sierra research

Report No. SR97-11-02

Development of a Proposed Procedure for Determining the Equivalency of Alternative Inspection and Maintenance Programs

prepared for:

U.S. Environmental Protection Agency

November 10, 1997

prepared by:

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

Development of a Proposed Procedure for Determining the Equivalency of Alternative Inspection and Maintenance Programs

prepared for:

U.S. Environmental Protection Agency
Regional and State Programs Division
Under Contract No. 68-C4-0056
Work Assignment No. 2-03

November 10, 1997

prepared by:

Thomas C. Austin
Laurence S. Caretto
Thomas R. Carlson
Philip L. Heirigs

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

DISCLAIMER

Although the information described in this report has been funded wholly or in part by the United States Environmental Protection Agency under Contract No. 68-C4-0056, it has not been subjected to the Agency's peer and administrative review and is being released for information purposes only. It therefore may not necessarily reflect the views of the Agency and no official endorsement should be inferred.

Development of a Proposed Procedure for Determining the Equivalency of Alternative Inspection and Maintenance Programs

Table of Contents

List of Tables

List of Figures

Development of a Proposed Procedure for Determining the Equivalency of Alternative Inspection and Maintenance Programs

Summary

Under Work Assignment No. 2-03 of Contract No. 68-C4-0056, Sierra Research, Inc. (Sierra) has assisted EPA in responding to the National Highway System Designation Act by developing a proposed protocol for the evaluation of state I/M programs in areas subject to the "enhanced" I/M requirements of the Clean Air Act Amendments of 1990. The objective of the proposed evaluation process is to determine whether state I/M programs are as effective as a "benchmark" program that meets the requirements of the Clean Air Act for enhanced I/M. After considering a variety of potential approaches, Sierra has concluded that the most practical method for estimating the effectiveness of various I/M programs will be to compare the emissions and evaporative system test results of vehicles that have gone through "alternative" programs to the emissions and evaporative system test results of vehicles subject to the benchmark program. This approach maintains some consistency with EPA's original procedure for determining the adequacy of alternative I/M programs by focusing on the average emissions achieved under the program rather than on the reduction in emissions achieved. As explained in more detail later in this document, determination of the actual emission reduction associated with a particular I/M program is usually not possible because of the lack of baseline (no-I/M) data. Additional details are also provided regarding the impracticality of measuring I/M program effectiveness with currently available remote sensing technology. (However, the proposed evaluation procedure accounts for any emission reductions achieved as the result of a remote sensing program.)

The proposed approach for evaluating the effectiveness of alternative I/M programs is feasible because some states have already implemented centralized IM240 programs and are expected to implement EPA-recommended standards (cutpoints) in the near future. Although no single I/M program may use precisely the combination of program features recommended by EPA, it is expected that the Phoenix, Arizona program will be close enough to the EPA recommendations to serve as a benchmark against which other programs can be evaluated.[*] Under the proposed approach, use of the MOBILE model can be limited to the development of adjustment factors needed to account for differences in the percentage of the fleet subject to I/M testing and differences in the fuel in the area with the alternative program and the area with the benchmark program. Because of the complexity this element of the procedure entails, states may need to seek assistance from

[*]For 1981 and later models, the Phoenix I/M program includes IM240 testing for exhaust emissons combined with pressure testing of the evaporative system and gas cap. For 1967-1980 models, the idle/2500 rpm test is used for exhaust emissions and pressure testing is limited to the gas cap.

EPA or contractors familiar with the intricacies of MOBILE to perform these analyses correctly. To facilitate the most efficient process, it would be useful for states to review detail testing and analysis plans with EPA prior to the initiation of program evaluation.

Table 1 summarizes the two program evaluation options that are proposed. As described under Option 1, the most straightforward means of comparing an alternative I/M program

Table 1
Proposed I/M Program Evaluation Options

Step	Option 1 IM240 Testing of Random Sample	Option 2 Alternative Test of Random Sample
1.	Recruit stratified random sample of 1,600 vehicles that have completed I/M program requirements.	Recruit stratified random sample of 800 vehicles that have completed I/M program requirements.
2.	Measure emissions and conduct pressure test of evaporative system and gas cap.	Test 800 vehicle sample using both IM240 and alternative short test and develop correlation equation.
3.	Calculate weighted average IM240 emissions and pressure test failure rate for sample based on the age distribution of the fleet.	Recruit stratified random sample of 4,000-8,000 vehicles from population that have already completed I/M program requirements.
4.	Determine weighted average IM240 emissions and pressure test failure rate for the benchmark program using the same age distribution.	Measure alternative test emissions and conduct pressure test of evaporative system and gas cap.
5.	Adjust weighted average results of sample from step 3 above to account for I/M program compliance rate.	Calculate weighted average IM240 emissions and pressure test failure rate for sample based on the age distribution of the fleet and the IM240/alternative test correlation.
6.	Compare results of steps 4 and 5.	Determine weighted average IM240 emissions and pressure test failure rate for the benchmark program vehicles using the same age distribution.
7.		Adjust weighted average results of sample from step 5 above to account for I/M program compliance rate.
8.		Compare results of steps 6 and 7.

to a benchmark program is with IM240 testing and functional testing of the evaporative emissions control system for a representative sample of vehicles. However, under certain circumstances, alternative test procedures can be used to establish comparisons with IM240 results obtained under a benchmark program.

Option 1 involves the recruitment and testing of a random sample of vehicles that will be subjected to IM240 testing and pressure testing at the completion of I/M program requirements. In the case of centralized programs that use alternative test procedures, this may be accomplished through the computer selection of a random sample for supplemental testing in a lane that has been equipped to perform IM240 tests. In the case of decentralized programs, the proposed approach requires independent testing of the random sample after the normal I/M program requirements have been completed. As shown in Table 1, the minimum sample size for a model year-stratified random sample is 1,600 vehicles. This sample size is necessary to achieve 90% confidence that the emissions are within 10% of the true mean.

To minimize the amount of IM240 testing required, correlations can be established between the IM240 and other test procedures, such as one of the ASM tests. However, analysis of available data has shown that the relationship between short test emissions and the IM240 are *program-specific*. For example, two different vehicle populations with the same idle emissions do not necessarily have the same IM240 emissions. Emissions during stop-and-go driving, as measured by the IM240 or the FTP, are not reduced as much in I/M programs where the pass-fail decision is based on an idle test.

As shown in Table 1, the minimum number of IM240 tests required to establish correlation with an alternative test is only 800. Furthermore, a correlation established by one state could be used by another state if both states are using the same test procedure (e.g., ASM 2525). Once the IM240-alternative test correlation is established, an additional 4,000-8,000 alternative tests (4,000 for the ASM, 8,000 for idle) are required to establish an adequate level of confidence.

Additional details regarding the sampling and testing procedures required to insure reliable results are described below.

Background

Ultimately, the effectiveness of a motor vehicle inspection and maintenance (I/M) program depends on how emissions from vehicles subject to the program compare to what the emissions from the vehicles would have been in the absence of *any* I/M program. A requirement for states to measure the effectiveness of enhanced I/M programs in terms of the *reduction* in vehicle emissions they achieve is embodied in one of the provisions of §182(c)(3)(C) of the 1990 Clean Air Act Amendments:

> *Each State shall biennially prepare a report to the Administrator which assesses the emission reductions achieved by the program required under this paragraph based*

on data collected during inspection and repair of vehicles. The methods used to assess the emission reductions shall be those established by the Administrator.

While simple in concept, determining the reduction in emissions achieved with an I/M program is complicated by the possible residual effects of prior I/M cycles on the reductions achieved during subsequent I/M cycles. When significant changes are made to a program, the emissions reduction achieved during the first cycle of the new I/M program depends on the effectiveness of the prior I/M program.[*] In order to determine the net effectiveness of new program, detailed information is needed regarding the effectiveness of the prior program, which is usually unavailable.

EPA's original procedure for determining whether state I/M programs would meet the performance standard for enhanced I/M avoided the need to measure the actual emissions reduction achieved by relying on the use of the agency's emissions simulation model (MOBILE). The compliance determination was based on whether estimated emissions of vehicles that had gone through a particular I/M program were less than or equal to the emissions predicted for vehicles subject to an I/M program with EPA-recommended program features. For areas requiring "high enhanced" I/M programs, those features are:

- Centralized Network Type;
- Annual Inspection Frequency;
- 1968 and Later Model Years Included;
- LDGV, LDGT1, LDGT2 Vehicle Types Included;
- Idle Emission Test on 1968-1980 Models;
- Idle/2500 Emission Test on 1981-1985 Models;
- IM240 Emissions Test on 1986 and Later Models;
- Evaporative System Pressure Test on 1983 and Later Models;
- Evaporative System Purge Test on 1986 and Later Models;
- Visual Check of Catalyst and Fuel Inlet on 1984 and Later Models;
- Pre-1981 Stringency of 20%;
- Pre-1981 Waiver Rate of 3%;
- Post-1980 Waiver Rate of 3%; and
- Compliance Rate of 96%

Depending on the ambient temperature, fuel volatility, and fleet age distribution of the particular area, MOBILE5a predicts that, in calendar year 2000, such a program will achieve a combined total of exhaust and evaporative hydrocarbon (HC) emissions from on-road vehicles that is approximately 32% lower than it would have been in the absence

[*]For example, if one state was operating a totally ineffective I/M program prior to 1997 and another state was operating a moderately effective program that significantly reduced excessive emissions in the vehicle fleet, the failure rate and the reduction in emissions that occurs if both states implement equally effective centralized IM240 programs in 1997 will be smaller in the state that previously had the more effective program. There is less of an emissions reduction available during the first cycle of the new program if the prior program was more effective in reducing the excess emissions of the fleet.

of an I/M program. (Evaporative emissions reductions are about one-third of the total.) Carbon monoxide (CO) and oxides of nitrogen (NOx) emissions reductions for the high enhanced program are approximately 35% and 13%, respectively. Areas with less severe air quality problems are allowed to meet a less stringent "low enhanced" I/M performance standard that requires emissions to be reduced by approximately 9% for HC, 16% for CO, and 1.5% for NOx.

Some objections were expressed to the use of the MOBILE model to estimate the benefit of alternative programs for the following reasons:

- the MOBILE model did not account for the potential benefits of supplementing a conventional I/M program with a remote sensing program;
- the model did not account for all conceivable test procedures; and
- the model was designed to reflect historical evidence that I/M programs in which inspections and repairs are performed at the same facility are inherently less effective than programs in which inspections are independently performed at a facility that does not perform automotive service and repair.

Although EPA has augmented the MOBILE model to address additional test procedures and remote sensing, all differences between various I/M program designs are not addressed and significant differences of opinion remain regarding the effectiveness of programs that combine the inspection and repair functions. It should also be noted that the modeling-based approach potentially introduces a large error in the calculated reduction in emissions because of uncertainty associated with the ability of models like MOBILE to accurately predict what emissions would have been in the absence of any I/M program.

In recognition of the above-described concerns with modeling-based approaches, EPA has been working closely with states subject to the enhanced I/M provisions of the Clean Air Act to develop alternative I/M program evaluation procedures that can be used to determine whether individual state programs are providing adequate emissions reductions. Ideally, final approval of I/M credits would be based on actual emissions test results obtained using a Mass Emissions Transient Test (METT) on a representative sample of vehicles subject to the state programs. However, the acceptable techniques for collecting and analyzing data to demonstrate compliance have not yet been specifically defined. This document describes the details of the proposed procedure for determining the equivalency of alternative inspection and maintenance tests that may allow many states to avoid the use of a METT.

Basic Approach

For the reasons summarized above, mass emission testing of even a large sample of vehicles will not produce data that can be used to determine the reduction in emissions

associated with changes to an existing I/M program. After considering a variety of potential approaches, the most practical method for estimating the effectiveness of various I/M programs is to compare the emissions of vehicles that have gone through "alternative" programs to the emissions of vehicles subject to a "benchmark" program meeting EPA's definition of enhanced I/M, described above. This approach maintains some consistency with EPA's original procedure for determining the adequacy of alternative I/M programs by focusing on the average emissions achieved under the program rather than on the reduction in emissions achieved. Use of this approach is feasible because some states have already implemented centralized IM240 programs and are expected to implement EPA-recommended standards (cutpoints) in the near future.* Under the proposed approach, use of the MOBILE model can be limited to the development of adjustment factors needed to account for differences in the portion of the fleet subject to I/M and differences in fuel specifications.

Test Procedures

The advantages, disadvantages, and feasibility of several possible test procedures for measuring the emissions of vehicles subject to I/M programs are described below.

Federal Test Procedure - Although the Federal Test Procedure (FTP) does not represent the full range of vehicle operation in customer service (particularly high acceleration rates and high speeds), it defines the range of operation over which all passenger cars and light-trucks are designed to effectively control emissions. Because the FTP includes both cold start and warmup operation and a broad range of speeds and acceleration rates, it better represents vehicle operation in customer service than any test procedure used in I/M programs. More representative test procedures have recently been developed to determine emissions from vehicles in customer service; however, the FTP still serves as the basis for defining average exhaust emissions from the in-use fleet. In addition, the FTP provides for direct measurement of evaporative emissions.

Notwithstanding the advantages of measuring emissions with the FTP, it is not a practical procedure for testing vehicles passing through an I/M test facility. The vehicle must be parked for at least 12 hours prior to the beginning of the test and then subjected to a 42-minute exhaust emissions test procedure, followed by a one-hour evaporative emissions test to measure "hot soak" emissions. A separate evaporative emissions test is required to measure "diurnal" emissions. Two days are required to complete all of these tests. Typical cost to conduct the full sequence of tests in approximately $1,000 and substantial additional costs may be associated with vehicle recruitment. Given the sample sizes needed to establish the average emissions from the vehicle fleet (discussed in more detail below), the use of FTP testing is not considered economically feasible for the

*Although no single I/M program may use precisely the test procedures recommended by EPA, it is expected that certain programs, like the Phoenix, Arizona program, will be close enough to the EPA recommendations to serve as a benchmark against which other programs can be evaluated.

routine evaluation of I/M programs. (I/M program evaluations based on the FTP have been previously conducted by both the California Air Resources Board (CARB) and EPA, but the expense associated with this approach is beyond the resources available to most states.) In addition, the basic approach being recommended avoids the need to determine FTP emissions by using IM240 and functional evaporative system test results as the benchmark.

IM240 - A four-minute subset of the FTP, the IM240 test produces results that are better correlated with the full FTP than any other I/M test procedure. Figures 1, 2, and 3 illustrate the correlation* between the IM240 and the FTP for 1981 and later model year vehicles based on tests performed by CARB.** As shown in the figures, the correlation between the IM240 and the FTP is relatively good. The coefficient of determination (r^2) value shown on each figure is a measure of how well the IM240 test can predict the FTP emissions for an individual vehicle, an r^2 of 1.0 being perfect correlation. The r^2 values are 0.89 for HC, 0.66 for CO, and 0.78 for NOx.

The imperfect nature of the correlation is primarily due to the lack of cold start and warm-up operation in the IM240. This is a small price to pay for shortening the time required for the exhaust emissions test from 13 hours to 4 minutes. The principal disadvantage of the IM240 is that it requires more sophisticated test equipment than other I/M tests, making it less practical for use in decentralized programs where the economies of scale are less suited to the use of expensive equipment. However, in a centralized environment, the cost of IM240 test equipment is not significant and the cost per test can be about $25.

Although the cost of IM240 testing in a high-volume test facility is in the range of other I/M tests, the economic feasibility of IM240 testing for program evaluation depends on the I/M network design. Costs are lowest in centralized networks that already use the IM240 test procedure. In this case, the only cost associated with program evaluation is that of subjecting a random sample of the vehicles to the full IM240 procedure, instead of the shorter “fast pass” and “fast fail” versions of the test. Subjecting a random sample to the full IM240 is generally a standard feature of centralized IM240 programs. The use of IM240 testing for program evaluation also appears to be economically feasible in hybrid and centralized programs that routinely use other test procedures. Installing IM240 capability in several test lanes provides economies of scale; relatively little additional capital investment and personnel costs are needed to provide IM240 testing for a random sample of the vehicles subject to the program.

For purely decentralized programs, there are two different approaches that can be taken to collecting IM240 data for program evaluation purposes. One option is to establish one or

---

*Because the objective of the I/M evaluation process is to determine FTP emissions, this analysis focuses on the correlation between various I/M tests and the FTP. However, the ability of a test to identify excessive emissions with a low rate of false failures cannot be determined only from its correlation with the FTP.

**Data plotted in the figures are for vehicles tested during the Enhanced I/M Pilot Project and were obtained directly from CARB on diskette.

