, since the State 5 emissions distributions are lower when the vehicles are new but not lower after the vehicles have aged.

Table 5-3. Make/Model/Engine Combinations for States 1 and 5

Combination	**Model Year**	**Make**	**Model**	**Engine Displacement [L]**	**Count, State 1**	**Count, State 5**
1	1984	Chev.	Cavalier	2.0	556	1339

2	1984	Chev.	Celebrity	2.8	422	745
3	1984	Ford	Tempo	2.3	435	615
4	1994	Ford	Escort	1.9	722	1843
5	1994	Honda	Accord	2.2	1220	685
6	1994	Toyota	Corolla	1.6	624	1490

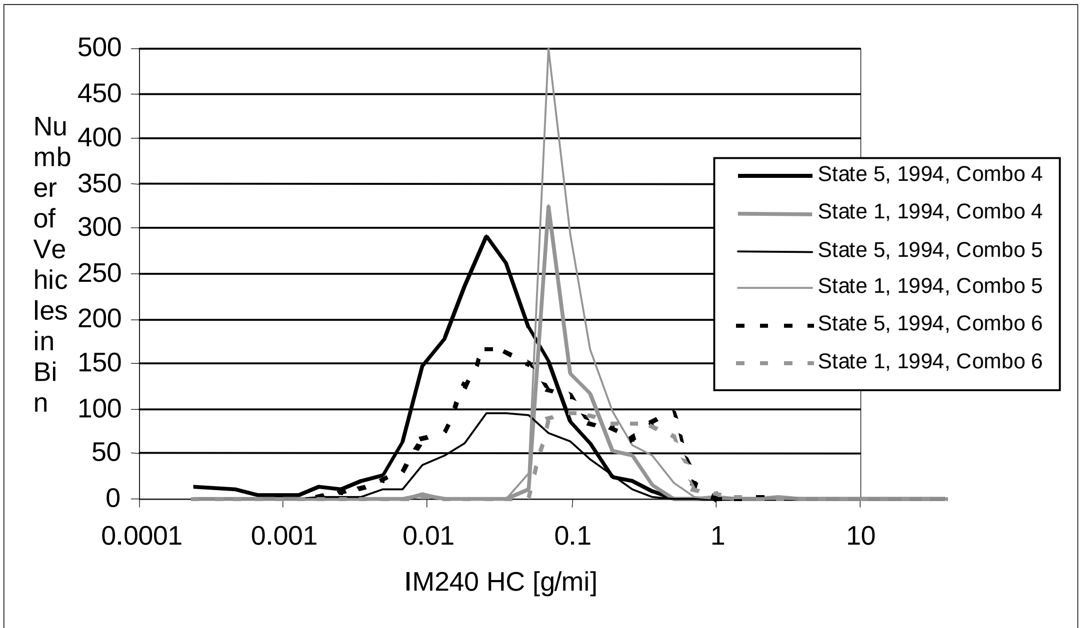


Figure 5-25. IM240 HC Emissions Distributions for 1994 Vehicle Combinations

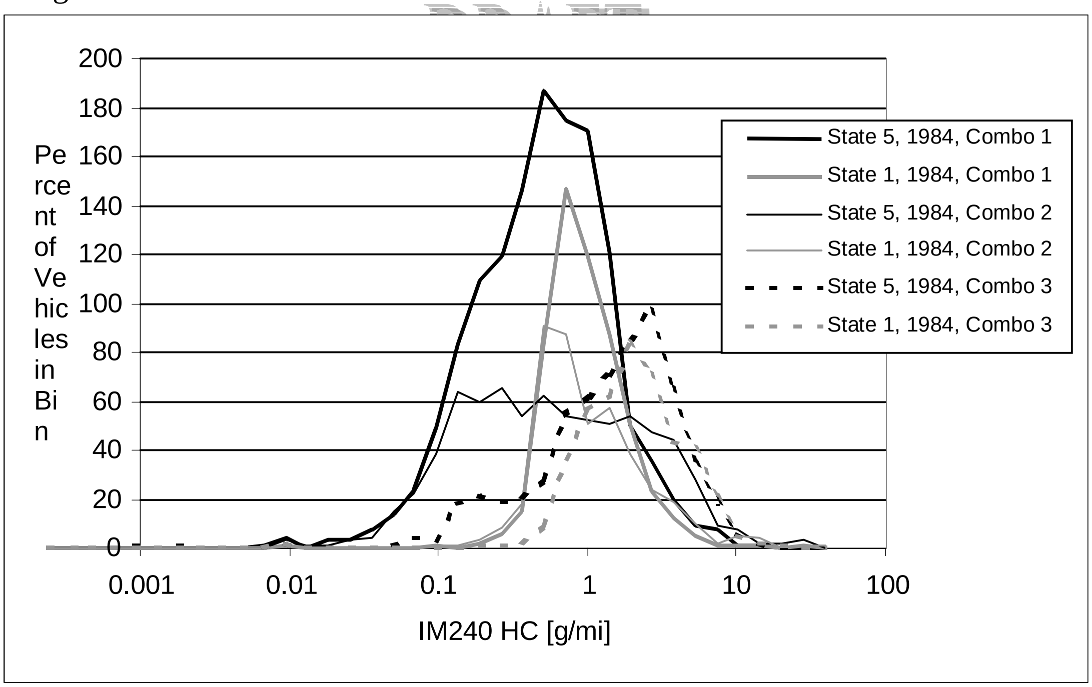


Figure 5-26. IM240 HC Emissions Distributions for 1984 Vehicles Combinations

5.3.5.1 Recommended Best Practices

The concept of tracer vehicles could be a valuable tool for benchmarking the results of one program against another, assuming states are willing to coordinate their efforts and support this concept. The effectiveness of I/M programs from two different areas may be compared using the emissions distributions of the tracer vehicles, without the need for correcting for regional differences (altitude, fuel, etc). However, fast-pass/fast-fail options seem to obscure the results. Additional work will be needed to determine the value of this type of analysis.

5.4 Evaporative Emission Reductions

This section outlines recent data from EPA and CRC studies to develop a first order estimate for the possible emission reductions from evaporative emissions for vehicles identified and repaired for evaporative emission control problems. Very small numbers of vehicles are included in these studies and clearly more data is needed to more accurately quantify the possible emission reductions from gas cap and pressure test results. A detailed discussion about the methodology used in the EPA and CRC studies is available elsewhere[23].

5.4.1 Estimate of Single Vehicle Gas Cap I/M Benefit

The first study (Reference XXX), conducted in 1997/1998 by Automotive Testing Labs, was performed under an EPA contract on vehicles recruited from the Arizona I/M Program. Vehicles were tested with the following conditions:

Fuel RVP of 6.3 psi.
38 hour 72-96°F diurnal.
1 hour hot soak at 95°F.
3x LA4 Running loss at 95°F.

The volatility of the fuel is described by Reid Vapor Pressure (RVP) with units of pounds per square inch. Diurnal emissions were measured with a 38-hour ambient temperature profile made up of a 72°F to 96°F increase, a 96°F to 72°F decrease, and another 72°F to 96°F increase. The specific temperatures are taken from the EPA 72-hour enhanced diurnal profile used for diurnal evaporative emissions testing. The hot soak emissions were measured for one hour with an ambient temperature of 95°F following an FTP driven at 95°F. Running losses were measured while driving three consecutive LA4 cycles. An LA4 cycle is the 1372-second cycle used for the first two bags (cold start + warm stabilized) of the FTP.

These conditions were considered appropriate for Arizona conditions in 1997/1998 because they were thought to be representative of in-use evaporative emissions generation. They are different from new vehicle certification test conditions, which are designed to be severe test conditions under which new vehicle emission control hardware and purge strategies must control emissions. Data from the EPA study includes the before-repair and after-repair evaporative emissions of the 26 vehicles tested. The estimated total evaporative emissions reduction was calculated using the 24-hour diurnal, hot soak, and running loss measurements before and after repair and using assumptions of 3 hot soaks per day and 30 miles traveled per day for each test vehicle. Evaporative emissions reductions for pressure, purge, or fuel cap based repairs are assigned for all the vehicles considered in the study.

Table 5-4 presents a summary of the emission reductions associated with the following categories:

- Pressure system repair;
- Purge system repair; and
- Gas cap repair.

Table 5-4. Summary of EPA WA1-8

	Evaporative Emission Reductions (g/mile)					
	Running Losses Only			All Evap Emissions (Hot Soak, Diurnal, and Running Losses)		
	Repairing Pressure Problems	Repairing Purge Problems	Repairing Gas Cap Problems	Repairing Pressure Problems	Repairing Purge Problems	Repairing Gas Cap Problems
Carbureted Vehicles	0.88 (1)*	2.50 (1)	0.56 (3)	1.01 (1)	2.50 (1)	1.07 (3)
Fuel Injected Vehicles	1.83 (10)	-	3.90 (9)	2.47 (10)	-	4.37 (9)
All Vehicles	1.75 (11)	2.50 (1)	3.07 (12)	2.34 (11)	2.50 (1)	3.55 (12)

*Numbers in parenthesis denote vehicle sample size.