Figure 1

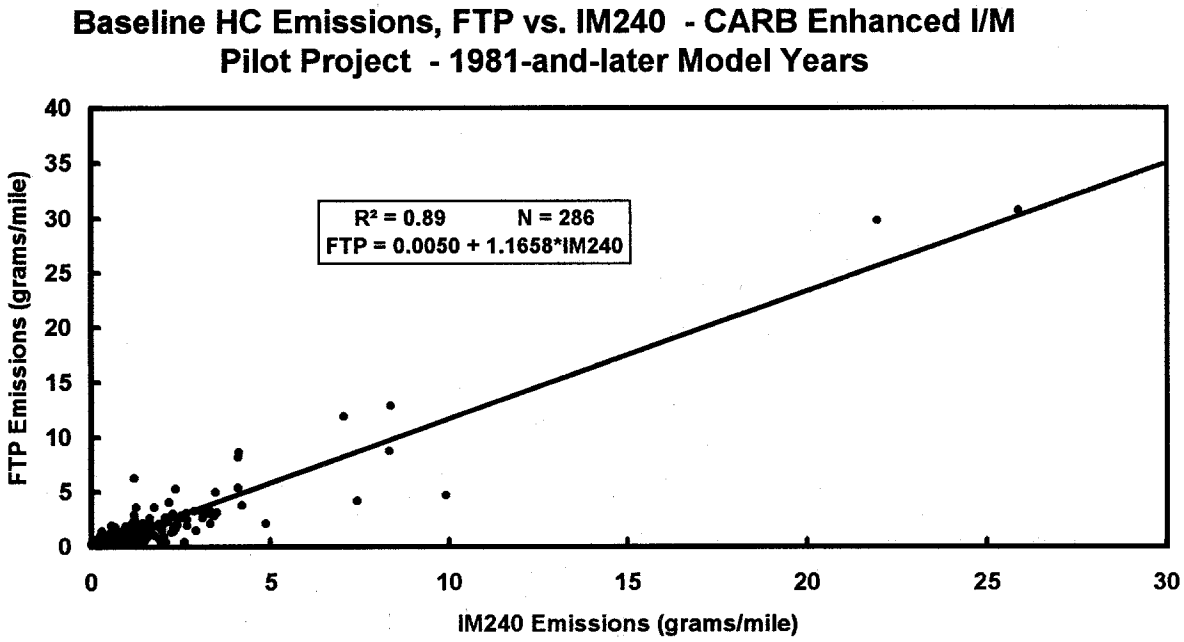


Figure 2

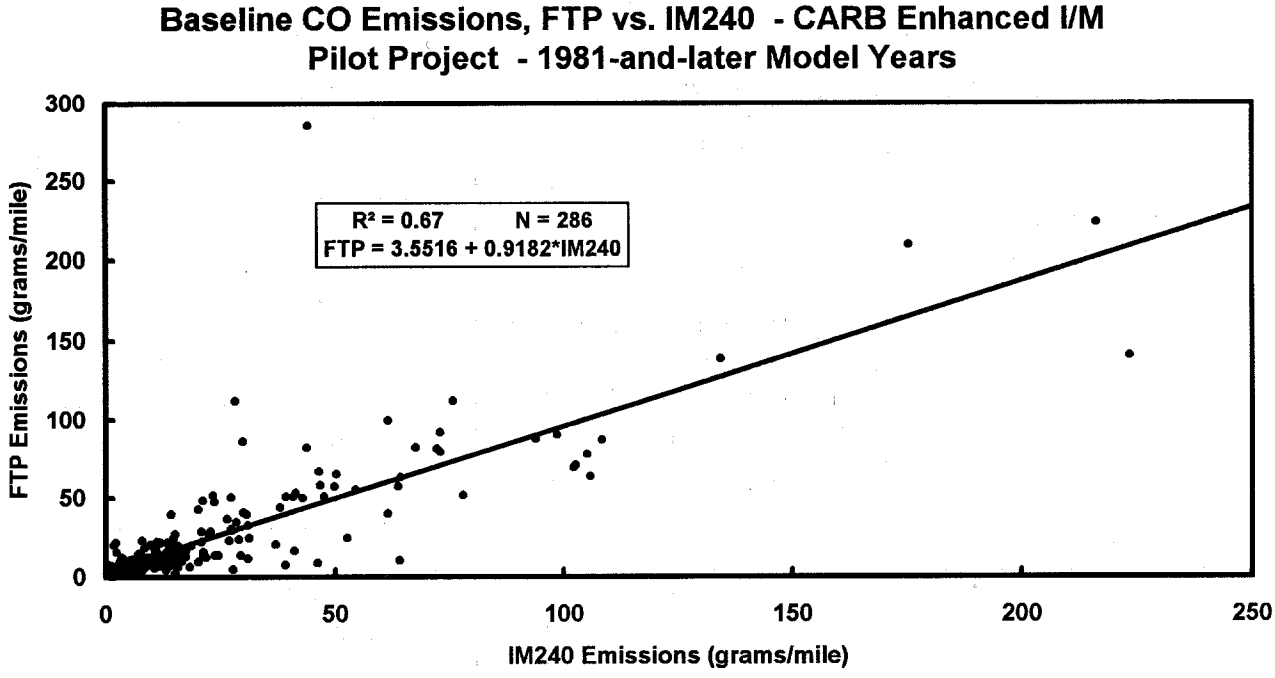
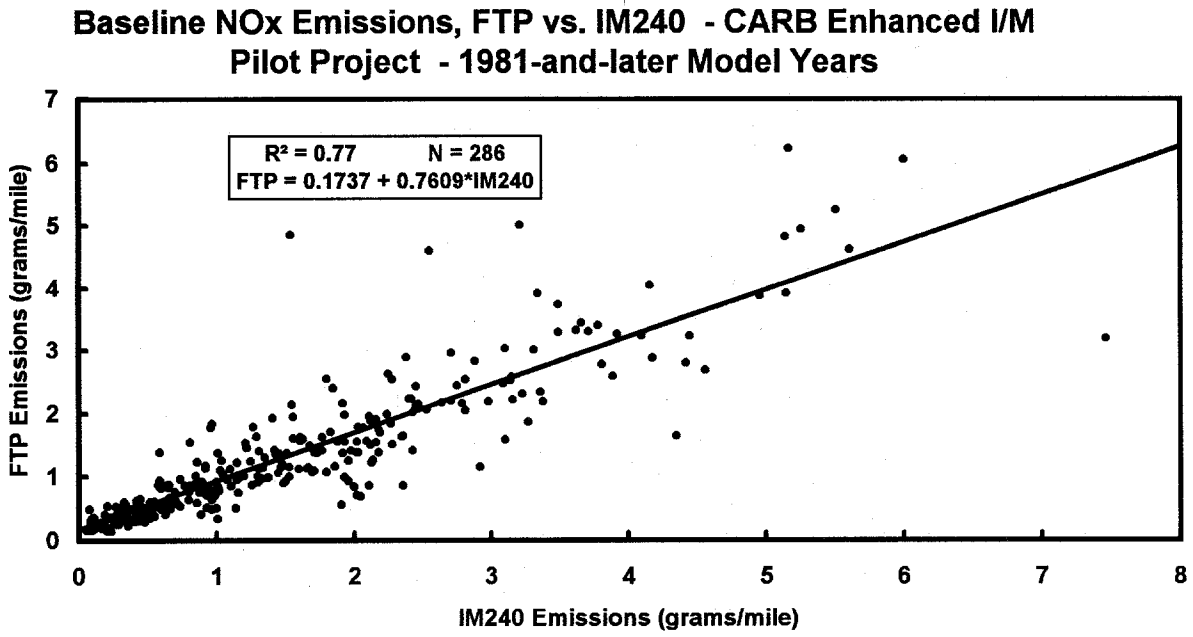


Figure 3



more centrally located IM240 test facilities and to require a random sample of vehicles to report to such facilities for the I/M test. The alternative approach is to establish a larger number of temporary facilities located adjacent to a representative sample of decentralized inspection facilities. Sierra's preliminary analysis indicates that the cost of using IM240 testing for program evaluation under either of these approaches may be prohibitive.* In the case of centrally located facilities, the cost per test for establishing and staffing a special purpose, high-volume test facility for a short-term testing program will be much higher than a typical I/M test. In the case of temporary facilities located adjacent to decentralized inspection stations, the personnel and equipment cost per test is likely to be high because of the low daily test volume. Because of these cost concerns, alternative

*Ignoring the cost of facilities, the capital cost to set up each independent IM240 test lane is estimated at approximately $200,000. Labor cost to install, maintain, and operate the system for a six-month period is estimated at approximately $150,000. Although these costs are significant, substantially greater costs are associated with the recruitment of truly random samples of vehicles for testing that will be subjected (or already have been subjected) to testing and repair at decentralized facilities wihout the knowledge of the participating garage. The cost of this element of the program exceeded $1 million during program evaluations previously conducted by the state of California.

program evaluation approaches for decentralized programs (discussed below) have been developed.

ASM - The Acceleration Simulation Mode (ASM) test procedures, ASM 5015 and ASM 2525, are steady-state dynamometer tests that simulate moderate rates of acceleration by loading the engine more than required to maintain a steady cruise speed. The tests were designed to provide the maximum possible correlation with the FTP using a simple, steady-state test procedure. Figures 4, 5, and 6 illustrate the correlation between the ASM 2525 and the FTP based on tests performed by CARB during the Enhanced I/M Pilot Project. The r^2 values are 0.77 for HC, 0.72 for CO, and 0.53 for NOx.

Like the IM240, the ASM tests have imperfect correlation with the FTP because of the lack of cold start and warm-up operation. In addition, correlation is limited by the fact that the ASM tests do not involve the wide range of operation contained in the IM240. As discussed in more detail below, the poorer correlation between the ASM procedures and the FTP requires that I/M program-specific correlations be developed. The correlations shown in the above figures are not valid for universal application because they were not constructed using data from vehicles subject to an ASM-based I/M program. (Evidence regarding the need for I/M program-specific correlations to be developed is presented below under the discussion of idle testing.)

Figure 4

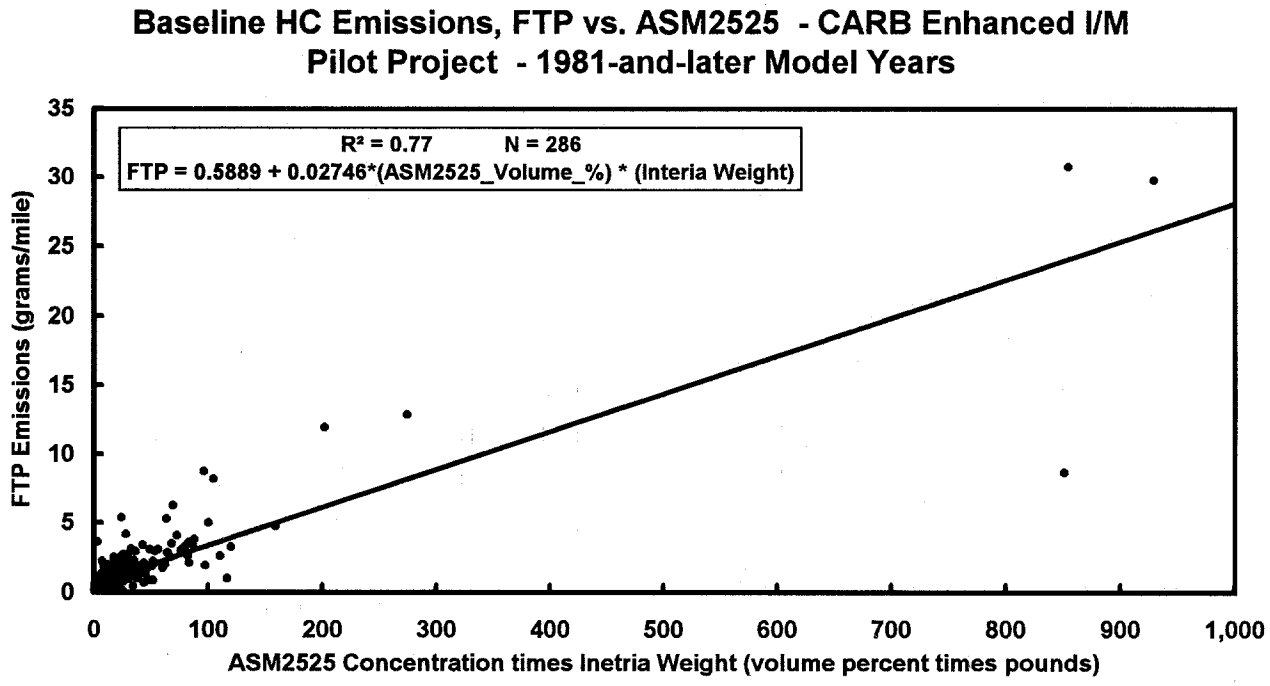


Figure 5

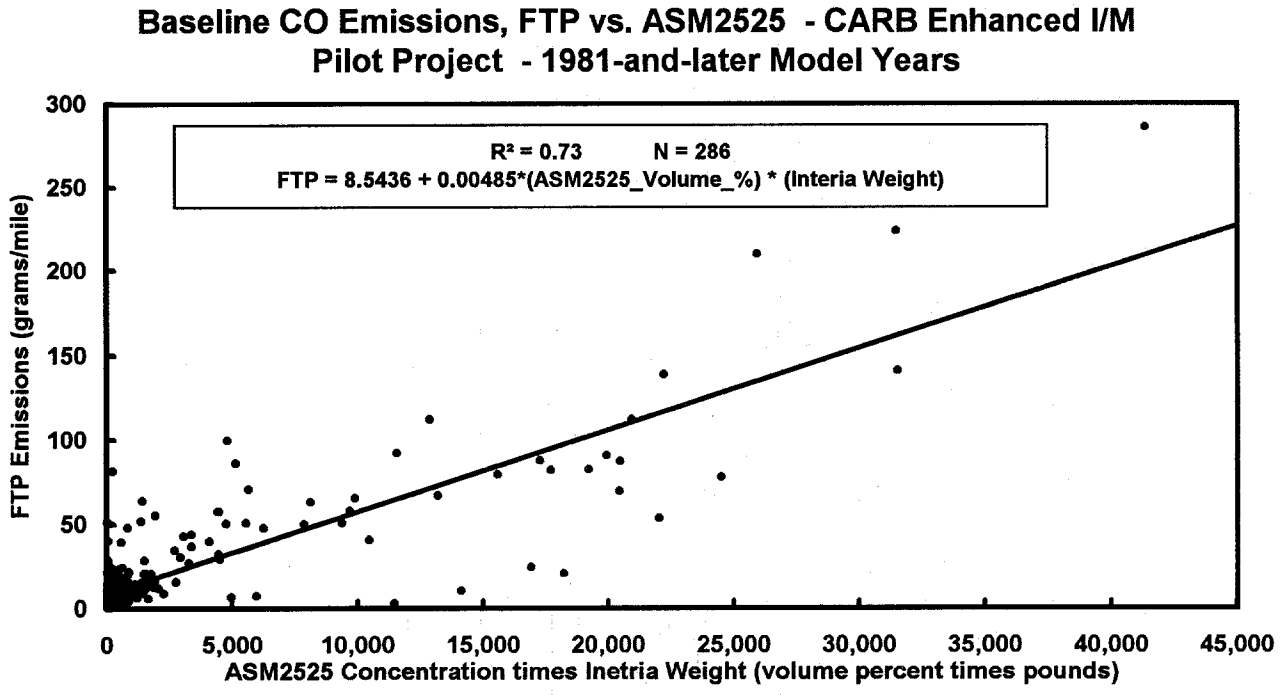
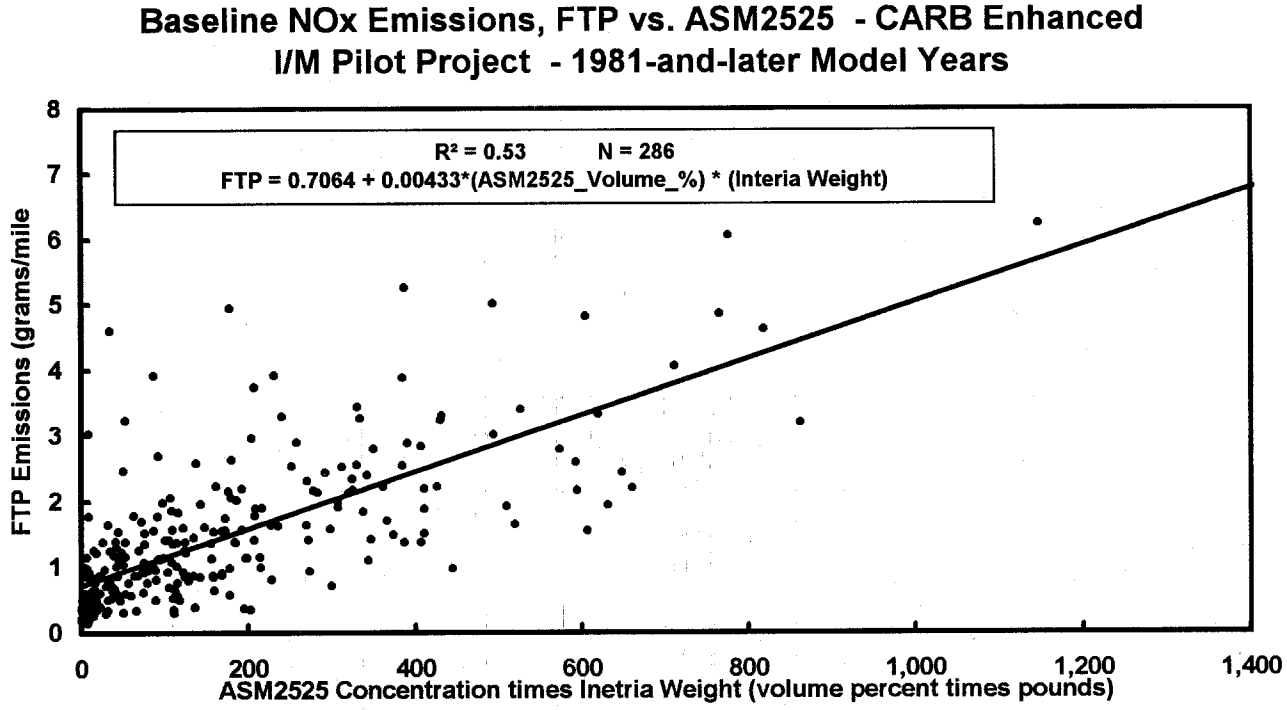


Figure 6



The principal advantage of the ASM procedures is that they can be run using less sophisticated test equipment than required for the IM240, making them more practical for use in decentralized programs where the economies of scale are less well suited to the use of expensive equipment. In a decentralized environment, the cost per test depends on the testing volume over which the capital costs are amortized. An average cost per test of $40 has been estimated if the required capital investment is spread over half the number of garages as currently participate in idle testing programs. In a centralized environment, the ASM tests can be run for less than $20 per test.