The table shows the emission reductions for carbureted and fuel-injected vehicles. In addition, running loss and total evaporative emissions reductions are shown. For this analysis the gas cap emission benefits are most important because State 1 uses a gas cap test to identify gas cap failures. The sample size for the data represented in Table 5-4 is small. EPA, California ARB, California BAR, and CRC all plan to conduct more SHED tests to quantify the change in evaporative emissions due to evaporative system repair. However, SHED tests are expensive and time consuming (compared to IM240 tests), so not as many tests are performed. The data in Table 5-4 is a best-estimate of evaporative emissions given the available data at this time.

The scatter of before- and after-repair evaporative emissions can provide additional insight beyond simply comparing means. Figures 5-27 and 5-28 show plots of the total evaporative emissions after repair versus before repair. The plots use different symbols for repair type and fuel metering system type, respectively. Logarithmic scales are used so that the data scatter can be seen more clearly. Figure 5-27 indicates that the scatter of data points for pressure fails and fuel cap fails is about the same. On the other hand, Figure 5-28 indicates that different but overlapping regions may characterize different fuel metering types. The 1:1 line in the figures is drawn to assist the readers in interpreting the data. Data points below the line represent vehicles whose emissions prior to repair were higher than emissions after repair. The magnitude of the distance of the points away from the 1:1 line denotes the amount of emission reduction caused by evaporative system repair.

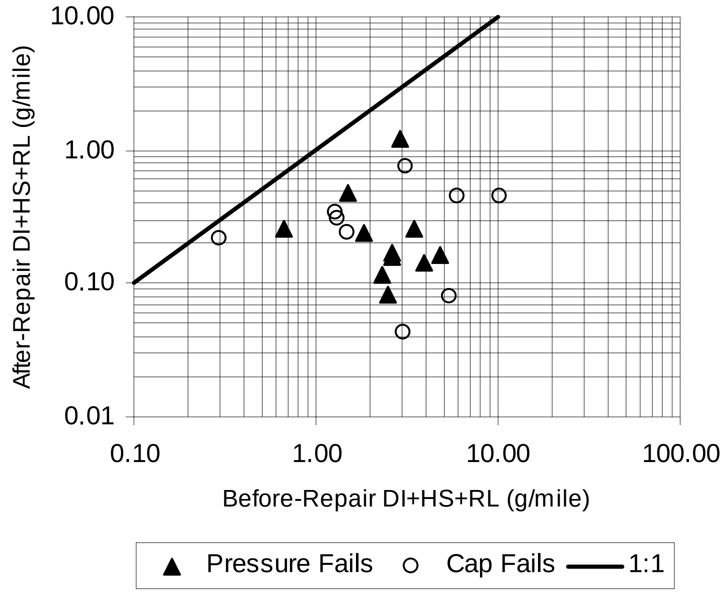


Figure 5-27. Effect of Repair Type on Evaporative Emissions System Repairs

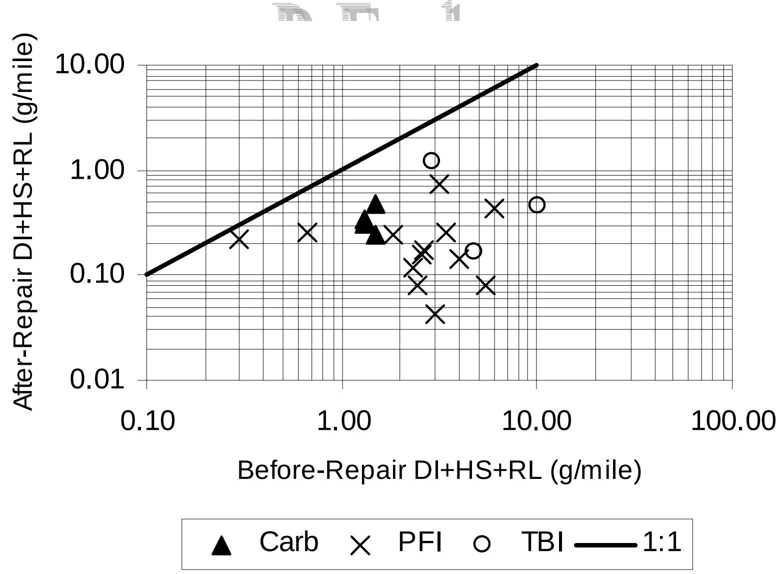


Figure 5-28. Effect of Fuel Metering Type on Evaporative Emissions Systems Repairs

In a second study (CRC Project E-35 Reference XXX), testing was conducted by CRC using approximately 7 psi RVP and 1 LA4 to measure running losses in the SHED. Thus, test conditions for this study were also different from new vehicle certification tests, and they were different from the test conditions used in the EPA study. Procedures, vehicle recruitment, and small sample sizes all contribute to the numerical results in emissions from the two studies. Only running losses were measured in this study. The vehicle testing was also conducted by Automotive Testing Laboratory (ATL). Running loss emissions for 29 vehicles were diagnosed with faulty gas caps or filler neck problems. The average running loss emissions of vehicles prior to gas cap repair was 0.36 g/mile.

In the CRC study, only running losses were considered in the evaporative emissions estimate. The fuel cap related failure was diagnosed but post-repair emissions are not available. A comparison of the measured before-repair running loss emissions from the CRC and EPA studies is shown in Figure 5-29. A logarithmic scale was used to help see the data scatter more clearly. The figure shows that pre-repair running losses were about 7 times higher in the EPA study, but the amount of scatter was comparable in the two studies. Higher running loss emissions can be expected from the EPA study, which used 3 LA-4s for its test cycle, in comparison with the CRC study, which used 1 LA-4.

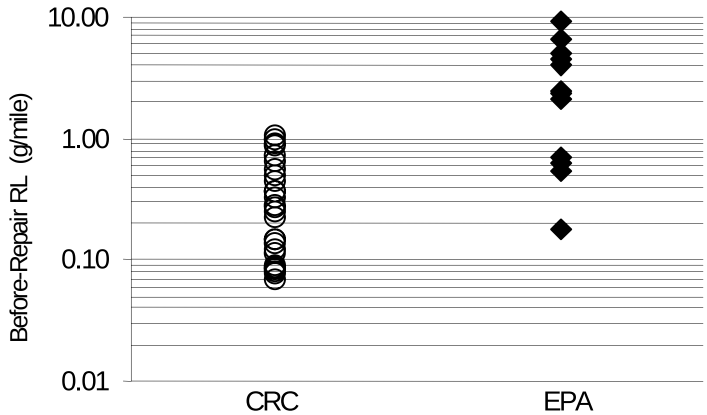


Figure 5-29. Comparison of Before-Repair Running Losses from EPA and CRC Studies

To calculate the emission reductions from the CRC data, it was assumed that the average post-repair running loss emissions for the CRC test vehicles is the same as that for the EPA test vehicles, i.e. it was assumed the CRC vehicles could be repaired to the same levels as the repaired vehicles in the EPA study. This is shown in Table 5-5. Average post-repair running loss emissions were calculated from the EPA data described in Table 5-4. Finally, the estimated running loss emission reductions associated with gas cap repair is calculated by difference for the CRC sample in Table 5-5. Because the CRC test fleet and the EPA test fleet are not the same, the subtraction of the EPA post-repair average from the CRC pre-repair average provides large uncertainty. However, failure to subtract some estimate of post-repair emission values would surely over-estimate the size of emissions reductions due to gas cap repairs.

Table 5-6 presents calculations to estimate the total evaporative emissions reductions for gas cap repair. The results of both studies are combined to arrive at estimated reductions. The top of Table 5-6 presents the emission reduction estimates from the two studies. The table shows the measured running loss reductions from the EPA study, the estimated running loss reductions from the CRC study, and the total evaporative emissions reductions from the EPA study. Total evaporative emission reductions are not available from the CRC study. The average running loss reductions for carbureted and fuel-injected vehicles are calculated by averaging the average running loss reductions for the CRC and EPA studies with the number of vehicles (in parentheses) as weighting factors. Then, total evaporative emissions (for hot soak, diurnal, and running losses) are calculated by using the weighted running loss estimates instead of only the EPA estimates. As shown, the best estimate of possible evaporative emissions reductions for vehicles that fail the gas cap test are 1.00 g/mile for carbureted emissions and 3.25 g/mile for fuel-injected vehicles. It would be beneficial to provide error bars on these estimates but due to the small sample it is difficult to quantify these with any degree of confidence. EPA and California are conducting additional studies to improve these estimates.

It is readily recognized that the sample sizes used to arrive at these estimated evaporative emissions are small; however, these are the only measurements available to states to make these estimates. Accordingly, the uncertainties of these estimates are large.

Table 5-5. Estimate of Running Loss Emission Reductions from Gas Cap Repairs in CRC Study (g/mile)

	Carbureted Vehicles	**Fuel-Injected Vehicles**
CRC Pre-Repair Running Loss Average	0.538 (25)	0.334 (4)
EPA Post-Repair Running Loss Average	0.059 (5)	0.088 (20)
Estimated CRC Running Loss Reduction	0.48	0.25

*Numbers in parentheses denote vehicle sample size.

Table 5-6. Total Evaporative Emission Reduction Calculation for Gas Cap Repairs

	Running Loss Reductions for Gas Cap Repair		Total (RL + DI + HS) Evap Reductions for Gas Cap Repair	
	EPA	CRC	EPA	CRC
Carbureted Vehicles	0.56 (3)	0.48 (25)	1.07	-
Fuel-Injected Vehicles	3.90 (9)	0.25 (4)	4.37	-

*Numbers in parenthesis denote vehicle sample size.

	Weighted Running Loss Reductions
Carbureted Vehicles	0.49 g/mile
Fuel-Injected Vehicles	2.78 g/mile

	Estimated Total Evaporative Reductions
Carbureted Vehicles	1.07 – 0.56 + 0.49 = 1.00 g/mile
Fuel-Injected Vehicles	4.37 – 3.90 + 2.78 = 3.25 g/mile

5.4.2 Fleet I/M Evaporative Benefit

In the last section, two studies were discussed to estimate the evaporative emissions benefit associated with the repair following a gas cap test failure. Before this data is used to project fleet benefits, several issues need to be discussed. These include the following:

- Emissions Deterioration
- Repair Effectiveness
- Collateral Defects

Evaporative Emissions Control Technology/OBD.

Emissions Deterioration: The previous section estimates the emissions reduction that are achieved immediately after repair. As vehicles go back into their normal usage following this repair, the emissions can creep up as the emissions control system degrades. Emissions can also increase if the fuel cap is not tightened or replaced following refueling. The frequency of occurrences of these events is not fully known at this time.

Repair Effectiveness: In the real world not all identified defects get repaired. Based on conversations with state I/M staff, it is assumed that 90% of the emissions reductions estimated for the roadside fleet associated with gas cap repair will actually be realized by the I/M program; however, this estimate is not based on any observed data. As more VID and roadside data is collected this assumption will be re-considered.

Collateral Defects: Vehicles which have a gas cap defect can also have other evaporative emissions control problems. In a small sample of roadside data in which 1992 and older vehicles were considered, 62.6% of the vehicles that failed the gas cap test also failed the fuel evaporative pressure test. Since the pressure test is conducted after removing the gas cap, this implies that these vehicles had other pressure leaks in addition to a gas cap defect. It is possible that some of these vehicles, which have gas cap and pressure defects, would benefit from a gas cap repair. For this analysis, State 1 assumed that 70% of the possible emissions reduction from gas cap repair will be achievable. This implies that 30% of the emissions reduction will be negated due to other evaporative emissions problems with the vehicle.

Evaporative Emissions Control Technology/OBD: Newer vehicles have more robust evaporative control systems and have fewer defects. In addition, 1996 and newer vehicles with OBDII system checks set an engine malfunction indicator light (MIL) if the evaporative control system fails the on-board test. The evaporative system monitors were optional/experimental on 1996-1997 Federal vehicles; monitors were required on at least 20% of 1996 model year vehicles and on at least 40% of 1997 model year vehicles. It is expected that future gas cap benefits may be reduced as more vehicles with OBD systems penetrate the fleet. New issues with OBD systems may occur over time but this issue will need to be studied as OBD equipped vehicles age. In this analysis, no gas cap emissions benefit is assumed for 1996 and newer vehicles.

Fleet Emissions Reduction Calculation: Figure 5-30 shows the gas cap failure rates observed in State 1 roadside data. Vehicles which had undergone an inspection were observed to have a lower fail rate than vehicles which were tested prior to their inspection. The results of the gas cap repair benefit from Table 5-6 and fail rates from Figure 5-30 were used to estimate the fleet emissions benefit. This calculation is shown in Table 5-7. The table shows the evaporative emission calculations for each model year. The fail rates shown in Figure 5-30 and repeated in this table are calculated from the roadside data. The percent of fuel injected vehicles are taken from EPA estimates. The evaporative benefit for carbureted and fuel injected vehicles is calculated as follows:

Evaporative emissions benefit for carbureted vehicles in any model year =

$$(1\text{-}FI) * (FR_B - FR_A) * EV_{carb}$$

Where:

FI = Fraction of fuel injected vehicles in the model year

$FR_{B,A}$ = Failure rate for roadside vehicles before and after Smog Check

EV_{carb} = Evaporative benefit for carbureted vehicles estimated in Table 3.

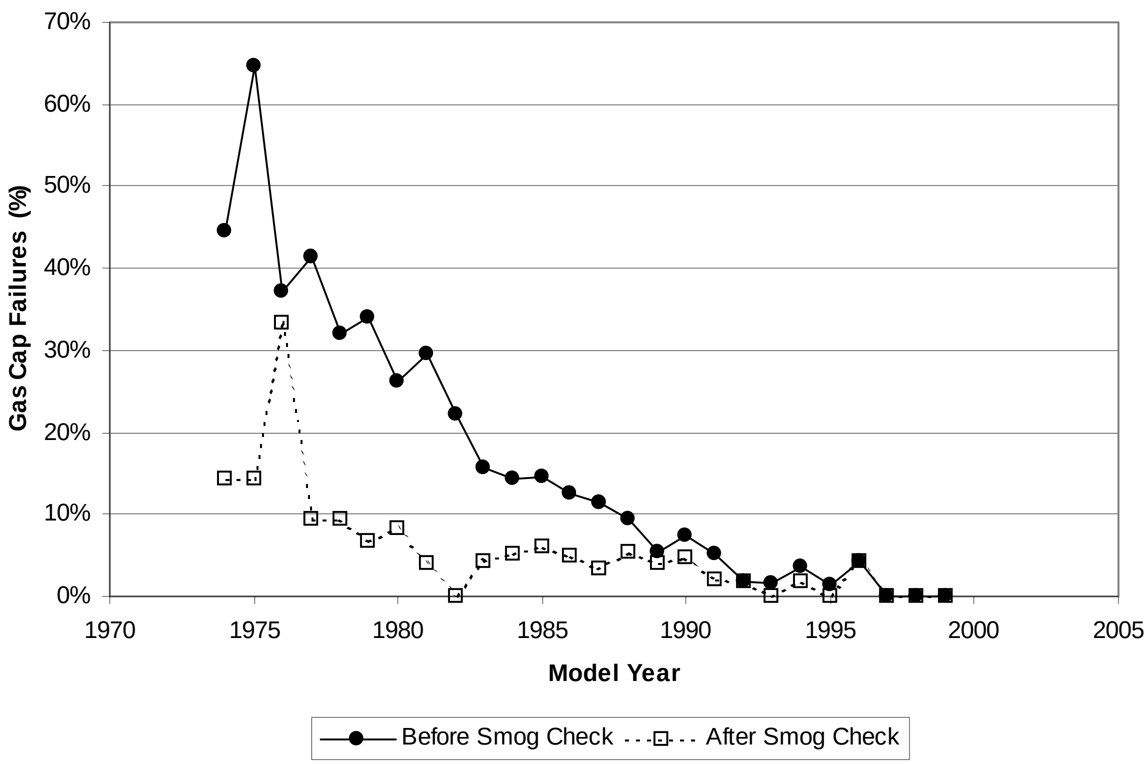


Figure 5-30. Gas Cap Failure Rates from California Roadside Data

The emissions benefit associated with fuel injected vehicles is calculated in a similar fashion using the emissions estimates for fuel injected vehicles from Table 5-6. The total model year evaporative emissions estimate is then calculated and multiplied by the travel fraction to estimate the weighted model year emissions benefit. The total calendar year 1999 estimate is then calculated by summing the emissions benefit for all the model years. This is calculated to be 0.076 g/mile in Table 5-7. The net evaporative emission estimate is then calculated by using the assumptions discussed above that 90% of the emissions associated with vehicles with gas cap defects are actually repaired and that 30% of the emissions benefit is negated due to collateral defects. The net evaporative estimate was hence calculated to be 0.048 g/mile.

Other states that do not conduct roadside tests could use gas cap fail rates in their states and compare them to a No-I/M area gas cap fail rates. Data for No-I/M areas would have to be developed by EPA and other stakeholders in order for this to be viable.