The use of ASM testing for program evaluation is feasible in centralized programs that routinely use ASM procedures, provided that ASM-IM240 correlations are available from a sample of vehicles that have been through an ASM-based I/M program. ASM may also be used for the evaluation of decentralized programs; however, as discussed in more detail below, a representative sample of vehicles must be independently tested to insure that falsification of test results has not occurred.

Idle - Idle and 2500 rpm ("high idle") no-load tests have been popular in I/M programs because they do not require the expense associated with testing a vehicle under load with a chassis dynamometer. Correlation with the FTP is relatively poor because idle and 2500 rpm tests do not represent a wide range of operation. Although idle operation frequently occurs in urban driving, correlation is generally inferior to the ASM tests because emissions are not measured with a load on the engine, the condition that accounts for most of the emissions. This is especially a problem in the case of NOx emissions, which are insignificant when there is no load on the engine.

Figures 7 and 8, which illustrate the correlation between the idle test and the FTP, are also based on tests performed by the CARB during the Enhanced I/M Pilot Project. The r^2 values are 0.64 for HC and 0.26 for CO. NOx emissions were not measured by CARB because it was recognized that the results would not have been meaningful.

The correlation between the 2500 rpm test and the FTP is illustrated in Figures 9 and 10. The r^2 values are 0.59 for HC and 0.66 for CO. Consistent with results that have been reported elsewhere, the 2500 rpm test is much better correlated with the FTP than the idle test for CO emissions. Although running the engine at 2500 rpm with the transmission out of gear is not representative of operation in customer service, the procedure puts some load on the engine, which makes it possible to identify elevated emissions levels that do not show up at low idle.

The primary limitation to the use of idle (and 2500 rpm) testing for program evaluation is the lack of data on NOx emissions. Data collected under I/M programs that use idle or idle/2500 rpm testing cannot be used to estimate the NOx emissions from vehicles subject to the program. In a decentralized environment, another concern is the potential falsification of test results, which is easier with idle testing than any other emissions test procedure. Methods to address this problem are discussed in more detail below. As in the case of ASM tests, I/M program-specific correlations must be developed to use idle test results to predict FTP emissions.

Figure 7

Baseline HC Emissions, FTP vs. Idle - CARB Enhanced I/M Pilot Project - 1981-and-later Model Years

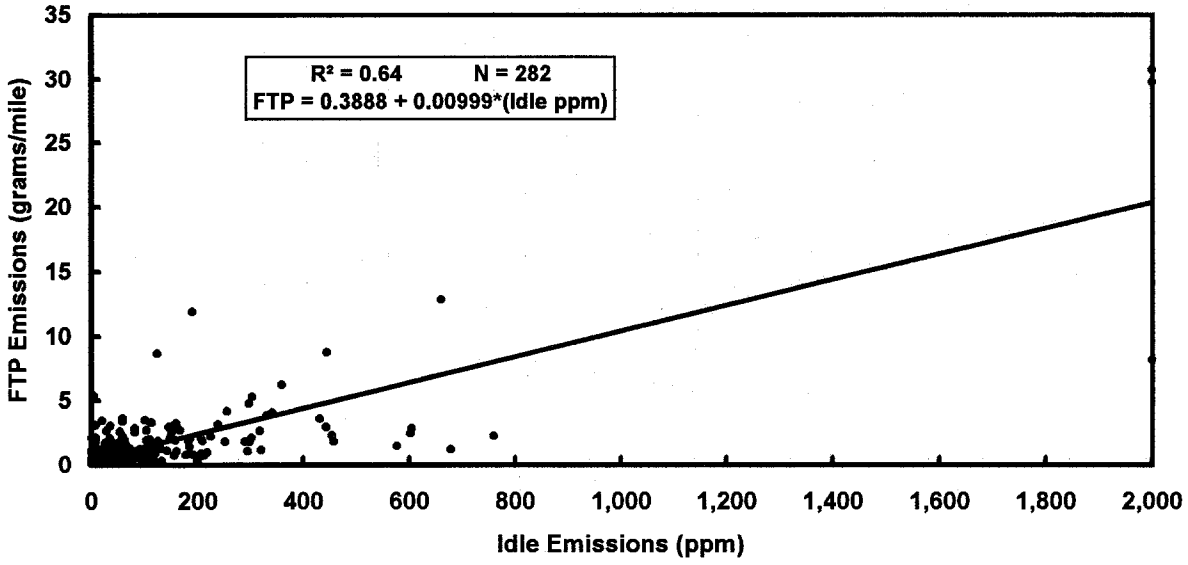


Figure 8

Baseline CO Emissions, FTP vs. Idle - CARB Enhanced I/M Pilot Project - 1981-and-later Model Years

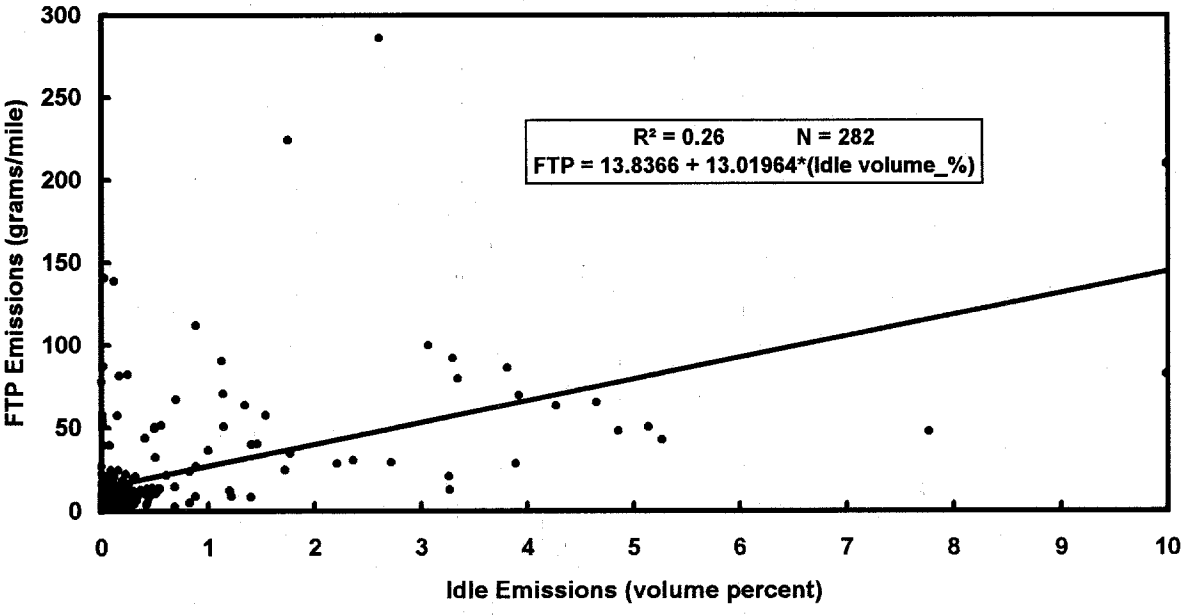


Figure 9

Baseline HC Emissions, FTP vs. 2500 RPM - CARB Enhanced I/M Pilot Project - 1981-and-later Model Years

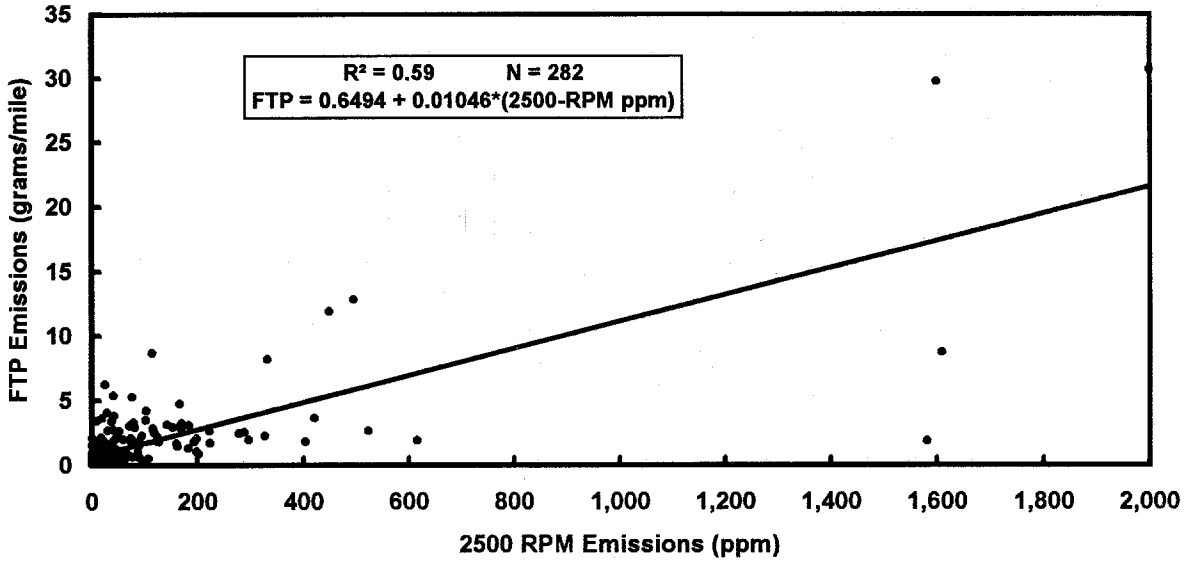
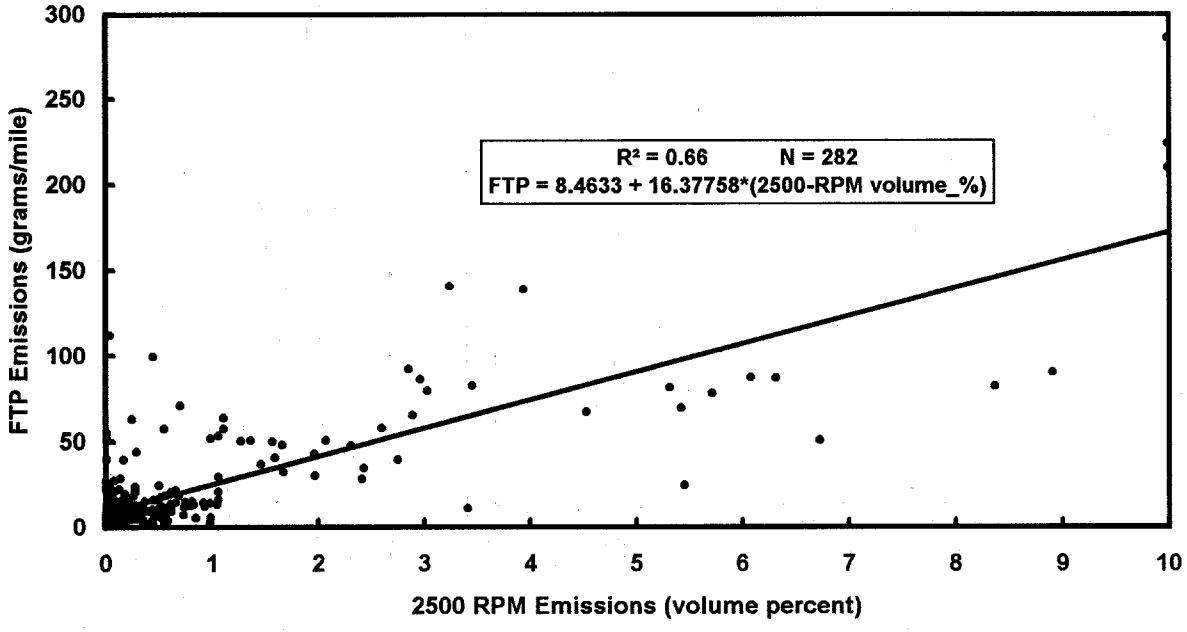


Figure 10

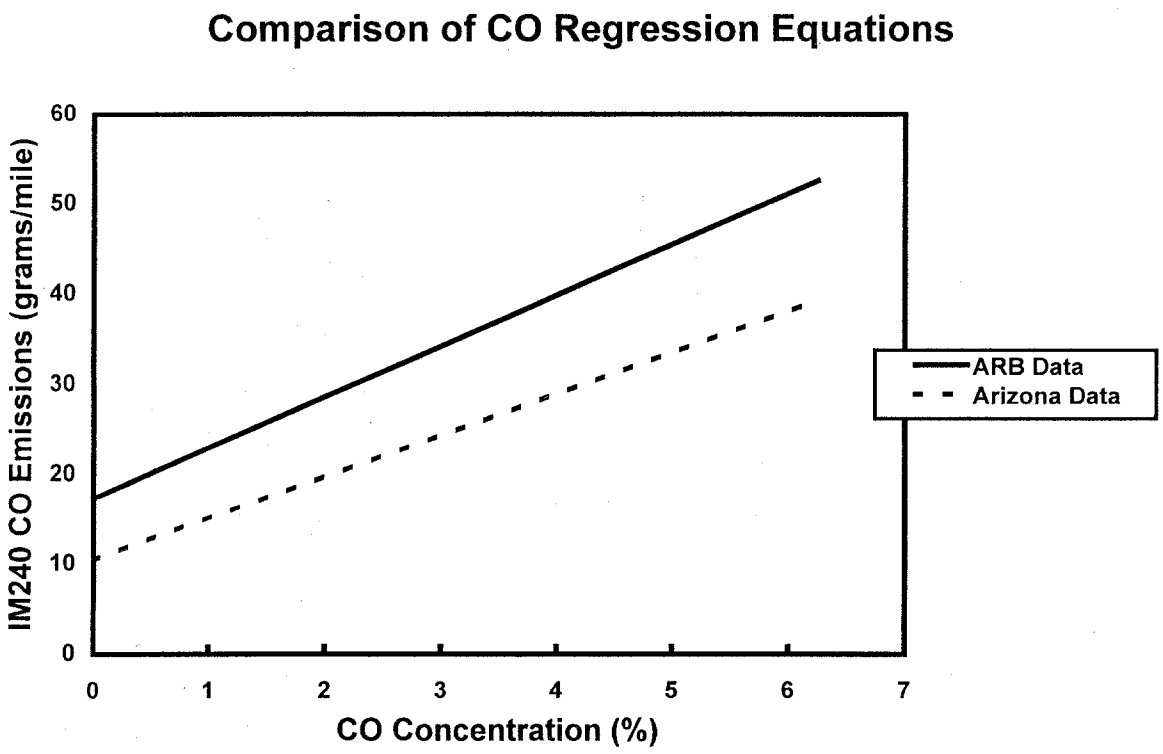
Baseline CO Emissions, FTP vs. 2500 RPM - CARB Enhanced I/M Pilot Project - 1981-and-later Model Years



The need for I/M program-specific correlations is illustrated in Figure 11. Figure 11 shows the correlation between idle CO emissions and IM240 CO emissions for vehicles subject to two different I/M programs: the California "Smog Check" program, which uses an idle test, and the Arizona I/M program, which uses the IM240 test. Data represented in the figure are from after-repair test results of vehicles that were reported to have passed the I/M test.*

In the case of the California program, many of the vehicles did not actually pass the I/M test. The results reported by the I/M test facility appear to have been falsified. However, all of the data from the California program used to construct the figure were obtained from undercover vehicles that were independently tested for idle and FTP emissions at the CARB laboratory.

Figure 11



*The Arizona results have been normalized to reflect the same model year distribution as the California results. FTP test results have been translated to IM240 emissions using a regression developed from data collected by CARB. Idle emissions for vehicles tested in Arizona were extracted from the idle period between the two hills of the IM240 test and converted to tailpipe concentration values by assuming stoichiometry.

The figure shows that for any given idle CO emission concentration, the average IM240 emissions of vehicles from the Arizona program are significantly lower. As expected, this seems to indicate that the IM240 emissions of vehicles that have been repaired to pass an idle test are higher than the IM240 emissions of vehicles with the same idle emissions that have been repaired to pass an IM240 test. This analysis demonstrates the need for program-specific correlations to be developed between short test and IM240 emissions.

Functional Test Results - As described above, there are a number of short exhaust emissions tests that demonstrate varying degrees of correlation with the emissions of vehicles under the wide range of conditions that occur in customer service, as represented by the FTP. In the case of evaporative emissions, a short test that correlates with the FTP has not yet been developed. For that reason, EPA has recommended the use of functional test procedures to identify vehicles with evaporative emissions control system defects. A pressure test of the fuel system can be used to identify a variety of vapor leaks. A test of the purge system can be used to determine whether HC vapors stored in the charcoal canister are being removed during normal driving. Based on data collected by EPA during experimental programs, significant emissions reductions can be achieved through the identification and correction of defects that cause a vehicle to fail the pressure test or the purge test.

Practical experience with the original versions of the pressure and purge tests recommended by EPA has uncovered several problems with their use. Because the tests require access to the vapor lines and vacuum lines connected to the charcoal canister, canister location makes it difficult to perform the test in a timely manner on about one-third of the vehicles. The intrusive nature of the purge test (which requires the removal and reinstallation of vacuum lines) raises the concern that some vehicles will have evaporative emissions control system defects introduced by the test.

EPA is currently studying alternative approaches to evaporative control system testing that have the potential for achieving the same degree of benefits at lower cost and with lower risk of introducing additional defects. A separate test of the gasoline cap in a special test fixture appears to be very effective and inexpensive. Testing for purge system function also appears to have potential; however, several procedural issues need to be resolved. A more comprehensive test for evaporative emissions defects involving the direct measurement of evaporative emissions (using a high flow rate sampling hood) may eventually be developed.

Until technology for directly measuring evaporative emissions is proven feasible and cost-effective, the evaluation of evaporative emissions is limited to the expensive FTP or analysis of functional testing performed in I/M programs. The effectiveness of functional evaporative emissions tests in a particular I/M program should be based on the percentage of vehicles exiting the program that pass a properly performed functional inspection. Because no I/M program currently requires routine purge testing, comparisons with a benchmark program will have to be based on pressure testing only (which may include a separate pressure test of the gas cap).

Remote Sensing - Unlike the short tests used in I/M programs, emission measurements made with the use of remote sensing devices (RSDs) do not involve the use of a precisely defined mode of vehicle operation. As a result, emission measurements obtained using remote sensing are poorly correlated with FTP emissions. This makes it difficult to use remote sensing results to identify individual vehicles with emissions-related defects.

A detailed analysis of data collected during the California "pilot project" has determined that there is potential for estimating average emissions from vehicles in customer service using data collected by remote sensing. The analysis examined how remote sensing results were correlated with IM240 measurements for various subsets of the motor vehicle fleet in Sacramento, California. One indication of the potential to estimate average emissions from remote sensing results is illustrated by Figure 12. As shown in the figure, there is a clear relationship between average CO emissions measured by remote sensors and vehicle age. However, more careful examination of the data shows that the average CO emissions for any particular model year are significantly affected by the site at which the remote sensing is done.

Figure 12

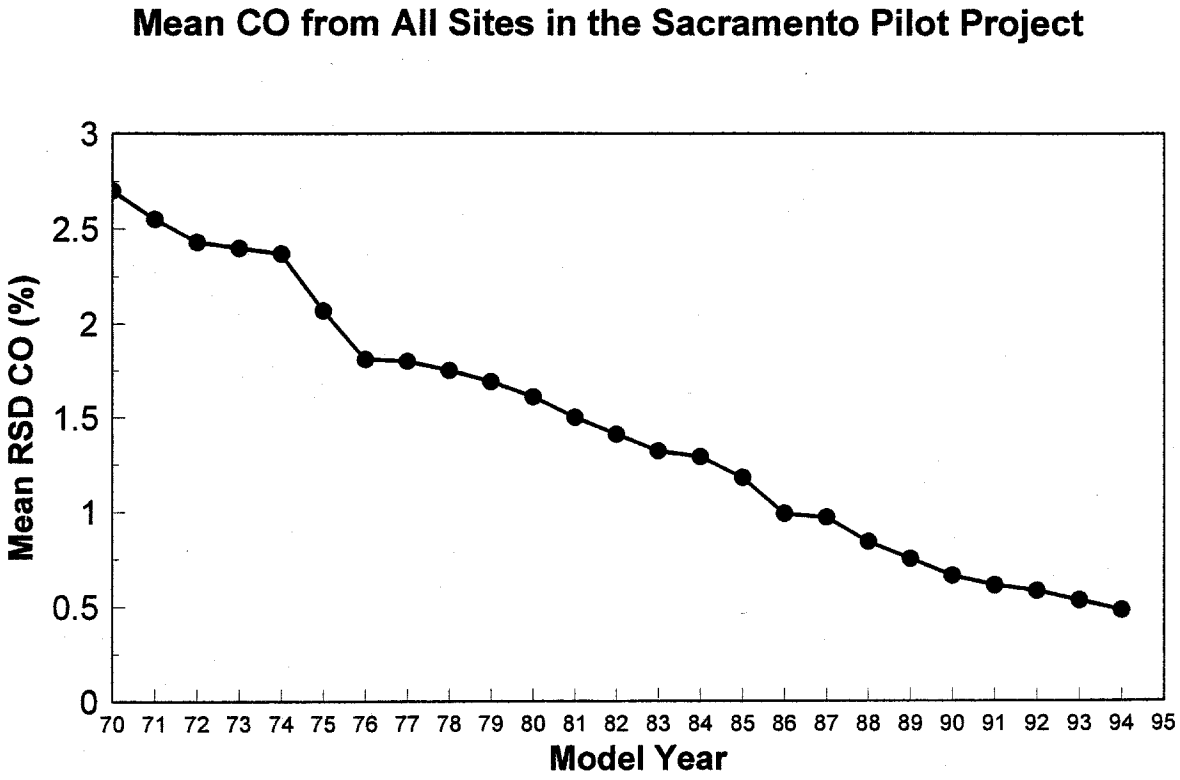
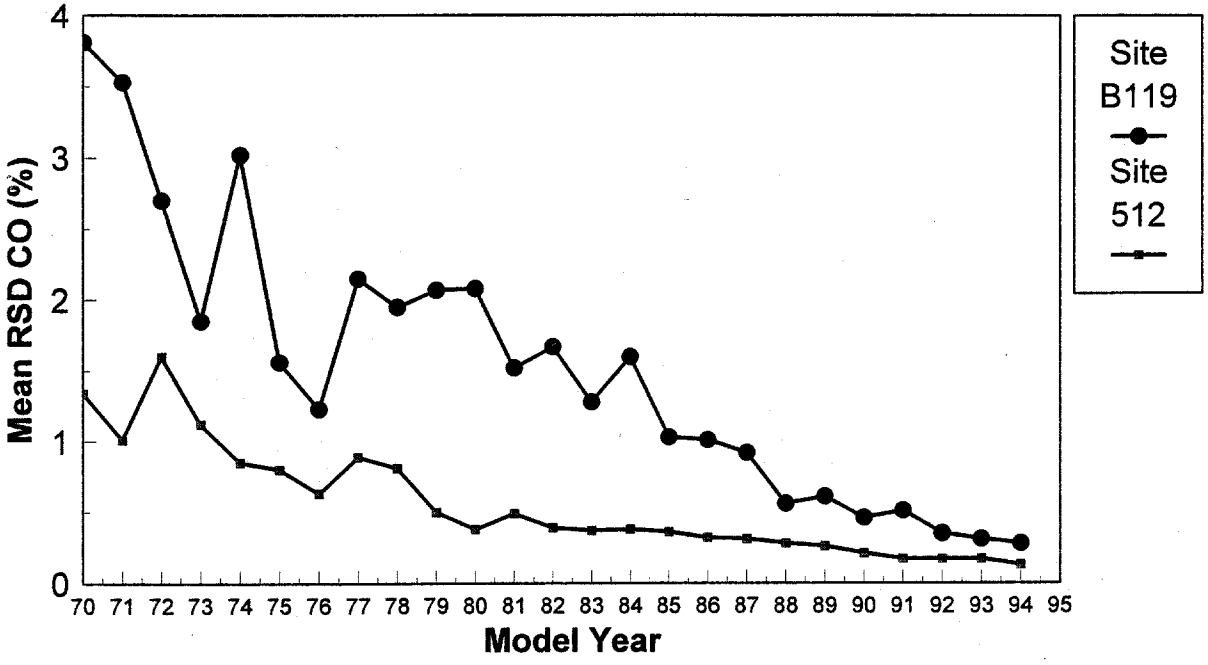


Figure 13 shows the variation in CO vs. model year for two of the sites that the California Bureau of Automotive Repair (BAR) considered to be "A" sites. "A" sites were considered the most representative because less than 5% of relatively new cars exhibited drive-by emissions in excess of 1% CO. BAR concluded that sites meeting this criterion minimized the percentage of false failures. As shown in Figure 13, the variation in average emissions between the two sites exceeds the difference in emissions between old cars and new cars at one particular site. This indicates that the relationship between CO emissions measured by RSDs and other test procedures is site-specific, making it impossible to accurately predict FTP (or IM240) emissions from RSD data unless site-specific correlations have been developed.

Figure 13

Comparison of Mean CO from 2 Individual "A" Sites in the Sacramento Pilot Project



During California's evaluation of remote sensing, IM240 data were collected on vehicles that had been measured via RSD. Because of the site-specific influences and uncontrolled vehicle operating conditions at the time of RSD measurements, the correlation between CO emissions measured by RSD and subsequent IM240 measurements on the same vehicles was very poor. For 1981 and later models, the r^2 for CO emissions was 0.07. This is further evidence that site-specific correlations with other tests are needed to use remote sensing data to evaluate emissions from vehicles in customer service. In states

with centralized programs using IM240 testing, such correlations for HC, CO, and NOx exhaust emissions could be developed using RSD data in combination with I/M program data. However, the high variability in RSD measurements means that large sample sizes would be required to establish the correlation. For states that don't routinely use IM240 testing, the costs to develop such correlations would be prohibitive.

The need for site-specific correlations could conceivably be eliminated if RSD technology and data analysis techniques are improved. Such improvement might be possible if RSD emissions measurements are supplemented with high-resolution vehicle speed measurements. By collecting both speed-time profiles, RSD data, and transient and composite emissions measurements obtained during dynamometer testing (e.g., IM240 tests), it could be possible to develop a relationship between the speed-time profile immediately preceding emissions measurement by RSD and the average emissions (as represented by the IM240 or FTP) of vehicles that have emissions-related defects. Once such a technique is developed, the need for site-specific RSD-FTP correlations would be eliminated. This would dramatically improve the accuracy of using RSDs to detect high-emission vehicles while lowering the costs of determining average emissions from vehicles in customer service. Until such technology is thoroughly demonstrated, RSD does not appear practical for I/M program evaluation.

Further data are needed to determine the feasibility of accurately estimating HC and NOx emissions from remote sensing data. Data collected during the California Pilot Project indicated difficulty in measuring HC emissions from vehicles that were not gross emitters. NOx emissions measurements were not routinely conducted. The ability to use remote sensing to estimate HC and NOx emissions will depend on the success of development efforts currently underway.

Test Applicability

Based on the considerations discussed above, Table 2 summarizes the applicability of feasible test procedures for demonstrating equivalence to the benchmark program. Exhaust emissions test procedures listed in the table are those for which there were adequate data to calculate sample size requirements (discussed below). Other test procedures (e.g., BAR31) may also be acceptable provided it can be demonstrated that they produce a reasonable degree of correlation with the IM240 for the pollutants of concern in a particular I/M area.

Sample Sizes

One option for the evaluation program outlined above calls for the development of a regression equation between the IM240 emissions and an alternative short test. The regression equation will allow the prediction of the IM240 emissions from the alternative test. Once the regression equation is developed, it is necessary to obtain test results from

Table 2
Test Applicability

	Pollutants Covered			
	HC Evap	HC Exhaust	CO	NOx
IM240	No	**Yes**	**Yes**	**Yes**
ASM*	No	**Yes**	**Yes**	**Yes**
Idle/2500*	No	**Yes**	**Yes**	No
Evaporative System Function	**Yes**	No	No	No

* Short test vs. IM240 correlations must be based on testing vehicles subject to same I/M program type and test procedures.

a sample of vehicles that will provide the average emissions of the vehicle fleet as measured by the short test. This average short test value can then be used in the regression equation to develop an estimate of the average IM240 emissions for the vehicle fleet that can be used to determine the overall effectiveness of the I/M program.

The above-described procedure requires two samples: (1) the sample used to obtain the regression equations, and (2) the sample used to obtain the average emissions for the alternative short test. Details of the approach used to develop the sample size for each of these cases is described in the appendices.

Table 3 shows the sample sizes required in order to determine the regression equation between ASM tests or idle tests and IM240. The ASM regressions may be performed with either the ASM 5015 or the ASM 2525 or the combination of both tests. The idle regressions are assumed to be based on the combination of the normal "curb" idle test and the 2500 RPM idle in neutral test. Table 4 presents the sample sizes required for tests conducted using the alternative test, once correlation with the IM240 test has been established. Also presented in the table are the sample sizes for programs using IM240 testing exclusively.

Sample sizes shown in the tables are based on the assumption that regressions will be developed for a stratified sample involving three different model year groups: 1974 and earlier, corresponding to non-catalyst vehicles; 1975-1980 model years, corresponding to oxidation catalyst vehicles; and 1981-and-later model years, corresponding to three-way catalyst-equipped vehicles. The sample sizes in the tables correspond to a relative error of 10% and a confidence level of 90%. Tables are provided in the appendix that allow sample size determinations for alternative values of relative error and confidence level. The difference in sample size for different types of non-attainment areas are due to

Table 3

IM240 Correlation Testing Sample Sizes

for 10% Relative Error and 90% Confidence Level

Analysis Assumes Stratified Sampling with Use of the Lognormal Distribution

Test Type	Sample Size vs. Nonattainment Designation		
	Ozone Only	CO Only	Ozone and CO
ASM 5015	493	698	698
ASM 2525	484	680	680
Both ASM Tests	426	661	661
Idle Tests	768	791	791

Table 4

Basic Testing Sample Size

10% error with 90% Confidence

Analysis Assumes Stratified Sampling with Use of the Lognormal Distribution

Test	Ozone Only	CO Only	Ozone and CO
IM240	1,640	1,639	1,640
ASM 5015	1,512	3,693	3,693
ASM 2525	1,897	3,965	3,965
2500 RPM	3,585	6,245	6,245
Idle	4,195	7,943	7,943

differences in the number of pollutants of concern and the variation in emissions of those pollutants.

The sample size estimates provided in Tables 3 and 4 are based on using a lognormal distribution. For the sample data sets used here, this distribution provided a better fit to the data than the normal distribution and reduced the sample size required. Sierra recommends that any future regression analysis be done using both the raw data and logarithms of the raw data to see which gives the least error. When regressions on the logarithms are used the regression errors on the original untransformed data should be computed to determine which regression approach provides the least error.

Sample Selection

Centralized Programs - Whether data are being collected to determine the correlation between an alternative test and the IM240 or to evaluate the effectiveness of the alternative program, recruitment of a truly representative sample is critical. This is a relatively simple task under a centralized program. To establish alternative test/IM240 correlation, one or more test lanes can be modified to allow IM240 testing and software modifications can be made to randomly select a subset of vehicles for IM240 and evaporative system pressure testing after testing using the alternative test has been completed. Once correlation has been established, or in cases where the IM240 test is the standard test, results can be based on the entire population of tested vehicles. Because falsification of computer-monitored test results in centralized programs is not a concern, data collected in centralized lanes can be considered reliable, as long as proper quality assurance procedures are being used. However, limited use of audit vehicles will be required to evaluate the accuracy with which visual and functional inspections that are not computer-monitored are performed. A minimum of 30 different vehicles should be used to establish the accuracy of each visual and functional inspection. Table 5 summarizes the key features that need to be contained in the audit program. Each audit vehicle should

Table 5
Key Features of a Covert Audit Program

1.	Audit vehicles must appear to be owned by ordinary motorists. Vehicles must be, or appear to be, registered to individuals living in the vicinity of the inspection facility.
2.	Individual audit vehicles and audit vehicle drivers should never be used more than once at a decentralized inspection facility and must be periodically replaced to avoid identification by inspection facility operators.
3.	Identification of audit vehicles (e.g., VIN, license plate number, model and model year, etc.) and audit vehicle drivers should be on a “need to know” basis.
4.	Audit vehicles should contain emissions-related defects that are representative of defects occurring in customer service. Implantation of defects should be done in a manner that eliminates signs of any recent service work on the vehicle (e.g., oil mist and dust should be used to coat components that have been recently changed).
5.	The condition of all audit vehicles before and after inspection should be documented by visual inspection (including photographs), functional inspections, and emissions tests.
6.	Audit vehicles should be sent to each inspection facility at least once per year. Multiple evaluations should be performed annually at high-volume inspection facilities.

contain defects representative of those found in customer service. If the defects are induced, efforts should be made to ensure that the defect is no more obvious than in typical examples of vehicles in customer service with the same category of defect. For example, vehicles failing a gas cap pressure test should be equipped with gas caps that are not obviously defective based on a visual inspection. Procedures must also be used that insure the inspectors in the centralized lane are unaware of the fact that the vehicle is an audit vehicle. (This may require the recruitment of vehicles due for inspection.)