Table 5-7. Calculation Summary for Estimating California Fleet Evaporative Emissions Benefit for Gas Cap Repairs

Model Year	Travel Fraction*	Gas Cap Fail Before	Gas Cap Fail After	% of Fuel-Injected Vehicles	Evap Benefits for Carbureted Vehicles (g/mile)	Evap Benefits for Fuel-Injected Vehicles (g/mile)	Total Model Year Evap Benefits (g/mile)	Weighted Evaporative Emissions Benefit
1974	0.00326	44.44%	14.29%	0.000%	0.30159	0.00000	0.30159	0.00098
1975	0.00275	64.71%	14.29%	0.000%	0.50420	0.00000	0.50420	0.00139
1976	0.00428	37.04%	33.33%	0.000%	0.03704	0.00000	0.03704	0.00016
1977	0.00632	41.27%	9.38%	0.000%	0.31895	0.00000	0.31895	0.00202
1978	0.00816	32.00%	9.38%	0.000%	0.22625	0.00000	0.22625	0.00185
1979	0.00979	33.94%	6.82%	0.000%	0.27127	0.00000	0.27127	0.00266
1980	0.00857	26.17%	8.33%	0.000%	0.17835	0.00000	0.17835	0.00153
1981	0.01000	29.57%	4.08%	9.000%	0.23190	0.07454	0.30644	0.00306
1982	0.01275	22.07%	0.00%	16.800%	0.18361	0.12050	0.30411	0.00388
1983	0.01622	15.76%	4.30%	27.100%	0.08354	0.10093	0.18447	0.00299
1984	0.02713	14.33%	5.16%	39.200%	0.05575	0.11682	0.17258	0.00468
1985	0.03274	14.63%	6.06%	51.500%	0.04158	0.14350	0.18508	0.00606
1986	0.03907	12.50%	5.00%	67.600%	0.02430	0.16478	0.18908	0.00739
1987	0.04284	11.46%	3.37%	74.100%	0.02094	0.19469	0.21563	0.00924
1988	0.04621	9.40%	5.26%	89.900%	0.00418	0.12092	0.12510	0.00578
1989	0.05222	5.38%	4.07%	87.200%	0.00167	0.03708	0.03875	0.00202
1990	0.04967	7.30%	4.63%	98.100%	0.00051	0.08513	0.08564	0.00425
1991	0.05314	5.25%	1.91%	99.800%	0.00007	0.10825	0.10832	0.00576
1992	0.04733	1.80%	1.80%	99.800%	0.00000	0.00000	0.00000	0.00000
1993	0.05763	1.63%	0.00%	100.000%	0.00000	0.05285	0.05285	0.00305
1994	0.06222	3.62%	1.87%	100.000%	0.00000	0.05701	0.05701	0.00355
1995	0.07303	1.42%	0.00%	100.000%	0.00000	0.04599	0.04599	0.00336
1996	0.06497	4.17%	4.17%	100.000%	0.00000	0.00000	0.00000	0.00000
1997	0.08405	0.00%	0.00%	100.000%	0.00000	0.00000	0.00000	0.00000
1998	0.10832	0.00%	0.00%	100.000%	0.00000	0.00000	0.00000	0.00000
1999	0.07732	0.00%	0.00%	100.000%	0.00000	0.00000	0.00000	0.00000
						Evap benefit g/mile =		**0.07564**
						Discount for repairs =	**90.0%**	**0.06808**
						Discount for Collateral Defects =	**30.0%**	**0.04766**

*From California BAR May 1999 Travel Fraction Calculator.

5.4.3 Other Evaporative Control Measures

In addition to pressure tests and gas cap tests, recent CRC and EPA studies have also led researchers to identify and repair liquid leaking vehicles (Reference XXX). The CRC study, CRC-E35, has pointed towards the existence of a very small fraction of vehicles which can be designated as liquid leakers. Drops of fuel are seen to be leaking from these vehicles. Experienced mechanics can usually identify these vehicles due to the strong gasoline smell emanating from these vehicles. California BAR is developing a testing protocol to identify, repair, and quantify the emission reductions possible from repairing such vehicles. EPA's MOBILE6 model also includes these vehicles in the fleet and includes estimates of both the frequency and the evaporative emissions estimates of these vehicles. However, procedures for quantifying the emission benefits realized by identifying and repairing these vehicles must still be developed.

6. Summary

A number of methods for estimating I/M program effectiveness using in-program data were outlined in this guidance. Effort was made to document, reference or provide examples for data collection procedures, QA/QC protocols, analysis methods, and sources of error or possible bias associated with a given method; however, it is recognized that improvements to the methods outlined in this document will continue to evolve. Therefore, it is strongly recommended that any state considering the use of in-program data for program evaluation purposes work closely with their respective regional EPA office and the Office of Transportation and Air Quality to ensure the most up-to-date practices are incorporated into the evaluation. Furthermore, states interested in using in-program data for program evaluation must recognize the need within their own agencies to develop a minimum level of expertise with the technology and procedures to ensure reliable data are collected and analyses performed.

It should also be recognized, given the difficulties associated with I/M program evaluations, that an evaluation based on both out-of-program data (e.g. RSD or roadside pullovers) and in-program data, will provide a more accurate estimate of overall program performance than simply relying on one method alone.

7. References

---

[1] IM240 & Evap Technical Guidance, April 2000, EPA420-R-00-0007 available on-line at www.epa.gov/oms/im.htm

[2] ASM Technical Guidance DRAFT, July 2000 available on-line at www.epa.gov/oms/im.htm

[3] Clean Air Act, 1970

[4] Clean Air Act Amendments, 1977

[5] EPA Guidance, 1978

[6] Clean Air Act Amendments, 1990

[7] IM Rule, November 5, 1992

[8] National Highway Systems Designation Act, 1995

[9] EPA Rule, January, 1998

[10] EPA Memo, October 30, 1998

[11] Klausmeier, Evaluation of Test Data Collected in 1999 from Connecticut's I/M Program, "DRAFT", July 2001.

[12] New York State Enhanced I/M Program Evaluation Report for the Period of 11/16/98 to 12/31/98, New York Department of Environmental Conservation, January 2001.

[13] T.H. DeFries, C.F. Palacios, and S. Kishan, "Models for Estimating California Fleet FTP Emissions from ASM Measurements," December 25, 1999, Eastern Research Group, Inc., Austin, Texas, ERG report BAR-991225

[14] T.H. DeFries and D.A. Westenbarger, "Evaluation of I/M Programs and Modeling Techniques to Predict Fleet Average FTP and IM240 Emission Rates," presented at Tenth CRC On-Road Vehicle Emissions Workshop, March 27-29, 2000, Coordinating Research Council, Atlanta, Georgia.

[15] ECOS/STAPPA/EPA Inspection and Maintenance Workgroup, Final Meeting Summary for March 25, 1998, Workgroup Meeting and Attached Background Materials.

[16] T.C. Austin, L.S. Caretto, T.R. Carlson, and P.L. Heirigs, "Development of a Proposed Procedure for Determining the Equivalency of Alternative Inspection and Maintenance Programs," prepared for U.S. EPA, Regional and State Programs Division, by Sierra Research, Inc., Sacramento, California, November 10, 1997.

[17] T.H. DeFries, A.D. Burnette, S. Kishan, and Y. Meng, "Evaluation of the Texas Motorist's Choice Program," prepared for the Texas Natural Resource Conservation Commission, by the Eastern Research Group, Inc., Austin, Texas, May 31, 2000, ERG report TNRCC-000531.

[18] S. Kishan, C.F. Palacios, "Comparison of California's I/M Program with the Benchmark Program," prepared for California Bureau of Automotive Repair, by the Eastern Research Group, Inc., January 19, 2000.

[19] A.W. Ando, W. Harrington, and V. McConnell, "Estimating Full IM240 Emissions from Partial Test Results: Evidence from Arizona", J. Air & Waste Management Association (49), 1153, 1999.

[20] T. Wenzel, "Analysis of LBNL's Short Test Adjustment Factors," Lawrence Berkeley National Laboratory, April 1998

[21] P. McClintock, Environmental Analysis private communication with J. Lindner, US EPA, 2000.

[22] P. McClintock, Environmental Analysis private communication with J. Lindner, US EPA, 2000.

[23] "Evaporative Emissions Impact of Smog Check," report by California BAR and ERG, September 2000.

Appendix A: Development of a Model to Predict IM240 Emissions Concentrations from Two-Speed Idle Data

Although the performance standard for an I/M Program is the IM240 test, many states choose to administer different types of tests such as the two-speed idle test (TSI) or the ASM test. Sierra Research proposed that if results of alternative types of tests are to be compared to baseline program results or results from other states, models must be built to predict IM240 emission rates from the measured alternative test emissions concentrations. This appendix contains an outline of the specific procedures to develop such a correlation, based on work done by Eastern Research Group for the Texas Natural Resource Conservation Commission.[1] Since Texas uses a two-speed idle test, development of a correlation between IM240 measurements and a different type of test will differ in the details of the fundamental procedures outlined below.