Decentralized Programs - Information available from the California Smog Check program and from other decentralized I/M programs indicates that falsification of computer-monitored test results remains a concern. Figures 14 and 15 illustrate the pattern of results obtained from the last comprehensive evaluation of the California I/M program (called "Smog Check"). The data presented in the figures are from 831 vehicles for which initial and after-repair inspection results were obtained both at CARB's laboratory and from decentralized inspection facilities. All of the vehicles contained emissions-related defects that should have been identified during a properly conducted inspection.

Figure 14 shows the pattern observed in cases where the results reported by the inspection facility were accurate, which was the case for 77% of the vehicles. On average, the initial test reported by the inspection facility (1.20% CO) was lower than the value recorded by CARB (1.63% CO) just before the vehicle was sent to a private garage for inspection. However, some of the difference is known to be caused by pre-inspection maintenance. The final inspection result reported by the private garage (0.66% CO) is only slightly lower than the results obtained when the vehicle was returned to the CARB lab (0.73% CO).

Figure 15 shows the pattern observed in cases where the results reported by the inspection facility appear to have been falsified, which occurred 23% of the time. On average, the initial test reported by the inspection facility (2.04% CO) was lower than the value recorded by CARB (3.00% CO) just before the vehicle was sent to a private garage for inspection. The available data indicate this difference was not the result of pre-inspection maintenance. The final inspection result reported by the private garage (0.69% CO) is dramatically lower than the results obtained when the vehicle was returned to the CARB lab (3.36% CO).

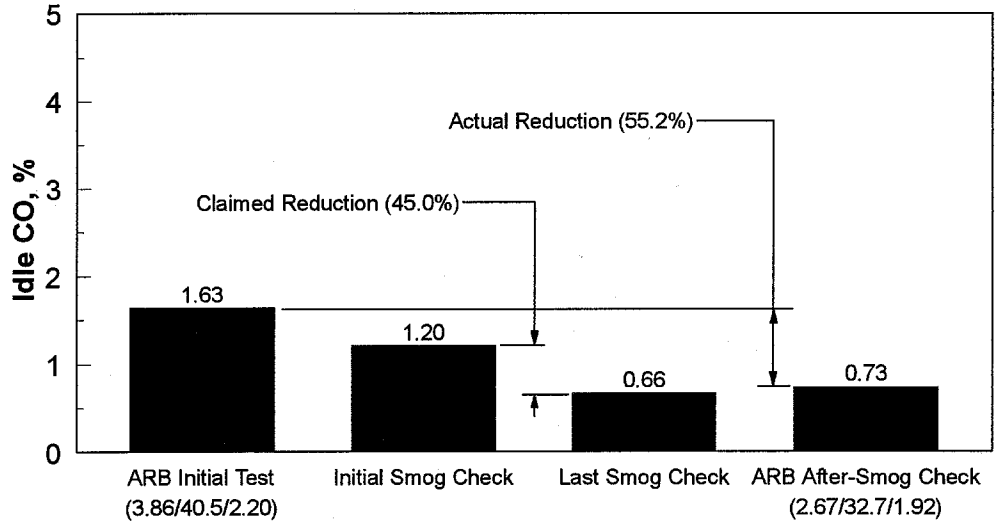
The data presented in Figures 14 and 15 indicate that falsely reported test results occur frequently and false results are more likely to be reported for vehicles with the highest emissions.* These data indicate that results reported from decentralized testing facilities will need to be verified independently through testing at a test-only facility operated by the state or a state contractor. For this independent testing to be meaningful, it will have to be performed on a representative sample of vehicles recruited from customer service.

---

*These figures are reproduced from Sierra Research Report No. SR94-04-01, "Analysis of Invalid Emission Testing in the California Smog Check Program," prepared for the U.S. Environmental Protection Agency, April 27, 1994.

Figure 14

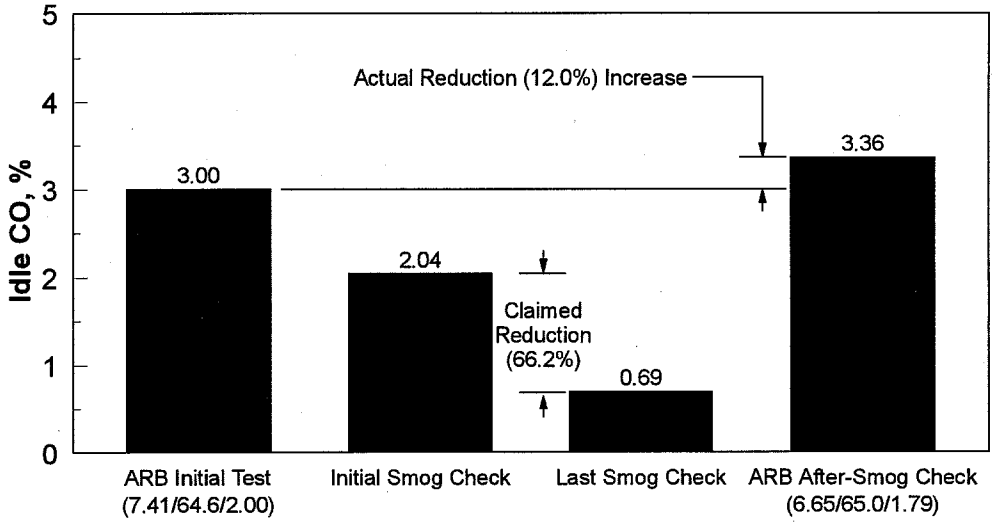
Average Idle CO Readings for Undercover Vehicles With Last Smog Check Test Confirmed by ARB (FTP Results, g/mi HC/CO/NOx, in parentheses)



ARB After-Smog readings do not exceed Last Smog Check readings by more than 400 ppm/2.00%.
641 vehicles met the criteria.

Figure 15

Average Idle CO Readings for Undercover Vehicles With Last Smog Check Test Much Lower Than ARB Retest (FTP Results, g/mi HC/CO/NOx, in parentheses)



ARB After-Smog readings exceed Last Smog Check readings by at least 400 ppm HC or 2.00% CO.
190 vehicles met the criteria.

Experience in California indicates that programs that depend on the voluntary participation of vehicle owners have significant limitations. Even with the offer of replacement vehicles and financial incentives, most motorists are unwilling to provide their vehicles for testing. This leaves two options available to collect a valid sample, both of which have been used in California. The simplest option involves randomly selecting vehicles from registration lists and making participation in the program an absolute condition of registration renewal. The other option involves recruiting a much larger sample than required and then screening vehicles voluntarily submitted for testing to achieve a sample that matches the fleet in terms of age, manufacturer representation, and state of repair (i.e., patterns of emissions-related defects). In order to determine the target pattern of emissions-related defects within each model year range, it is necessary to have inspection results from a truly random sample of the vehicle fleet. In California, such data have been collected using a roadside inspection program conducted with the assistance of the Highway Patrol.

Under the option involving recruitment of a random sample from registration lists there are two different mechanisms that can be used to obtain independent verification of test results. One involves notifying vehicle owners that a second inspection of the vehicle is required at a special test-only inspection facility operated by the state or a contractor. It is necessary for this notification to occur after the vehicle has completed the normal inspection process to minimize the possibility that selection of the vehicle for verification testing is affecting the I/M process. Although this notice could be given by mail, the preferred approach would be to randomly select vehicles for verification testing at motor vehicle registration offices where motorists are attempting to complete the registration renewal process in person. With a temporary test facility in the immediate vicinity of the registration office, the verification testing could be completed expeditiously. This approach would minimize inconvenience to the motorist and minimize the risk that changes to the vehicle occur because the vehicle is selected for verification testing.

Under the option involving voluntary recruitment and screening of a larger sample, the procedure routinely employed by the California Air Resources Board should be used. State employees posing as the owner of the vehicle should take the vehicle to randomly selected inspection facilities to complete the I/M requirement and the verification testing should be subsequently performed at a facility operated by the state or a contractor.

Modeling Considerations

There are several additional issues that must be considered when comparing the results of a particular I/M program to the "benchmark" program. First, because the I/M rule requires that combined exhaust and evaporative HC emissions be evaluated when determining compliance with the performance standard, a means to translate the pressure test failure rate into an emission rate is needed. Second, programs that are deficient in some areas can make up for that shortfall by including additional vehicle ages or classes in their program that may not be included in the benchmark program. Finally, the proposed benchmark program used to determine compliance with the enhanced I/M requirements is a biennial program, while states have the option of implementing annual testing (or other

test frequency). As described below, these issues can be addressed through modeling exercises performed with EPA's MOBILE emission factors model.

Evaporative HC Emission Rates - Because it is not possible to measure evaporative emissions through an I/M short test procedure, a way to translate a parameter that is measured (e.g., pressure test failure rates) to an emission rate is needed. Once an evaporative emission rate is determined, the combined exhaust + evaporative HC emission rate can be compared to the benchmark program to determine whether equivalent results have been achieved.

The evaporative emission rate for the benchmark program is determined by running MOBILE with the same pressure test failure rate (by model year) as observed in that program. (Although the option to allow user-input evaporative control system failure rates is not included with the standard version of MOBILE5, such an analysis is not overly complex.) This model run would be performed under RVP and temperature conditions reflective of the sample program being analyzed. Similarly, a model run would also be performed using the pressure test failure rates under a non-I/M case. (The emission rates and corresponding failure rates for the benchmark program and the non-I/M case would be provided to the states by EPA.) The evaporative HC emission rate for the sample program would then be determined through linear interpolation based on the fleet-weighted pressure test failure rate observed in the sample program relative to the benchmark program and the non-I/M case.

Once the evaporative HC emission rates are determined as described above, they are combined with the IM240 exhaust HC emission rates determined by comparing the results obtained for a particular program to the benchmark program (the procedure for which is described in more detail below). However, to be consistent with MOBILE, the IM240-based exhaust HC emission rates for the benchmark program and the sample program are first converted to an FTP basis with IM240-to-FTP correlations provided by EPA. The combination of exhaust and evaporative emission rates calculated as described above is then used to determine compliance with the benchmark program (and, therefore, compliance with the performance standard).

Accounting for the Effects of Heavy-Duty Vehicle I/M - To provide additional emission reductions, or to make up for shortfalls in other program areas, some states include testing of heavy-duty gasoline vehicles (HDGVs) in their I/M programs. However, the basic program evaluation concept described above is focused on only light-duty cars and trucks to avoid the practical problems associated with the recruitment of heavy-duty vehicles. Thus, a method to account for the impact of other vehicle classes (specifically, HDGVs) is needed. To avoid the cost of additional testing, a state may use the modeled results from MOBILE to account for vehicle classes not tested in the program evaluation program.

The method proposed for this analysis requires MOBILE to be run under a non-I/M case and under the I/M conditions applicable for the heavy-duty component of the program. The difference in emission rates between the non-I/M case and the I/M case is then multiplied by the VMT fraction attributable to heavy-duty vehicles. This value represents

the fleet-weighted emission reduction attributable to the heavy-duty component of the program. For example, assume that the non-I/M HDGV exhaust HC emission rate calculated by MOBILE is 5.0 g/mi and the I/M emission rate is 4.5 g/mi. Also assume that this vehicle class accounts for 5% of the travel in the area. Thus, the heavy-duty component of the I/M program reduces fleet-average exhaust HC emissions by:

HDGV HC Fleet-Average I/M Reduction = (5.0 - 4.5) × 0.05 = 0.025 g/mi.

If the light-duty fleet accounts for 85% of the travel in an area, the 0.025 g/mi fleet-average reduction from HDGV testing translates to an effective reduction of 0.025 ÷ 0.85 = 0.029 g/mi from light-duty vehicles. Thus, inclusion of HDGVs in this program would reduce the reductions needed from the light-duty fleet by 0.029 g/mi.

Accounting for Annual Test Frequency - Because the benchmark program used to assess compliance with the enhanced I/M performance standard utilizes a biennial test frequency, some adjustment to account for the possibility of annual testing is necessary. Under the proposed I/M evaluation concept, compliance with the benchmark program is determined immediately after the I/M process is completed, which can be thought of as the bottom of the sawtooth in a classical deterioration, inspection, and repair I/M cycle. If the sample program being evaluated is an annual program, the bottom of the sawtooth does not have to be as low as the benchmark program to achieve the same reduction over time since the fleet will be inspected (and repaired) again the following year, while vehicles in the benchmark program continue to deteriorate until the next inspection cycle at year 2.

To account for an annual test frequency, the MOBILE model is first exercised using the sample program I/M parameters, which would include an annual inspection frequency, and then using a biennial inspection frequency. The fleet-average sample g/mi values (calculated as described below) are then reduced by the ratio of the annual MOBILE results to the biennial MOBILE results prior to comparing them with the benchmark program.

Proposed Program Evaluation Options

Based on the considerations outlined above, two different options are also proposed for demonstrating compliance with the EPA performance standard. Option 1 involves the following steps:

1. Recruit stratified random sample of 1,600 1967 and later model year passenger cars and light-duty trucks from population.* Recruited vehicles subject to the I/M program shall have just completed applicable I/M program requirements (passed initial test, passed after-repair test, received waiver, or exempted).

*Regardless of the age of vehicles included in program being evaluated, the sample is to include all 1967 and later models because that is the model year range included in the benchmark program.

2. Measure IM240 emissions of 1981 and later model year vehicles and conduct pressure test of evaporative system and gas cap. Measure idle/2500 emissions of pre-1981 model year vehicles and conduct pressure test of the gasoline cap. Convert idle/2500 emissions results for pre-1981 model year vehicles to IM240 using regression equations developed as outlined in this report.*

3. Calculate weighted average IM240 emissions and pressure test failure rate for sample based on the VMT distribution of the fleet for each model year range.

4. Based on the same VMT distribution, obtain weighted average IM240 emissions and pressure test failure rate for the benchmark program vehicles from EPA.

5. Adjust weighted average IM240 emissions and pressure test failure rate of sample based on difference between measured compliance rate and compliance rate of 96%. Do adjustment based on assumption that non-complying vehicles have HC and CO emissions 50% higher than emissions from complying vehicles, NOx emissions 10% higher,** and pressure test failure rates equal to the initial inspection failure rate for the benchmark program (provided by EPA). Adjust emission rate of sample to account for inclusion of heavy-duty vehicles or pre-1967 cars and light-trucks (as described earlier).

6. Compare results of steps 4 and 5 after adjusting results to account for fuel-related differences in emissions using the latest available fuel effects model approved by EPA.***

Based on Sierra's experience in I/M evaluation and in-use surveillance testing, the recommended recruitment options are as follows:

1. Test vehicles may be obtained from a random sample of computer-selected vehicles at a test-only inspection facility.

2. Test vehicles may be obtained from random samples of vehicles that are denied registration renewal unless they are submitted for testing.

---

*Although not recommended, the use of idle-only data could be considered if the sample size is adjusted accordingly.

**These are conservative assumptions based on the average difference in emissions predicted by MOBILE5 between vehicles subject to I/M and vehicles not subject to I/M.

***Results for the benchmark program should be adjusted by EPA to reflect a typical gasoline formulation for non-attainment areas (e.g., Phase 1 reformulated gasoline). A fuel effects model (e.g., the Complex Model) can then be used to adjust the results obtained for a particular program from local fuel specifications to the assumed use of the same (e.g., Phase 1) fuel.

3. Test vehicles may be obtained from random samples of vehicles that are selected by law enforcement officers at demographically balanced number of roadside locations.

4. Test vehicles may be selected from vehicles volunteered for testing to match the pattern of emissions-related defects in the fleet determined by a mandatory roadside inspection program.

Regardless of the recruitment option selected, testing must be done in a manner that prevents the personnel involved in the inspection process from knowing whether a vehicle has been selected for testing until such testing is completed. The alternatives that may be used are as follows:

1. In a test-only inspection program, testing must be done in a manner that prevents inspectors from knowing that the vehicle has been selected for testing until the vehicle has completed all program requirements (i.e., passed initial inspection or passed after-repair test).

2. In a test-and-repair program, testing must be independently performed at a special facility being operated by the state or a contractor to the state. The contractor may be a test-and-repair station different from the station that performed the official test, provided the station has passed audit testing by the state and provided the inspection is witnessed by a representative of the state. (This procedure is required to address the fact that test results are sometimes falsified at certain decentralized I/M stations.)

The compliance rate used in step number five of Option 1 shall be estimated using field survey data (e.g., parking lot surveys) analyzed in conjunction with vehicle registration records or from the results of random roadside inspections of a demographically-balanced sample.