The correlation of TSI and IM240 results is based on emissions data from a sample of Texas vehicles that received both the TSI and IM240 tests. Procedures for selecting a suitable vehicle sample, developing a correlation model, testing the model for bias, quantifying uncertainty, and the limitations of applying the model to the fleet will be described.

<u>A.1 Data Collection</u>

Two types of data were acquired for the development of the dataset on which models can be developed: two-speed idle (TSI) data, and IM240 dynamometer data. The two-speed idle measurements were made at two I/M inspection stations according to normal station procedures. Since valid data is critical for successful model development, the TSI instruments were calibrated and zeroed as usual, and then independently checked at the beginning of each workday with zero and span audit gases separate from the I/M station's normal supply. IM240 tests were performed using a portable dynamometer located just outside the I/M station. All equipment was calibrated and operated according to EPA specifications.

Selection of vehicles to participate in the test program was based on a stratified random sampling scheme using model year group, TSI test results, and vehicle type. Stratification is used to prevent selection of predominantly new, relatively low-emitting vehicles. While a stratified random sample does not represent the vehicle distribution in the fleet, it does provide a model-building dataset containing the full range of emissions levels.

Model year groups are used as a stratification category instead of individual model years to reduce the number of stratification levels. For the TNRCC work, four model year groups were used: 1981 to 1984, 1985 to 1988, 1989 to 1992, and 1993 to 1997. Also, for each of the four TSI measures (high-speed idle HC and CO and low-speed idle HC and CO), bins were created based on these model year groups and TSI concentration groups. Historical Texas TSI data were used to define TSI concentration groups so that each represented approximately a quintile of the TSI distribution for each model year group. The goal of vehicle selection was to achieve an equal number of vehicles in each model year group/TSI concentration group bin for each type of TSI. In addition, the vehicles in each bin were targeted to be 64% passenger cars and 36% light-duty trucks

(trucks, vans, MPVs, SUVs) as in the Texas fleet. When the TSI results, model year, and vehicle type of a vehicle at the I/M station indicated that the vehicle would be a suitable candidate for the stratified sample, the vehicle owner was offered an incentive in exchange for allowing the vehicle to be receive an IM240 test following the TSI test. A smaller stratified sample of repeat two-speed idle measurements was also collected to cover the range of HC and CO low-speed idle and high-speed idle. Additional incentive was offered to the vehicle owner for allowing a second IM240 and TSI test to be performed.

TSI measurements were performed using the I/M station's BAR90 analyzers. All TSI and I/M CO and CO_2 measurements were determined by non-dispersive infrared (NDIR). IM240 NO_x was determined by chemiluminescence. In the case of hydrocarbons, the IM240 hydrocarbon was measured by flame ionization detector (FID) and the TSI hydrocarbon was measured by NDIR. Major differences in response factors to different types of hydrocarbon compounds are known to exist between FID and NDIR. Therefore, proper application of the models that were developed requires that TSI hydrocarbon be measured by NDIR.

The overall goal for the TSI/IM240 data set was to acquire test pairs of test results for 800 vehicles, divided among the four model year groups, five emissions level quintiles, and two vehicle types.

A.2 Model Development

The steps involved in developing the models for TNRCC were:

- General quality assurance of the raw data including review of the TSI analyzer calibration and gas audit results;

- Data preparation consisting of humidity corrections for IM240 NO_x values, correction of TSI values for vehicle exhaust system dilution, removal of suspect observations from the dataset, and special handling for low TSI values;

- Investigation of transformations of the variables to be used in the models to make the variance across the range of values homogeneous;

- Various types of variable screening techniques to determine variables which could be expected to be important to the prediction of IM240 values and to discover any major curvature that might be present;

- Variable screening through the use of model building using ordinary least squares modeling techniques. With ordinary least squares modeling, the independent variables are assumed to have no measurement error;

- Estimation of the error variances of IM240 measurements and the error variances and covariances of TSI measurements; and

- Using the independent variables which produced the best ordinary least squares models, to develop the final models using the measurement error model building technique. In this technique, the error variances and covariances of the TSI measurements and the error variances of the IM240 measurements were used to build models which are less biased than the ordinary least squares models.

Each of these different steps in the modeling approach is discussed below.

Data Preparation

After an exhaustive quality assurance check was performed on the TSI and IM240 data, the TSI data was corrected for dilution. IM240 data does not require a dilution correction, although the NO_x values are corrected for ambient humidity when collected.

Adjustment of Low Two-Speed Idle Values

The presence of negative two-speed idle values is known to exist in the Texas VID system. Therefore, during field data collection in this project, we were aware that negative values might occur. Negative values can be expected in any instrumental measurement. Even though negative concentration values make no physical sense, it is important to remember that the output of instruments is simply a voltage or current which can have negative values. Thus, a small error in zeroing the instrument can produce negative values in the dataset. During model building, negative and zero values need to be handled appropriately to arrive at a model which is unbiased on the low concentration end.

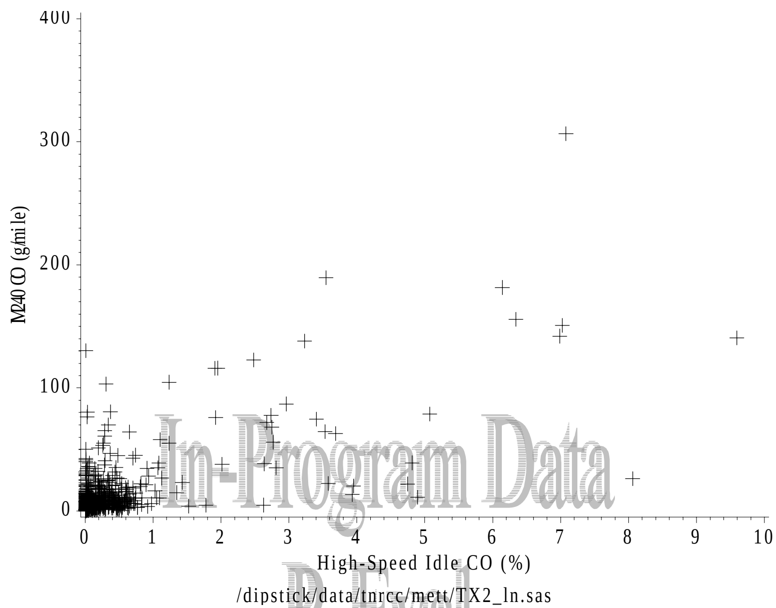
In the dataset collected in this study, no negative two-speed idle values were obtained. The smallest non-zero values reported by the TSI analyzers were 1 ppm HC and 0.01% CO. Many zero values (0 ppm HC and 0.00% CO) for two-speed idle concentrations were measured (11 low-speed idle HC zeroes, 246 low-speed idle CO zeroes, 55 high-speed idle HC zeroes, and 324 high-speed idle CO zeroes) for the modeling dataset. For model building purposes, zero two-speed idle HC values were set to 1 ppm, and zero two-speed idle CO values were set to 0.01%. These changes are well within the measurement error of the TSI method and instruments. The changes are necessary to allow logarithmic transformations of TSI values for model building purposes.

Negative and zero IM240 values were not reported on the test vehicles.

Selection of Appropriate Variable Transformations

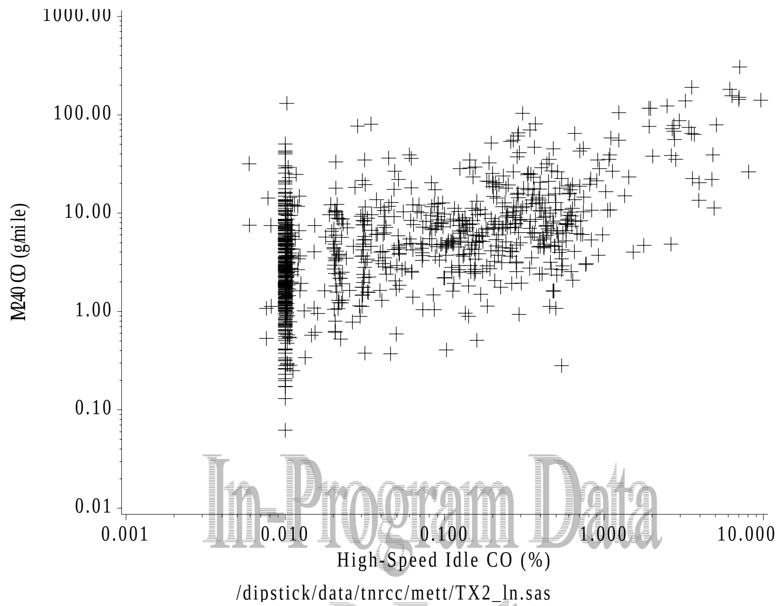
Plots of IM240 emission rates versus dilution-corrected TSI concentrations indicate that the values of both variables are highly positively skewed and the variance of any relationship between the two variables is inhomogeneous. Inhomogeneous variance means that the scatter at high emissions levels is much different than the scatter at low emissions levels. This difference in scatter can be seen in the sample plot in Figure A-1 for IM240 CO versus high-speed idle CO in linear space. The figure shows much larger scatter at high emissions than at low emissions. Another serious problem with building variables in linear space for this data is a result of the “kite and string” nature of the data. Because of the highly skewed distribution for the dependent and independent variables, as is seen in Figure A-1, any regression line will be anchored near the origin by the large number of data points there. Then, the presence or absence of the few high values on the upper right portion of the plot will influence the position of the regression line far out of proportion to their abundance in the data set.