Option 2 involves the following steps:

1. Recruit stratified random sample of 800 1967 and later model year vehicles that have just completed any applicable I/M requirements.

2. Test the 800 vehicle sample using both IM240 and an alternative short test and develop a correlation equation.

3. Recruit stratified random sample of 4,000-8,000 1967 and later model year vehicles from population. Recruited vehicles subject to the I/M program shall have just completed applicable I/M program requirements (if any).

4. Measure alternative test emissions and conduct pressure test of evaporative system and gas cap.

5. Using correlations developed in step number 2, calculate weighted average IM240 emissions and pressure test failure rate for sample based on the age distribution of the fleet.

6. Based on the same VMT distribution, obtain weighted average IM240 emissions and pressure test failure rate for the benchmark program vehicles from EPA.

7. Adjust weighted average IM240 emissions and pressure test failure rate of sample based on difference between measured compliance rate and compliance rate of 96%. Do adjustment based on assumption that non-complying vehicles have HC and CO emissions 50% higher than emissions from complying vehicles, NOx emissions 10% higher, and pressure test failure rates equal to the initial inspection failure rate for the benchmark program (provided by EPA). Adjust emission rate of sample to account for inclusion of heavy-duty vehicles or pre-1967 cars and light-trucks (as described earlier).

8. Compare results of steps 6 and 7 after adjusting results to account for fuel-related differences in emissions using the latest version of EPA's model for fuel effects.

Recruitment options and testing procedures used under Option 2 should be the same as those described for Option 1.

Data Analysis

Because the above-described procedure includes non-standard use of the MOBILE model to account for various differences between the benchmark program and other programs, states may need to seek assistance from EPA or contractors familiar with the intricacies of MOBILE to perform these analyses correctly. To facilitate the most efficient process, states should be encouraged to review detail testing and analysis plans with EPA prior to the initiation of program evaluation.

###

Appendix A
Sample Size Considerations

Overview

This appendix discusses the details involved in determining sample size. This overview section is intended to provide a summary of the important considerations in sample size determination.

Information Required for Sample Size Determination - A simple example of sample size determination is provided below. This equation applies to the determination of the sample size where the normal distribution applies. The variables in the equation are the sample size, n; the mean, $\bar{x}$; the standard deviation, s; the relative error, e_r; and the ordinate of the normal distribution, $z_{\alpha/2.}$ The value of $z_{\alpha/2}$ is found from tables of the cumulative normal distribution for a specified level of significance, α. The significance level, which is one minus the confidence level, must be specified by the experimenter. This is discussed further below.

$$n = \left(\frac{z_{\alpha/2}\, s}{e_r\, \bar{x}} \right)^2$$

The equation gives the sample size required to have the relative error in the mean be less than the specified relative error. This sample size does not guarantee that the relative error will *always* be less than the value specified. Instead, the user must specify a confidence level, say 90%. The confidence level is the probability that the sample size calculated by the equation will provide a relative error which is less than the value specified. Typical values of confidence level range from 80% to 99.5%. Once the confidence level is specified by the user, the value of $z_{\alpha/2}$ can be determined.*

To use this equation to determine the sample size, some estimate of the ratio $s/\bar{x}$ (the coefficient of variation, COV) is required. This can be found from previous experiments

*The values of $z_{\alpha/2}$ for commonly used confidence levels are shown in the table below.

Confidence Level, 1-α	80%	90%	95%	99%	99.9%
Significance level, α	0.2	0.1	0.05	0.01	0.001
Normal Ordinate, $z_{\alpha/2}$	1.282	1.645	1.960	2.576	3.290

Since the sample size depends on $(z_{\alpha/2})^2$, the significance level is of critical importance. Moving from a 90% to a 95% confidence level increases the sample size requirements by a factor of 1.42 [= $(1.96/1.645)^2$], i.e., a 42% increase in sample size.

on similar emissions measurements. In some sample size equations, based on the lognormal distribution, the standard deviation of the logarithms is used instead of the COV. However, all sample size determinations require the same kinds of inputs by the user: (1) the desired experimental error; (2) the confidence limit or the probability that the sample size will produce an error less than the one specified; and (3) some measure of the expected dispersion of the results, such as the COV.

The results presented in the body of the report are for a relative error of 10% and a confidence level of 90%. Tables provided in this appendix allow sample size to be determined for various values of relative error and confidence level.

Use of Nonnormal Statistics - Most statistical approaches are based on the use of the normal or Gaussian distribution. This distribution is symmetric about the mean and, in theory, extends from minus infinity to plus infinity. Emissions data do not follow the normal distribution. Except in the cases of instrument error, their values are all positive. In addition, they are not symmetric about the mean. Instead they have a "long tail" caused by a very small fraction of the vehicles with very large emissions. Such data are better described by the lognormal distribution. In this distribution, the logarithms of the original data are normally distributed. Figures A-1 to A-4 show the distribution of actual emissions data compared to theoretical distributions. The entire distribution is shown in Figure A-1. Figures A-2 to A-4 show portions of the distribution so that the differences between the actual and theoretical distributions can be seen more clearly. These figures show three theoretical distributions: (1) the normal distribution, (2) the lognormal distribution, and (3) the gamma distribution. The gamma distribution is another distribution with a long tail that has been proposed for emissions data. These figures show that the actual emissions data best follow the lognormal distribution. These figures are for the IM240 CO emissions from the CARB pilot program, but similar comparison plots can be obtained for other pollutant species and other emission tests. Consequently, the lognormal distribution was used for the sample size estimates presented here. An expanded discussion of these distributions, including the results of a χ^2 goodness-of-fit test, is provided in Appendix B.

Use of Stratified Sampling - The required sample sizes can be reduced by the use of stratified sampling. In this technique, the fleet is divided into different groups. Typically different ranges of model years are used, which act as a surrogate for different emission control technologies. By over-sampling the groups with high standard deviation, the overall standard deviation for a given sample size can be minimized. The charts and tables that give the sample size as a function of standard deviation allow the user to determine the sample size based on whatever sampling plan is being used. The specific examples provided below assume that stratified sampling has been used to reduce the required sample size. In these cases, separate regression equations are assumed to be determined for each model year group.

Figure A1

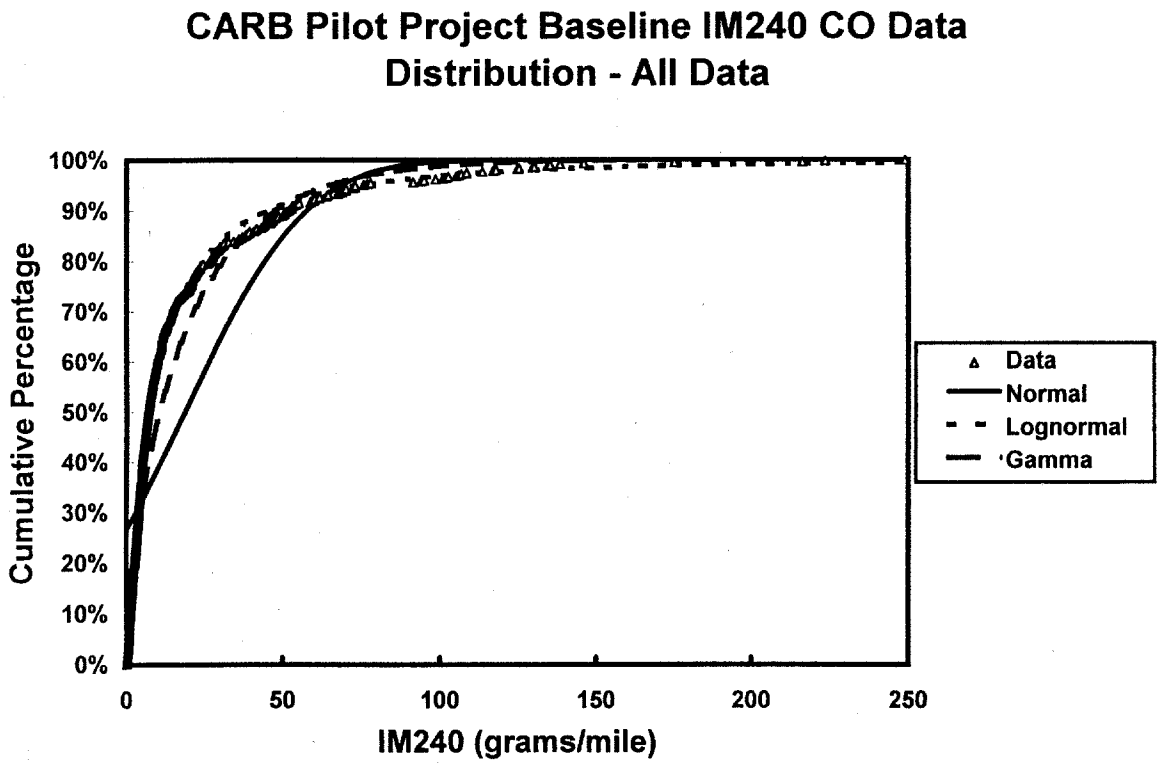


Figure A2

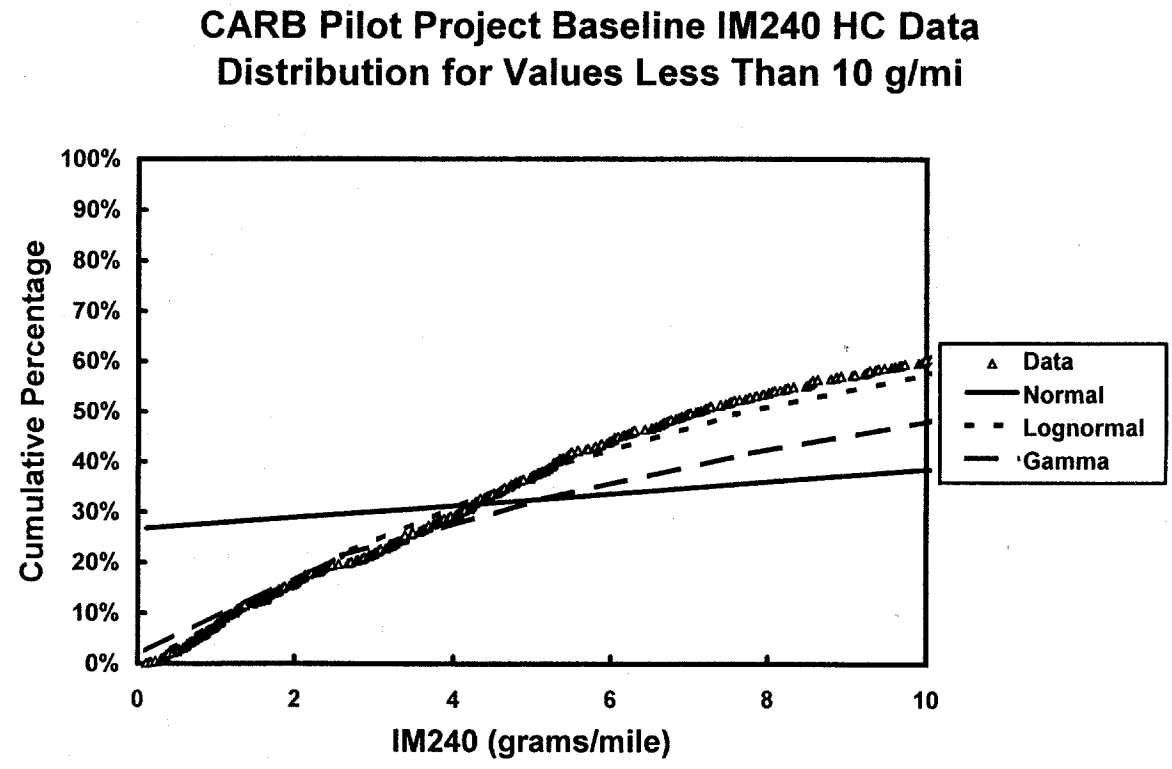


Figure A3

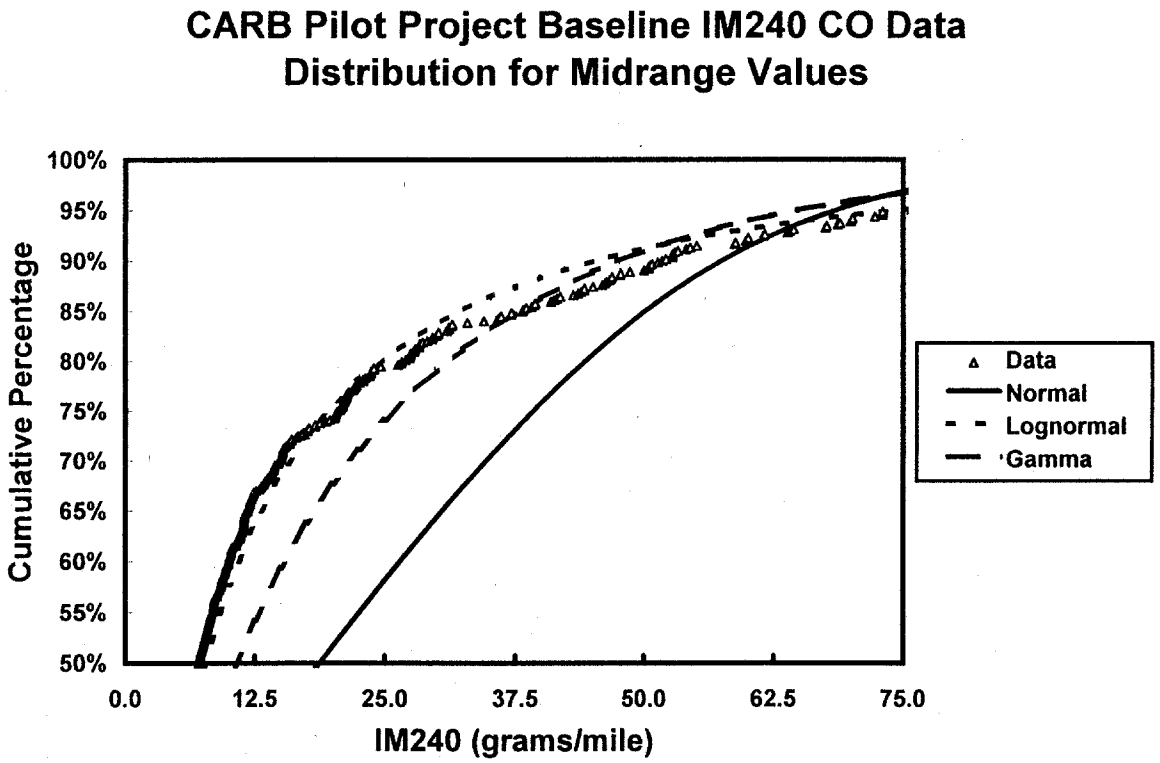
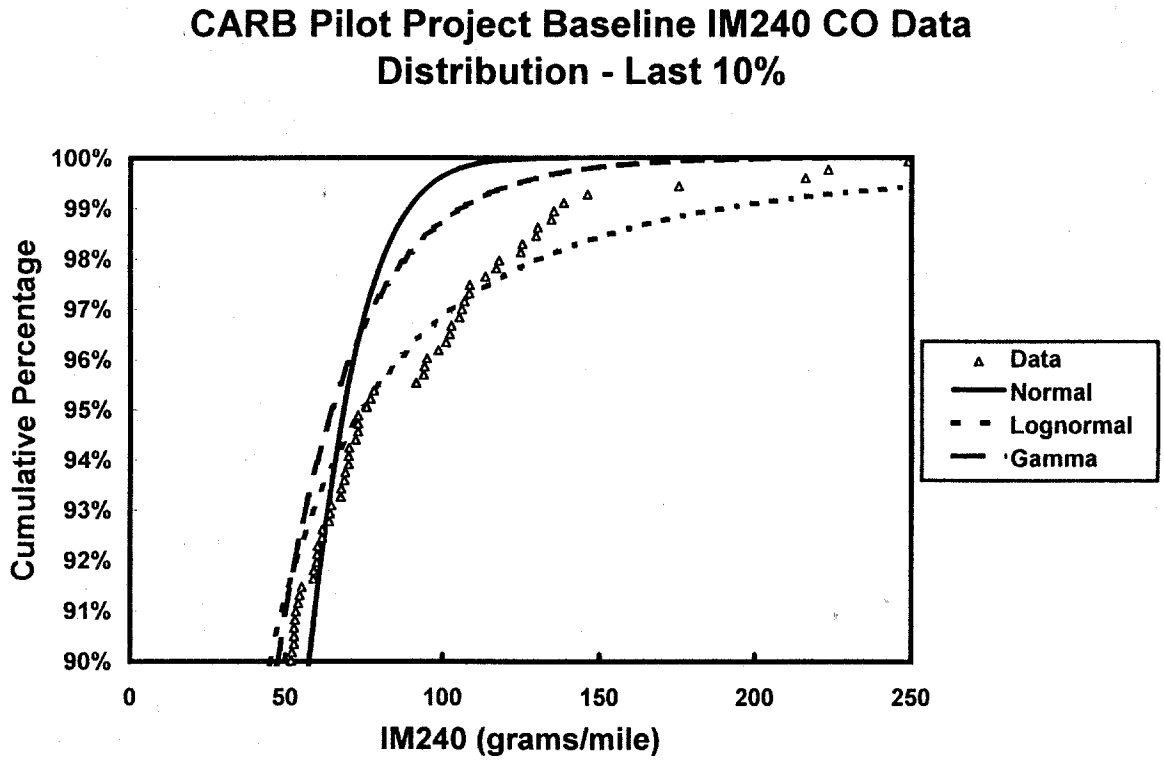


Figure A4



Organization of This Appendix - This appendix presents a derivation of the sample size equations. Following this overview section, a brief background section illustrates the method of obtaining sample size equations and demonstrates the application of such equations. This is followed by sections discussing nonnormal statistics, sample size considerations in regression equations, and stratified sampling. Finally, a set of tables is presented showing the required sample size as a function of the standard deviation, the confidence level and the relative error.