Figure A-1. IM240 CO versus High-Speed Idle CO



Other transformations were sought to help correct these problems. The natural logarithm of both the IM240 emission rates and the TSI concentrations was chosen. Figure A-2 shows the scatter plot in log-log space for IM240 CO versus high-speed idle CO. The plot shows that the data for both variables in log space is not highly skewed and that the variance (the scatter of points) is nearly homogeneous across the range of the variables. The log-log plots for all combinations of IM240 emission rates and TSI emission concentrations were also examined.

Figure A-2. Comparison of IM240 CO and High-Speed Idle CO



Investigate Independent Variables
As the first step in model building, correlation coefficients were calculated and plots were made to investigate the relationships among the different variables in the dataset. The R^2 values were tabulated and the strongest relationships noted. The R^2 between IM240 emission rates and different variables which are candidates for predictors were also calculated.

Statistical Variable Selection Using Conventional Regression
The second step in the selection of variables to be used to predict IM240 emission rates is the development of ordinary least squares regression models. Unlike correlation coefficients and scatter plots that can only consider the influence of one independent variable at a time on the IM240 emission rate, multiple linear regression can consider the influences of many variables at the same time on IM240 emission rates.

In the process of performing ordinary least squares regression, dozens of models were created and evaluated in an effort to find the best model for predicting IM240 emission rates. The PROC REG procedure in SAS was used with the stepwise option to select input variables from the TSI measurements and vehicle characteristic descriptors. Main effects, two-factor interactions, and squared effects of the following variables were considered for inclusion as terms in the models:

High-Speed Idle HC (ppm)
High-Speed Idle CO (%)
Low-Speed Idle HC (ppm)
Low-Speed Idle CO (%)
Engine Displacement (L)
Age (year)

Truck/Car Indicator (+0.5, -0.5)
Carbureted/Fuel-Injected Indicator (+0.5, -0.5)
Oxy-Catalyst Indicator (+0.5, -0.5)
Three-Way Catalyst Indicator (+0.5, -0.5)
Exhaust Gas Recirculation Indicator (+0.5, -0.5)
Air Injection Reactor Indicator (+0.5, -0.5)

Only terms which had coefficients that were significant at the 99.9% confidence level were retained for further consideration. The terms which survived this test were then used to develop the measurement error models.

Estimation of IM240 and TSI Measurement Error

The measurement error variances of IM240 HC, CO, and NO_x and of TSI HC and CO are needed for development of measurement error models and for evaluation of the influences of measurement error on model predictions. In the context of this study, measurement error is used in the statistical sense and includes all sources of error that would cause the emissions measurement of a vehicle to be different if the vehicle were tested at different I/M stations. Correctly determining the measurement error would involve measuring the emissions of a set of vehicles at different times and at different stations and instruments. Instead of using this type of comprehensive effort, we used repeat measurements on a set of vehicles to estimate measurement error. IM240 repeat measurements were performed following each other on the same dynamometer. TSI repeat measurements were performed at the same I/M station within about one hour of each other; some repeats were performed on the same BAR90 analyzer, and some were performed on different BAR90 analyzers. In any case, the repeat measurements will under-estimate the true measurement error since variability contributions of different stations, dynamometers, and days are not present. Nevertheless, the use of estimated measurement error values is significantly better than ignoring measurement error in model development, which would essentially be assuming all measurement errors are zero.

For the emissions of each repeat-tested vehicle, the variance of each repeat pair was calculated, and then the variances for all vehicles getting repeat tests were pooled to arrive at the overall variance for the test. IM240 measurement errors for HC, CO, and NO_x were calculated using 127, 127, and 125 repeat pairs, respectively. In a somewhat similar manner, the TSI measurement error variances were calculated for the TSI HC and CO values using the repeat TSI data. High-speed idle HC and CO and low-speed idle HC and CO had 146, 101, 159, and 111 repeat pairs, respectively.

The pooling of measurement variances for the repeat-tested vehicles must be performed in a transformed space where measurement error is homogeneous, that is, where the scatter from measurement error is constant across the range of emissions levels. We searched for the optimum transformation using the following procedure. Each set of repeat pairs was divided into low-valued pairs and high-valued pairs. Pairs were assigned to low if their transformed-space average was below the transformed-space value corresponding to 100 ppm for HC or 1.0% for CO; otherwise, they were assigned to high. Then, we considered different power transformations from λ= 0.1 to 0.9 until the pooled standard deviations of the within-pair differences were the same for the low set and the high set.

The same approach was used to estimate the measurement variance of the IM240 HC, CO, and NO_x. Table A-1 shows the measurement variances for TSI and IM240 tests.

Table A-1. Measurement Variances for IM240 and TSI Measurements

	Space	**Variance**
IM240 HC (g/mile)	natural log	0.0798
IM240 CO (g/mile)	natural log	0.284
IM240 NO_x (g/mile)	natural log	0.126
High-Speed Idle HC (ppm)	0.38 power	1.30
High-Speed Idle CO (%)	0.60 power	0.042
Low-Speed Idle HC (ppm)	0.32 power	0.92
Low-Speed Idle CO (%)	0.75 power	0.037

To put the measurement error variances in perspective, the variances given in Table A-1 have been converted to the 95% confidence limits in linear space shown in Table A-2. The confidence limits can be interpreted as follows. The exact value of a vehicle’s emission rate is unknown; the measured value is just an estimate of the emission rate. The probability that the exact value falls within the confidence limits in the table is 95%. For example, if a measured IM240 CO value were 10 g/mile, we would be 95% confident that the exact IM240 CO would be between 3.5 and 28 g/mile.

Table A-2. Measurement Error 95% Confidence Limits for IM240 and TSI in Linear Space

Measured Emission Value	IM240 HC (g/mile)		IM240 CO (g/mile)		IM240 NOx (g/mile)		High-Speed HC (ppm)		Low-Speed HC (ppm)		High-Speed CO (%)		Low-Speed CO (%)	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
0.01											0.00	0.28	0.00	0.30
0.1	0.06	0.17	0.04	0.28	0.05	0.20					0.00	0.49	0.00	0.46
1	0.57	1.72	0.35	2.84	0.50	2.00	0	22	0	27	0.42	1.76	0.53	1.53
10	5.75	17.20	3.52	28.40	5.00	20.00	0	57	0	74	8.38	11.74	9.12	10.90
100			35.21	284.00			27	237	17	306				
1000							628	1485	486	1796				

An examination of the resulting measurement error magnitudes might lead the reader to question the ability of the TSI (especially at low TSI values) to be useful to predict the average IM240 emission rate of a fleet. In fact, the evaluation of sources of error when applying these models to a fleet reveals that the TSI measurement error is one of the smaller sources of error.
In any case, the non-negligible error variances for the TSI values, which were used as predictor variables, provide a motivation for using measurement error models. This topic is discussed in the following subsection.

Measurement Error Method for Final Models
In conventional regression analysis, it is assumed that the dependent variable (the IM240 HC, CO, or NO_x value in this study) has error, but the independent variables have no error. The TSI variables included as predictor variables in the models have, as we have shown above, non-negligible measurement errors. Since the assumptions of conventional regression analysis are not satisfied for this problem, if this method had been used to develop the final models, there would have been biases in the regression coefficients. To avoid this problem, statistical methods designed to handle situations with errors in both the dependent and independent models were used. The type of model in which there are errors in both the dependent variable and one or more of the independent variables are called "measurement error models." Measurement error models were developed using EV CARP software. This program is a product of the Statistical Laboratory at Iowa State University.

As with conventional regression, EV CARP requires the value of the dependent variable and the values of the independent variables for each observation to be used in the model development. Other inputs are required also, depending on the option of EV CARP that is selected. The option used by ERG is called EV1. We elected to supply the variance of the measurement error in the dependent variable, which is an optional input with EV1. Additionally, the variances of the measurement errors in the predictor variables were input. Covariances quantify the relationships between measurement errors for different variables. Error covariances were also calculated from the repeat emissions tests and supplied to the software.