Background

Various methods are available to estimate the sample size required for a particular experiment. These methods differ because they are based on the use of different statistical tests to analyze the results of the experiment. All the methods have three common requirements:

- a specification of the confidence level for a type I error,
- a specification of a desired accuracy or minimum detectable difference, and
- an estimate of the standard deviation of the experiment to be performed.

In addition, some approaches require specifying the probability of a type II error.

Relation Between Sample Size and Accuracy - The required sample size can be estimated from the equation for the confidence limits on a population mean value, μ (sometimes called the true mean), as compared to the sample mean $\bar{x}$. The equation for these confidence limits is as follows:

$$\mu = \bar{x} \pm \frac{t_{\alpha/2,\, n-1}\, s}{\sqrt{n}} \qquad [A\text{-}1]$$

where

n is the sample size;

α is the significance level;

$t_{\alpha/2,n-1}$ is the ordinate of the Student's t distribution for a cumulative probability of $1 - \alpha/2$ and a sample size of n; and

s is the standard deviation of the sample.

The significance level, α, represents the probability that the difference between the sample mean and the population mean will be *larger* than the interval given by equation [A-1]. An alternative term, the confidence level, represents the probability that the actual difference between the sample and population mean will be *smaller* than the value calculated by equation [A-1]. This confidence level is $1-\alpha$.

For large sample sizes, the ordinate, $t_{\alpha,n}$, of the Student's t distribution is essentially the same as the ordinate of the normal distribution, z_α. This is a more convenient variable to use, since it does not depend on the sample size. With this substitution, the confidence interval is given by the following equation.

$$\mu = \bar{x} \pm \frac{z_{\alpha/2}\, s}{\sqrt{n}} \qquad \text{[A-2]}$$

Rearranging this equation, dividing by the sample mean, and taking the absolute value of the result gives an equation for the relative error, e_r, defined as follows:

$$e_r = \left| \frac{\mu - \bar{x}}{\bar{x}} \right| = \left| \frac{z_{\alpha/2}\, s}{\bar{x}\sqrt{n}} \right| \qquad \text{[A-3]}$$

Solving this equation for n gives the following result:

$$n = \left(\frac{z_{\alpha/2}\, s}{e_r\, \bar{x}} \right)^2 \qquad \text{[A-4]}$$

To use this equation, some estimate of the ratio $s/\bar{x}$ (the coefficient of variation, COV) and a specification of the desired relative error and the confidence level, $1-\alpha$, is needed. Typical values for the relative error and confidence level are 10% and 90%, respectively. Alternatively, the difference between the sample mean and the true mean can be defined as the absolute error, e_a. From equation [A-2] then,

$$e_a = \left| \mu - \bar{x} \right| = \left| \frac{z_{\alpha/2}\, s}{\sqrt{n}} \right| \qquad \text{[A-5]}$$

Solving equation [A-5] for the sample size gives an alternative to equation [A-4].

$$n = \left(\frac{z_{\alpha/2}\, s}{e_a} \right)^2 \qquad \text{[A-6]}$$

Equations [A-4] and [A-6] are similar equations. Each provides an estimate of the sample size that depends on the significance level, α; the acceptable error; and an estimate of the standard deviation or the coefficient of variation of the experiment to be performed.

If a comparison is made between the mean value of two procedures, the required sample size would be based on the confidence interval for the difference between the means of the samples from each procedure. This equation is written using the following notation: the sample means are denoted as $\bar{x}_b$ and $\bar{x}_a$; the corresponding population means are μ_b and μ_a respectively. The following equation gives the limits for the difference between the true means.

$$\mu_b - \mu_a = \bar{x}_b - \bar{x}_a \pm t_{\alpha/2,\, n_b + n_a - 2}\; s \sqrt{\frac{1}{n_b} + \frac{1}{n_a}} \qquad \text{[A-7]}$$

In this equation, n_b and n_a represent the number sampled from each procedure. The standard deviation, s, is the pooled estimate from both samples. It is computed from the individual standard deviations, s_b and s_a, by the following equation.*

$$s = \sqrt{\frac{(n_b - 1)\, s_b^2 + (n_a - 1)\, s_a^2}{n_b + n_a - 2}} \qquad \text{[A-8]}$$

Following the same analysis as shown above, a relative error can be defined as follows:

$$e_r = \left| \frac{(\mu_b - \mu_a) - (\bar{x}_b - \bar{x}_a)}{\bar{x}_b - \bar{x}_a} \right| = \left| \frac{z_{\alpha/2}\, s}{(\bar{x}_b - \bar{x}_a)} \sqrt{\frac{1}{n_b} + \frac{1}{n_a}} \right| \qquad \text{[A-9]}$$

If both the before- and after-repair samples have the same size, $n = n_b = n_a$, then the required sample size is given by the following equation:

$$n = 2 \left[\frac{z_{\alpha/2}\, s}{e_r\, (\bar{x}_b - \bar{x}_a)} \right]^2 \qquad \text{[A-10]}$$

*This equation assumes that both samples have the same variance, σ^2. The pooled estimate of the variance, s, is the estimate of this common variance.

It is also possible to define an absolute error by the following equation:

$$e_r = \left|(\mu_b - \mu_a) - (\bar{x}_b - \bar{x}_a)\right| = \left| z_{\alpha/2}\, s \sqrt{\frac{1}{n_b} + \frac{1}{n_a}} \right| \qquad \text{[A-11]}$$

The required sample size would be given by the following equation in the case where $n = n_b = n_a$.

$$n = 2\left[\frac{z_{\alpha/2}\, s}{e_r}\right]^2 \qquad \text{[A-12]}$$

Equations [A-10] or [A-12] would be used in place of equations [A-4] or [A-6], respectively, when the comparison between two separate processes is being made.

Examples of Sample Size Determinations - For emissions data, the standard deviation is usually larger than the mean value. In considering data from more than one pollutant, the one with the largest value of $s/\bar{x}$ is used to compute the sample size. An estimate of the required sample size was determined using data obtained by the California Air Resources Board in a study of I/M programs.* Shown in Table A-1 are the mean values and standard deviations for the portion of that program in which repairs were made on vehicles (the 2S94V2 program). The "after-repair" columns in this table are an average of the after-repair data for failing vehicles and the before-repair data for passing vehicles. Thus, these data represent the after-repair average for the entire fleet of passing and failing vehicles. The mean values for the differences of individual vehicles are the same as the differences in the means for the before-repair and after-repair fleets. However, the standard deviations for these differences are not related to the standard deviations for the individual fleets.**

For all three sets of data, the largest coefficient of variation occurs for hydrocarbons. These COV values are used to compute the sample size.

* Philip L. Heirigs and Thomas C. Austin, "Analysis of Data from the California Enhanced I/M Program," Sierra Research Report SR95-06-01 prepared for U. S. Environmental Protection Agency under Contract 68-C4-0056, Work Assignment 0-03, June 29, 1995.

**Not all passing vehicles in the pilot program were given full FTP tests. In order to account for this and still obtain the results for a representative vehicle fleet, the FTP results for the passing vehicles were approximated by using their IM240 results and a regression equation giving FTP results as a function of IM240 results.

Table A-1 Data from CARB Pilot Program									
	Before-Repair Results			After-Repair Results			Difference of Individual Vehicles		
Data Item	HC	CO	NOx	HC	CO	NOx	HC	CO	NOx
Count	469	469	469	469	469	469	469	469	469
Mean	2.240	23.719	1.426	1.523	17.345	1.080	0.717	6.374	0.346
Standard Deviation	6.967	35.216	1.344	5.891	29.131	0.819	3.448	23.001	0.967
Coefficient of Variation	3.110	1.485	0.942	3.867	1.679	0.758	4.812	3.609	2.794

This example uses a desired relative error of 10% with a confidence interval of 90%. The significance level, $\alpha = 0.1$, and the corresponding ordinate of the normal distribution, $z_{\alpha/2} = z_{0.05} = 1.645$, are found from tables of the Normal distribution.* For a study in which differences are available on individual vehicles, the required sample size is calculated from equation [A-4] as follows:

$$n = \left(\frac{(1.645)\ (4.812)}{0.1} \right)^2 = 6{,}264 \qquad \text{[A-13]}$$

If paired data from individual vehicles are not available, data from a before-repair fleet and an after-repair fleet would have to be obtained. The sample size required in this case can be found from equation [A-10]. It is first necessary to use equation [A-8] to obtain an estimate of the pooled standard deviation. This is found from the hydrocarbon data** in Table A-1 as follows:

*The values of $z_{\alpha/2}$ for commonly used confidence levels are shown in the table below. The actual calculations were made using full significant figures of Excel spreadsheet calculations. The value of $z_{\alpha/2}$ is found by the function call NORMINV(0.5 + 0.5 * CL, 0.0, 1.0), where CL is the name of the cell containing the confidence level.

Confidence Level, 1-α	80%	90%	95%	99%
Normal Ordinate, $z_{\alpha/2}$	1.282	1.645	1.960	2.576

**The sample size was calculated for all three pollutants; only the hydrocarbon results, which resulted in the largest sample size, are shown here.

$$s = \sqrt{\frac{(469 - 1)\,6.967^2 + (469 - 1)\,5.891^2}{469 + 469 - 2}} = 6.451 \qquad \text{[A-14]}$$

This standard deviation is used in equation [A-10] to compute the required sample size.

$$n = 2\left[\frac{(1.645)\ (6.451)}{(0.1)\ (2.240 - 1.660)}\right]^2 = 19{,}445 \qquad \text{[A-15]}$$

Comparing this result with the sample size of 6,264 for before- and after-repair tests on the same vehicle shows the advantage of same vehicle tests. The required sample size is reduced by more than a factor of three.

The discussion here has focussed on sample sizes for determining emission reductions. If only a measurement of emissions is wanted, equation [A-4] could be used to determine the required sample size. The after-repair fleet hydrocarbon data have the highest coefficient of variation, 3.867. For this COV, the required sample size for 10% error at a 90% confidence level is 4,047 vehicles.

Consideration of Nonnormal Distributions

Emissions data often have long tails and are not normally distributed. Typically emissions data follow a lognormal distribution. In this case, the usual relations that are valid for a normal distribution may be used but the logarithms of the emission data, rather than the data themselves, are used.* If x_i represents an individual emission data point, the operations with the lognormal distribution are then done using the variables y_i, where

$$y_i = \ln x_i \qquad \text{[A-16]}$$

The usual formulas for the mean, $\bar{y}$, and standard deviation, s_y, are applied to the individual values of y_i. Since the distribution of y_i is normal, the usual confidence limit equations apply to the mean value, $\bar{y}$. However, the mean value of the actual variable, x, is not simply related to the mean value, $\bar{y}$. In the theoretical lognormal distribution, the quantities μ and σ^2 usually refer to the mean and variance, respectively, of y. In this appendix, these quantities will be denoted as μ_{lnx} and σ_{lnx}^2. The true mean and variance of the original variable x will be written as μ_x and σ_x^2, respectively. For the lognormal distribution,

*Equations in this section for lognormal distribution are taken from Karl V. Bury, *Statistical Models in Applied Science*, John Wiley & Sons, 1975.

$$\mu_x = e^{\mu_{lnx} + \frac{\sigma^2_{lnx}}{2}} \quad [A\text{-}17]$$

The arithmetic mean of the original variable, $\bar{x}$, is not the most efficient estimator of the true mean, μ_x. The minimum variance estimator of μ_x, which is denoted as μ_x', is given by the following equation:

$$\mu_x^{/} = e^{\bar{y} + \frac{s^2_{lnx}}{2}} \; g(\frac{s^2_{lnx}}{2}) \quad [A\text{-}18]$$

In this equation, s_{lnx} is the usual estimate of the variance computed for the transformed variable, i.e.,

$$s^2_{lnx} = \frac{\sum_{i=1}^{n} (\ln x_i - \bar{y})^2}{n - 1} \quad [A\text{-}19]$$

The correction term, $g(s_{lnx}^2/2)$, can be expressed in the functional form, $g(\omega)$, where $\omega = s_{lnx}^2/2$.*

$$g(\omega) = 1 - \frac{\omega(\omega + 1)}{n} + \frac{\omega^2(3\omega^2 + 22\omega + 21)}{6n^2} \quad [A\text{-}20]$$

For small values of s_{lnx}^2 or large values of n, this correction term is approximately one.

Equation [A-1] for the confidence limits of the mean now gives a relationship between μ_{lnx} and $\bar{y}$. Thus, in place of equation [A-2], we could write the following equation:

$$\mu_{lnx} = \bar{y} \pm \frac{z_{\alpha/2} \; s_{lnx}}{\sqrt{n}} \quad [A\text{-}21]$$

Taking logarithms of equations [A-17] and [A-19], with the correction term g = 1, gives the following results.

*The function $g(\omega)$ is also used to provide a correction factor to the estimate of the variance. In that equation, which is not used here, ω is defined differently.

$$\mu_{lnx} = \ln\mu_x - \frac{\sigma^2_{lnx}}{2} \qquad [A\text{-}22]$$

$$\bar{y} = \ln\mu'_x - \frac{s^2_{lnx}}{2} \qquad [A\text{-}23]$$

Substituting [A-22] and [A-23] into equation [A-21] gives the following result:

$$\ln\mu_x - \frac{\sigma^2_{lnx}}{2} = \ln\mu'_x - \frac{s^2_{lnx}}{2} \pm \frac{z_{\alpha/2}\, s_{lnx}}{\sqrt{n}} \qquad [A\text{-}24]$$

The logarithm terms can be combined, using the relation that log(x) - log(y) = log(x/y), to introduce the relative error defined previously.

$$\ln\mu_x - \ln\mu'_x = \ln\left[\frac{\mu_x}{\mu'_x}\right] = \ln\left[\frac{\mu_x}{\mu'_x}\right] = \ln\left[\frac{\mu_x - \mu'_x}{\mu'_x} + 1\right] = \ln(e_r + 1) \qquad [A\text{-}25]$$

Combining equations [A-24] and [A-25] and solving for n gives the following equation for the relative error.

$$\ln(e_r + 1) = \frac{\sigma^2_{lnx}}{2} - \frac{s^2_{lnx}}{2} \pm \frac{z_{\alpha/2}\, s_{lnx}}{\sqrt{n}} \qquad [A\text{-}26]$$

The relative error will depend on the error bounds for the estimate of the variance. Since this variance is from a normal distribution, we can use the usual chi-squared confidence limits for the estimated variance.

$$\frac{(n-1)\,s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)\,s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \qquad [A\text{-}27]$$

where $\chi^2_{a,\nu}$ is the ordinate of the chi-squared distribution with ν degrees of freedom and probability a. The confidence limits for this quantity are not symmetric. Thus the confidence limits for $\ln(e_r+1)$ in equation [A-26] will not be symmetric. Instead we can consider the following rearrangement of true and actual variance terms in equation [A-26].

$$\frac{\sigma_{lnx}^2 - s_{lnx}^2}{2} = \frac{s_{lnx}^2}{2}\left[\frac{\sigma_{lnx}^2}{s_{lnx}^2} - 1\right] \qquad \text{[A-28]}$$

From equations [A-28] and [A-27], the variance terms in equation [A-26] are seen to have the following bounds:

$$\frac{s_{lnx}^2}{2}\left[\frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}} - 1\right] \le \frac{s_{lnx}^2}{2}\left[\frac{\sigma_{lnx}^2}{s_{lnx}^2} - 1\right] \le \frac{s_{lnx}^2}{2}\left[\frac{n-1}{\chi^2_{1-\frac{\alpha}{2},n-1}} - 1\right] \qquad \text{[A-29]}$$

In addition to the bounds determined by the variance, the $\ln(e_r + 1)$ term in equation [A-26] has a range of $\pm z_{\alpha/2,n-1}\, s_{lnx} / \sqrt{n}$. Thus, the total range for $\ln(e_r + 1)$ is given by the following inequality:

$$\frac{s_{lnx}^2}{2}\left[\frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}} - 1\right] - \frac{z_{\alpha/2}\, s_{lnx}}{\sqrt{n}} \le \ln(e_r + 1) \le \frac{s_{lnx}^2}{2}\left[\frac{n-1}{\chi^2_{1-\frac{\alpha}{2},n-1}} - 1\right] + \frac{z_{\alpha/2}\, s_{lnx}}{\sqrt{n}} \qquad \text{[A-30]}$$