The EV1 option is especially suited for this application because it accounts for several separate sources of variability. Disagreement between individual measured IM240 values and the IM240 values predicted by the model occurs for three reasons. First, measurement errors in the dependent variable (the IM240 value) cause data scatter. Second, the TSI values measured with error are used in the model, so TSI measurement error also causes differences between measured and predicted IM240 values.

There is a third reason for data scatter. Even if the TSI values and IM240 values were measured with no error, there would still be some disagreement between the measured and predicted IM240 values. This is because of idiosyncrasies of individual vehicles that cannot reasonably be captured perfectly by the model.

EV CARP is especially appropriate for this application, since it provides an option that accounts for all three sources of data scatter mentioned above.

Ideally, the measurement variances and covariances of the predictor variables would be calculated for input into EV CARP in the transform space where the variances were homogeneous. These spaces were determined in the analysis described in the previous subsection. However, we found that when the measurement error models were built using these transformations for the input variables, the regression results for the measurement error models were unstable. This instability was characterized by large changes in the regression coefficients compared to the values obtained with the conventional regression analysis. In some cases, the regression coefficients changed sign. We found that to achieve a stable measurement error model it was necessary to change the transformations used for the two-speed idle measurements. We found that the natural log of the two-speed idle measurements produced measurement error models which were stable. Unfortunately, this means that the two-speed idle variances and co-variances used to develop the models were the average variances for the dataset when we know that the variances are not homogeneous in log space. By using these average variance values, the model "believes" that low TSI values carry more information and high TSI values carry less information then they actually do. Nevertheless, the use of these average variance values will provide models that should be superior to models built without considering measurement error at all.

A.3 Limitations of the Models in Applications

The models developed for TNRCC relate emissions from TSI concentrations to IM240 emission rates as they were determined: 1) in two specific Texas I/M stations for TSI measurements and in a portable IM240 dynamometer environment for IM240 measurements; and 2) on a specific set of vehicles. Therefore, as with any models, application of these models to other situations may result in the introduction of biases in the results. Biases can be introduced through the application of the model in situations with different TSI test conditions and/or different vehicle characteristics from those used in the dataset used to develop the models. Nevertheless, the variety of model years, technologies, vehicle types, and vehicle ages used in the model building data set should be sufficiently diverse to allow the model to be used successfully in many real situations.

In the discussion in this section, we present a summary of the test conditions and vehicle characteristics under which these models were built. The model user should consider how the model application dataset differs with respect to test conditions and vehicle characteristics when he uses the models reported in this study.

The following test conditions were used to acquire the model training dataset:

- TSIs were measured with Texas I/M station grade BAR90 equipment and procedures;
- TSIs were measured at ambient temperature and relative humidity;
- TSIs and IM240s were determined on vehicles with as-received fuel; and
- IM240s were measured on a portable dynamometer system.

If TSIs are collected for an application dataset with equipment and procedures other than those at the Texas I/M stations used to develop the model training dataset, then there is a possibility of a bias or a different variance for the TSI measurements between the training dataset and the application dataset.

The effects of ambient temperature and relative humidity on TSI HC and CO results to our knowledge, are not known. Therefore, TSI results at conditions other than the ambient temperature and relative humidity used for the training dataset could produce TSI values which are systematically different.

There are several vehicle characteristics of the training dataset which could affect the applicability of the models developed:

- Model year and vehicle age;
- Vehicle type;
- I/M program in place at the time of the training dataset collection; and
- Small, specific fractions of the fleet.

Application of the models to datasets which differ significantly from the training dataset in model year could be a step outside prudent application limits. This would also include application to datasets where vehicle ages were significantly different from those in the training dataset even though the model year distribution was similar. The model user should also be aware of the emission control technologies used on the vehicles in the application dataset although attention to the model year distribution should be adequate given the high correlation between emission control technology and model year.

Consequently, we expect that it will be beneficial to update the models as newer TSI and IM240 measurements on a set of vehicles become available.

The Texas models were built on vehicles with models years from 1981 to 1997. Smaller numbers of vehicles in the oldest model years mean that the uncertainty in the predicted IM240 values for vehicles in those model years is relatively larger than for the IM240 emissions in the later years. As far as predicting fleet emissions is concerned, for the middle 1990's model year vehicles, the very low IM240 and TSI emissions of these vehicles make the measurement and prediction of IM240 emissions with small relative errors difficult.

Perhaps a more subtle limitation on the application of the models developed in this study is the effect of the I/M program in force at the time of data collection. For the training dataset, the I/M program at the time was based on two-speed idle testing. Therefore, the vehicles which were tested for TSI and IM240 emissions were subject to a two-speed idle I/M program. As long as the models developed in this study are applied to vehicles subject to the same two-speed idle I/M program and cutpoints, there should be no question that the model application is appropriate from this perspective. However, if a different I/M program is instituted, then it is possible that the relationship between TSI and IM240 could be different. Under a new I/M program, vehicles would be tested and repaired based on other emissions results. There is no guarantee that the resulting changes in the emissions characteristics of the vehicle population would preserve the TSI to IM240 relationships discovered in this study.

The correlation models are intended to be used to estimate the <u>average</u> IM240 emissions of a large fleet of vehicles such as the Texas fleet. The estimates can be made for different cities and for different model years. The uncertainty of the average will increase for small fractions of a fleet since small fractions could not have been well represented in the model training dataset. For example, we would expect larger uncertainties for predicted IM240 emissions for 1985 light-duty carbureted trucks. Thus, as an investigator further sub-divides the application dataset when applying these models, the uncertainty of the mean predicted IM240 emission rates increases. In the extreme, the largest uncertainties are those for a single vehicle based on its TSI measurement.

<u>A.4 Accuracy of the Models in Their Application</u>

This section discusses application of the models and the roles of various sources of variability. Issues pertaining to model precision and bias and the effect on the estimation of the fleet average by using the models are also covered.

The role of several types of variance in using the model to estimate a fleet average is discussed below. The estimation of the fleet average involves first estimating the average IM240 emission rate in model year strata. These stratum-specific averages are weighted by their travel fractions and summed to obtain the estimated fleet average. Model refinement is necessary to achieve zero or insignificant biases in the strata. This in turn produces a zero or insignificant bias in the estimate of the fleet average.

Refinements in the model precision may not change the estimated precision in the fleet average. Refinements improve the estimate of the emissions of a specific vehicle. Thus, the unexplained part of the variance in the IM240 values decreases. However, the explained part of the variance of the IM240 values increases by an equal amount. The uncertainty in the estimate of an average emission level is a function of both types of variance. The details of these relationships are discussed further below.

Even if an enhancement to the model does not change the estimated precision in the fleet average, improving the model is still beneficial. As is mentioned above, enhancements in the model reduce the possibility of bias in the strata and therefore reduce the possibility of bias in the estimate of the fleet average. A detailed set of plots may be used to determine if any bias remains in the models and if it is small compared to the random scatter.

What is meant by bias in this context is lack of fit between the model and the data that could be eliminated by modifying the terms in the model in some manner (including additional terms or changing the functional forms of the existing terms). The issue here does not pertain to biases in the data or to biases resulting from inappropriate use of the models.

Variability as it Influences Precision and Bias

We will briefly review the different sources of variability and indicate the role of each source. The primary emphasis of this section pertains to the application of the models. However, this cannot be adequately discussed without some reference to the model development.

IM240 measurement errors, TSI measurement errors, and vehicle-to-vehicle idiosyncrasies that are not captured by the models all affect the model development. All of these sources of variability contribute to scatter of data points in the model-development dataset about the IM240 values predicted by the model.

Of these sources of variability, the IM240 measurement errors do not affect the application of the models. This is because the models are applied in situations in which TSI measurements are used to predict the IM240 values. The IM240 measurements are not actually made, so IM240 measurement error is not a factor at all in situations in which the models are applied.

The IM240 values predicted by the models are direct functions of the TSI measurements. Thus, the measurement errors in the TSI values in the application dataset affect the results obtained by using the models.

Vehicle-to-vehicle idiosyncrasies introduce another source of variability. This variability is real; that is, it represents variability among true IM240 emission rates that is not explained by the model and is not caused by measurement error.

These ideas are illustrated by the following conceptual equation:

Estimate of the Measured IM240

= TSI Terms + Terms without Error + Vehicle-to-Vehicle Term

The terms involving TSI values are self-explanatory. The terms without error include predictor variables such as vehicle age, vehicle type, fuel metering type, and displacement that are considered to be known essentially without error. The vehicle-to-vehicle term represents the effect of the vehicle's specific characteristics that are not captured by the model.

If one were interested in predicting the IM240 value for an individual vehicle, the predominant errors of concern would include (1) the effect of measurement errors in the TSI values used in the prediction and (2) the unexplainable vehicle-to-vehicle term. Since the IM240 value for a specific vehicle is needed, variability among true IM240 values in the fleet does not contribute to the relevant error in the estimate.

Alternatively, suppose we want to estimate the average emissions for a fleet or stratum within a fleet. Even if there were no TSI measurement error and no unexplainable vehicle-to-vehicle term, the average IM240 value based on a sample of size n would still have an error. The sample will not perfectly represent the population from which it is drawn. The imperfect representation of the population by the sample occurs because of random variability of the true IM240 values among vehicles in the fleet and because of the random sampling process.

To summarize, the following three sources of variation affect the estimation of IM240 average emissions on the basis of predictions made using one of the models:

(1) The effect of TSI measurement errors on the predictions;
(2) True variability of the IM240 values that is captured by the model; and
(3) True variability of the IM240 values that is not captured by the model.

The first two errors listed above are represented in the predicted IM240 values. If we compute the variance of the predicted IM240 values for our sample of size n, this variance will represent the effect of these two sources of variability. We call the variance of the n predicted values $s^2_{explained}$. This is the "explained" variance in the sense that it can be computed directly on the basis of the predicted values.

But, as is discussed above, the emission rate of a given vehicle deviates from the predicted value because of the vehicle-to-vehicle idiosyncrasy effect. Since the vehicle-to-vehicle effect is not "explained" in terms of the predicted IM240 values, we denote its variance $s^2_{unexplained}$. The total variance of a single IM240 prediction is as follows:

$$s^2_{total} = s^2_{explained} + s^2_{unexplained}$$

If we average the predicted IM240 values for n vehicles, the result will differ from the true average for the population sampled because of both the unexplained and the explained errors discussed above. The variance of the error in the mean associated with the explained part of the variance is $s^2_{explained}/n$. The variance of the error in the mean associated with the unexplained part of the variance is $s^2_{unexplained}/n$. Thus, the total variance of the error in the mean is as follows:

$$s^2_{mean} = \frac{s^2_{explained} + s^2_{unexplained}}{n}$$

Now, suppose we make some improvements to the model that allow the model to "explain" some of the variance that was previously unexplained and therefore included in the vehicle-to-vehicle term. The improvement, therefore, reduces $s^2_{unexplained}$ to some extent.

The total variance among the true IM240 values in the sampled population does not change as a result of the change to our model. That is, changing the model does not change s^2_{total}. The s^2_{total} is the same (except for the small effects of TSI measurement error and IM240 measurement error) as if IM240s had actually been measured. Thus, $s^2_{explained}$ is increased by the amount that $s^2_{unexplained}$ is decreased. Thus, the error variance s^2_{mean}, which is an estimate of the precision in the mean, is not changed by improvements in the model (one can contrive exceptions to this statement on the basis of trivial models).

This does not, however, imply that improvements in the model lead to no improvement in the estimation of the fleet average IM240 emission rate; model improvements lead to reduced biases. The process of making this estimation will be briefly summarized here. The average and error variance of the average is computed within each stratum, where a stratum consists of all data for a specific model year. The estimated average emission rate for the fleet equals the sum of stratum-specific averages, each weighted by its travel fraction.

Now, suppose we omitted a variable, such as model year, from the model. The mean residual (observed minus predicted value) in the model development dataset would still be zero, since this is a property of regression analysis. However, there would be biases in the strata. Similar comments would apply if model year were included in the model, but the functional form of the term involving model year did not fit the data. Avoiding prediction bias is much more complicated than simply being sure that all the necessary variables are included in the model in their simplest forms.

In this exercise, the sample size of the model development datasets HC, CO, NO_x were 897, 921, and 918 observations, respectively. Despite this large sample size, the counts in the strata can be small. For example, the largest number of vehicles for any model year is 66 for 1988 and 1993. Much smaller counts exist for some years. For example, 1981 has only 8 counts.

Even if the model development dataset were selected randomly from the fleet, because of sampling variability, one would not expect the travel fractions in the fleet to be exactly matched by the fractions of vehicles in the strata in the model development dataset. Thus, even if the biases in the different strata in the model development dataset produce an average residual value of zero, the biases will in all likelihood not balance when the fleet average is computed on the basis of the application dataset.

The solution is to develop the models so that the biases in the strata are zero or insignificant. If this is achieved, the bias in the fleet average is likely also to be zero or insignificant. There will be no necessity for the biases in the different strata to "balance" each other for the bias in the fleet average to be unbiased.

Evaluation of Bias in the Models

The importance of avoiding model bias is stressed in the discussion above. The considerable steps taken to avoid significant prediction biases are discussed in this subsection. Again, bias in this context refers to a systematic difference between the observed and predicted IM240 values, such that this systematic difference could be eliminating by including more or different terms in the models. Biases in the data or prediction biases resulting from improper use of the models are addressed in Section A.3.

Evaluation of the models to ensure that no significant biases exist is an important additional step and was performed by examining a large number of plots. Table A-3 presents a list of plots that was prepared for this purpose for the Texas models. Recall that a residual is the observed minus predicted value. The variables are in natural-log space unless otherwise noted. In addition to the scatter plots listed in Table A-3, several histograms were also prepared.

Table A-3. List of Model Validation Scatter Plots Examined for All Three Pollutants, for All Vehicles Combined, for Trucks and Cars Separately, and for Carbureted and Fuel-Injected Vehicles Separately

Y-Variable in the Plot	X-Variable in the Plot
Residual	Model Year
Residual	Natural Log of Displacement
Residual	Natural Log of High-Speed Idle HC Value
Residual	Natural Log of High-Speed Idle CO Value
Residual	Natural Log of Low-Speed Idle HC Value
Residual	Natural Log of Low-Speed Idle CO Value
Residual	Natural Log of Predicted Value of the Pollutant
Measured Value of the Pollutant	Estimated Value of the Pollutant
Ratio of Average Predicted to Average Measured Value by Model Year	Model Year
Ratio of Average Predicted to Average Measured Value by Model Year Group	Model Year Group
Mean Residual by Model Year	Model Year
Measured Value in Linear Space	Predicted Value in Linear Space
Measured versus Predicted Value in Linear Space, Expansion Showing Smaller Values	Expansion around Smaller Predicted Values in Linear Space

Table A-3 lists 13 types of plots. These were all produced for all three pollutants (HC, CO, and NO_x), resulting in 39 separate plots in a set. A complete set of plots was produced for five cases: All vehicles combined, trucks and cars separately, and carbureted vehicles and fuel-injected vehicles separately. Five sets times 39 plots per set results in 195 separate plots. The ERG staff examined all of these plots and reasonable sampling of the plots, which present the major results. A major objective in plotting residuals is to determine whether any remaining trend exists in the data. If so, it is possible that further improvement in the models can be made.

Further Discussion of the Role of Model Year

The performance of the models as a function of model year is important and warrants some discussion. One way to address this issue is to examine the average residual for each model year, as in the figures described above. However, the number of vehicles varies as a function of model year. The mean residuals for years with small numbers of data points are highly variable. One way to address this issue is to account for the different sample sizes for different years by using the t-statistic.

To address this problem, the t-statistic was computed for each model year. The t-statistic for a particular model year is as follows:

$$t = \frac{\bar{r}}{s / \sqrt{n}}$$

where

t = t-statistic;
$\bar{r}$ = mean log-space residual for this model year;

n = number of vehicles for this model year; and
s = pooled standard deviation.

The pooled standard deviation *s* is an estimate of the variability within a model year. However, the separate estimates from all model years were combined to obtain the most reliable common estimate. Pooling was necessary, since otherwise the standard deviations for some years with small numbers of vehicles were unreliable.

The t-statistic accounts explicitly for the different sample sizes in the different years. The residuals are expressed in units that are much more comparable for different years with different numbers of data points.

We have shown that the means of the log space residuals versus model year appear to be unbiased. However, when the predicted values of IM240 are considered in linear space, it is possible that biases with respect to model year can be present.

To evaluate the potential for bias in the linear space predictions as a function of model year, we calculated the average predicted and measured IM240 value for each model year in the dataset. Then we took the ratio of the average predicted IM240 value and the average measured IM240 value. These ratios were plotted as a function of model year, with a horizontal reference line at 1.0 on each graph. If there is an insignificant bias with respect to model year, the data points that apply should be scattered more or less randomly about this line. These plots showed data points that were scattered randomly about the 1.0 reference line.

Histograms for Residuals Revealing the Roles of Additional Variables
Additional plots were produced to reveal the role of other variables. These include the vehicle type (car or truck), the presence of a carburetor or fuel-injection, and the presence of exhaust gas recirculation. As is indicated in Section A.2, variables to account for the vehicle type, the carburetor versus fuel-injection dichotomy, and the exhaust gas recirculation dichotomy are included in the models. In view of this, a remaining bias with respect to these variables was not expected.