To determine a relationship between sample size and relative error, we define the average absolute value of the upper and lower bounds for the $\ln(e_r + 1)$ term. The upper limit, UL, is positive, giving the following result.

$$UL = \left|\frac{s_{lnx}^2}{2}\left[\frac{n-1}{\chi^2_{1-\frac{\alpha}{2},n-1}} - 1\right] + \frac{z_{\alpha/2}\, s_{lnx}}{\sqrt{n}}\right| = \frac{s_{lnx}^2}{2}\left[\frac{n-1}{\chi^2_{1-\frac{\alpha}{2},n-1}} - 1\right] + \frac{z_{\alpha/2}\, s_{lnx}}{\sqrt{n}} \qquad \text{[A-31]}$$

The lower limit, LL, is negative, so taking absolute values results in the following equation:

$$LL = \left|\frac{s_{lnx}^2}{2}\left[\frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}} - 1\right] - \frac{z_{\alpha/2}\, s_{lnx}}{\sqrt{n}}\right| = \frac{s_{lnx}^2}{2}\left[1 - \frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}}\right] + \frac{z_{\alpha/2}\, s_{lnx}}{\sqrt{n}} \qquad \text{[A-32]}$$

We seek a sample size, n, so that the $\ln(e_r + 1)$ term is less than the average of these two limits.

$$\ln(e_r + 1) \le \frac{s_{lnx}^2}{4}\left[\frac{n-1}{\chi^2_{1-\frac{\alpha}{2},n-1}} - \frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}}\right] + \frac{z_{\alpha/2}\, s_{lnx}}{\sqrt{n}} \qquad \text{[A-33]}$$

This equation cannot be explicitly solved for the sample size n. Instead, a trial and error solution is required, which means that it is not necessary to approximate the ordinate of the t distribution by replacing it with the ordinate of the normal distribution. The $z_{\alpha/2}$ term in equation [A-33] can be replaced by $t_{\alpha/2,n-1}$, and the resulting equation can be solved for the relative error.

$$e_r \le e^{\frac{s_{lnx}^2}{4}\left[\frac{n-1}{\chi^2_{1-\frac{\alpha}{2},n-1}} - \frac{n-1}{\chi^2_{\frac{\alpha}{2},n-1}}\right] + \frac{t_{\alpha/2,n-1}\, s_{lnx}}{\sqrt{n}}} - 1 \qquad \text{[A-34]}$$

For an estimated standard deviation, s_{lnx}, and a desired significance level, α, equation [A-34] can be evaluated for various values of n until the desired value of relative error is found. Tables showing the relationship between relative error, sample size, standard deviation, and confidence level are provided in the final section of this appendix.

The data used to derive the means and standard deviations shown in Table A-1 can also be used to estimate the sample size using the log-normal distribution. Taking logarithms of the emission data and determining the mean and standard deviation of the logarithmic data gives the results shown in Table A-2. The mean of the log data and the standard deviation of the log data rows are simply found by the usual equations for the arithmetic mean and standard deviation. These are applied to the individual data after taking (natural) logarithms of the data. The mean emissions and the standard deviation of the emissions are found from the lognormal formulae for these quantities in equations [A-18] and [A-19]. A comparison of the mean emission data in A-1 and A-2 shows that the mean emissions are lower when using the formulae for the lognormal distribution.

Table A-2
Results from CARB Pilot Project Using Log Transforms

	Before Repair Results			After Repair Results		
Data Item	HC	CO	NOx	HC	CO	NOx
Count	469	469	469	469	469	469
Mean of Log Data	-0.220	2.529	0.014	-0.442	2.304	-0.157
Standard Deviation of Log Data	1.251	1.040	0.816	1.096	0.918	0.683
Mean Emissions	1.750	21.495	1.413	1.170	15.252	1.078
Standard Deviation of Emissions	3.361	29.770	1.370	1.768	17.454	0.829

The data in Table A-2 can be used to estimate the required sample size for FTP measurements. The largest standard deviation is for HC emissions before repairs. For the previously used values of a 90% confidence level (significance level, $\alpha = 0.1$) and a 10% error, the sample size is found to be 1660. This can be checked with equation [A-34]. The required probability distribution parameters are found from Excel spreadsheet functions.*

$$\chi^2_{\frac{0.1}{2},1660-1} = 1754.87 \qquad \chi^2_{1-\frac{0.1}{2},1660-1} = 1565.40 \qquad t_{\frac{0.1}{2},1660-1} = 1.6458$$

With these values and the log standard deviation for before-repair hydrocarbons, $s_{lnx} = 1.251$, the relative error is found from equation [A-34] as follows.

$$e_r \le e^{\frac{1.251^2}{4}\left[\frac{1660-1}{1565.40} - \frac{1660-1}{1754.87}\right] + \frac{(1.6458)(1.251)}{\sqrt{1660}}} - 1 = 0.09999 \qquad \text{[A-35]}$$

Previous sample size estimates for FTP emissions, using the sample size equation for a normal distribution, predicted a sample size of 4,047 vehicles for the given confidence limit and relative error.

The sample size requirements are reduced when the more nearly correct lognormal distribution is used. This results from the fact that the assumption of a normal distribution, coupled with the estimates of the mean and standard deviation for the normal distribution, results in a very wide normal distribution. Reducing the error in sampling from this incorrectly formed normal distribution requires a large number of samples. However, the lognormal distribution accounts for the long positive tails in the emissions data. Because the underlying theoretical distribution is better suited to the actual data, the analyses based on this distribution allow a more accurate estimate of the mean. Consequently, the estimated sample size is smaller when the lognormal distribution is used.**

---

*The functions TINV and CHIINV are used to find the ordinates of the t distribution and the χ^2 distribution, respectively. In Version 7.0 of Excel, the CHIINV function does not operate for large sample sizes (above approximately n = 1100). For such sample sizes, the approximate formula from Abramowitz and Stegun, *Handbook of Mathematical Functions and Tables*, Applied Mathematics Series 55, National Bureau of Standards, December 1965, equation 26.4.18, was used to find the ordinates of the χ^2 distribution.

**Additional information on the confidence limits for the log normal distribution was found after the completion of this appendix: Charles E. Land, "Hypothesis Tests and Interval Estimates," *Lognormal Distributions*, Edwin L. Chow and Kunio Shimzu, editors, Marcel Dekker, Inc., 1988, pp. 87-112. Dr. Land provided Sierra with a copy of his code for the determination of the confidence limits on the lognormal distribution. The

(continued...)

Errors in Relating Other Emission Measurements to IM240 Emissions

Equations [A-4] and [A-8] are used to determine the sample size required for a given relative error (and confidence level) in a measured quantity. These equations would be applied if a standard baseline test (e.g., the FTP test or an IM240 test) were used to measure the effectiveness of the I/M program. If a short test were used to evaluate the I/M program, the change in FTP or IM240 emissions that are equivalent to the change measured by the short test must be determined. The relationship between standard emission tests and short tests is usually estimated by linear regression equations. For purposes of discussion, we assume that the IM240 test is used as the standard test and the regression equation is written in the following form:

$$E_{species,\,IM240} = a_{species} + b_{species}\, E_{species,\,st} \qquad \text{[A-36]}$$

where

E	is the emission rate in grams per mile;
species	is the individual pollutant, HC, CO, and NOx;
st	is the subscript denoting the short test;
IM240	is the subscript denoting the IM240 test;
$a_{species}$	is the intercept in the linear regression equation for the given pollutant; and
$b_{species}$	is the slope in the linear regression equation for the given pollutant.

The mean IM240 emissions computed from this equation have an uncertainty resulting from the regression.

$$\bar{E}_{species,\,IM240} = a_{species} + b_{species}\, \bar{E}_{species,\,st} \pm \Delta E_{species,reg} \qquad \text{[A-37]}$$

The error term, $\Delta E_{species,reg}$, is based on the confidence limits for the mean value of IM240 emissions for a given value of the short test. This corresponds to the difference between the (unknown) true value of IM240 emissions (at a specified short test value) and that

---

[**](...continued)
relative error predicted by Land's approach is higher, for a given sample size, standard deviation, and confidence limit, than the results obtained here. Thus, his approach would require a larger sample size than the ones presented here. Sierra is grateful to Dr. Land for providing us a copy of his confidence interval code.

computed from the regression. These confidence limits, CL, are given by the following equation:*

$$\Delta E_{Species,reg} = t_{\alpha/2,\, n_{reg}-2}\; s_{y|x} \sqrt{\frac{1}{n_{reg}} + \frac{[\, E_{species,st,spec} - \overline{E}^{*}_{species,st}\,]^2}{\sum_{i=1}^{n_{reg}} [\, E^{*}_{i,species,st} - \overline{E}^{*}_{species,st}\,]^2}} \qquad \text{[A-38]}$$

In this equation,

n_{reg}	is the sample size used to develop the regression equation,
$s_{y\|x}$	is the standard regression error of the estimated variable,
$E_{species,st,spec}$	is the specified short test emissions value at which the confidence limits are to be computed,
$\overline{E}^{*}_{species,st}$	is the mean value found in the regression analysis data set, and
$E^{*}_{i,species,st}$	is an individual emissions value in the regression analysis data set.

The regression sample size is selected to ensure that the overall regression error is less than some desired relative error. In this case, the relative error is defined in terms of the $\Delta E_{species,reg}$ term and the mean IM240 emissions.

$$e_r = \frac{\Delta E_{species,\, reg}}{\overline{E}_{species,IM240}} \qquad \text{[A-39]}$$

Squaring this equation and substituting equation [A-38] for $\Delta E_{species,reg}$ gives the following result:

$$e_r^2 = \frac{t^2_{\alpha/2,\, n_{reg}-2}\; s^2_{y|x}}{\overline{E}^2_{Species,\, IM240}} \left[\frac{1}{n_{reg}} + \frac{[\, E_{species,st,spec} - \overline{E}^{*}_{species,st}\,]^2}{\sum_{i=1}^{n_{reg}} [\, E^{*}_{i,species,st} - \overline{E}^{*}_{species,st}\,]^2} \right] \qquad \text{[A-40]}$$

*There are two confidence-limit equations for regressions. The one used here is the confidence level on the *mean of the dependent variable* (the IM240 result) corresponding to a particular value of the independent variable (the short test result.) The other equation, not shown here, is for the confidence limits corresponding to *one future observation* of the independent variable. The confidence limits for the single observation are greater than the confidence limits for the mean.

This equation is solved for the required regression sample size, n_{reg}, with the following simplifying assumptions: (1) the value of $t_{\alpha/2,n-2}$ is replaced by the normal distribution ordinate, $z_{\alpha/2}$; and (2) the specified value at which the confidence limits are evaluated is assumed to be the same as the mean value. With these assumptions, the regression sample size is given by the following result:

$$n_{reg} = \left(\frac{z_{\alpha/2} \; s_{y|x}}{e_r \; \bar{E}_{Species,\, IM240}} \right)^2 \qquad \text{[A-41]}$$

From data in the CARB pilot program, it is possible to obtain a regression of IM240 emissions as a function of ASM scores. For each species, regressions were done using only the ASM concentrations and repeated using the product of ASM concentration times inertia weight. The use of the inertia weight made only a small improvement in the regressions. Regressions were done using the ASM 50/25 and ASM 25/25 scores separately and using a single regression against both ASM tests. The combined regression did not give any improvement over the regressions obtained with the single ASM25/25 score. The regression of IM240 HC versus ASM 25/25 HC concentration had a $s_{x|y}$ value of 2.631g/mi when the mean IM240 emissions were 1.545 g/mi. With these values, the required sample size for a 90% confidence level ($z_{\alpha/2}$ = 1.645) and a 10% error is found as follows:

$$n_{reg} = \left(\frac{(1.645) \; (2.631)}{(1.545) \; (0.1)} \right)^2 = 786 \qquad \text{[A-42]}$$

Consideration of Nonnormal Distributions - The assumption of a lognormal distribution applied to emissions data above can also be used in regressions. This has an additional advantage in regression analysis. A basic assumption in regression analysis is that the independent variable (e.g., the variable x in the regression y = a + b x) is either known exactly or is a random variable from a normal distribution. When using regression analysis to consider two sets of emission data, as was done above, a log transformation on the data will satisfy the condition for regression analysis that a random independent variable is normally distributed. Using a log transformation on equation [A-37] gives the following result:

$$\ln \bar{E}_{species,\, IM240} = a^*_{species} + b^*_{species} \ln \bar{E}_{species,\, st} \pm \Delta(\ln E_{species,reg}) \qquad \text{[A-43]}$$

The asterisks on the regression coefficients in equation [A-43] show that these coefficients are different from the ones used in the original regression. The $\Delta(\ln E_{species,reg})$ term is still

obtained from equation [A-38],* but the $s_{y|x}$ term in that equation is now found from the regression of the logarithms and will be designated with a (*) superscript to show its difference from the regular term. With this notational change, the substitution of the normal ordinate, $z_{\alpha/2}$, for the t-distribution ordinate, and the evaluation of the final term at the mean value, $\Delta(\ln E_{species,reg})$ is written as follows:

$$\Delta(\ln E_{Species,reg}) = \frac{z_{\alpha/2}\, s^*_{y|x}}{\sqrt{n_{reg}}} \qquad \text{[A-44]}$$

This is the confidence limit in the mean of the log of the emissions (at a given value for the emissions from the short test.) We are interested in the sample size necessary to obtain a certain accuracy in the emissions themselves. This requires an analysis of the confidence limits on the emissions. The analysis steps between equations [A-21] and [A-34] that were used to determine the relationship between relative error and sample size for a single measured variable can be repeated here with the regression error for the logarithms, $s^*_{y|x}$, replacing the standard deviation, s_{lnx}, for the single variable. Thus, the regression case can be taken from equation [A-34] by substituting $s^*_{y|x}$ for s_{lnx} and n_{reg} for n. This gives the following result:

$$e_r \le e^{\frac{(s^*_{x|y})^2}{4}\left[\frac{n_{reg}-1}{\chi^2_{1-\frac{\alpha}{2},n_{reg}-1}} - \frac{n_{reg}-1}{\chi^2_{\frac{\alpha}{2},n_{reg}-1}}\right] + \frac{t_{\alpha/2,n_{reg}-1}\, s^*_{x|y}}{\sqrt{n_{reg}}}} - 1 \qquad \text{[A-54]}$$

The regression used to compute the sample size in equation [A-19] was repeated using a log transformation. The resulting regression had a standard error, $s^*_{y|x} = 0.8551$, when both the ASM 50/15 and ASM 25/25 results were used in the regression. The sample size for the same 90% confidence level and 10% relative error considered before is found by trial and error to be 565 vehicles. This can be verified as shown below. The ordinates of the χ^2 and t distributions for the significance level of 0.1 and the sample size of 565 are as follows:

$$\chi^2_{\frac{0.1}{2},565-1} = 615.11 \qquad \chi^2_{1-\frac{0.1}{2},565-1} = 505.16 \qquad t_{\frac{0.1}{2},565-1} = 1.6476$$

With these values and $s^*_{y|x} = 0.8551$, the relative error of 10% can be verified using equation [A-54] as follows.

*The emissions variables in the final term in Equation [A-15] would be replaced by the logarithms of the emissions. However, we will continue to evaluate the final term at the mean emissions point where this term is zero.

$$e_r \le e^{\frac{0.8551^2}{4}\left[\frac{565-1}{505.16}-\frac{565-1}{615.11}\right]+\frac{(1.6476)(0.8551)}{\sqrt{1660}}} - 1 = 0.10000$$

Stratified Sampling Error

Stratified Sampling Notation - In stratified sampling, the random selection of vehicles is weighted by their population in the overall vehicle fleet. These different groups are usually taken to be different model years, but they could also represent different vehicle emission control technologies. Defined below are the terms used in the discussion of stratified sampling.*

- N = the total number of vehicles in the fleet
- N_g = the number of vehicles in the fleet in group "g"
- $\bar{x}_g$ = the mean emissions for group "g"
- s_g = the standard deviation of emissions for group "g"
- n = the sample size for all groups
- n_g = the sample size for group "g"

The values of N_g and N always enter the calculations as the ratio N_g/N; this means that it is sufficient to know the fraction of the vehicles in each group. For convenience, we denote the fraction of vehicles in the fleet as F_g and the fraction of vehicles in the sample as f_g. These are defined as follows:

$$F_g = \frac{N_g}{N} \qquad f_g = \frac{n_g}{n} \qquad \text{[A-55]}$$

The mean value, $\bar{x}$, and the standard deviation, s, for the entire fleet are computed from the individual means and standard deviations for each group by the following equations:

$$\bar{x} = \sum_g F_g \bar{x}_g \qquad \text{[A-56]}$$

and

$$s = \sqrt{\sum_g \frac{F_g^2 s_g^2}{f_g}} \qquad \text{[A-57]}$$

* The basic formulae in this section are taken from Chapter Five in W. G. Chochran, *Sampling Techniques* (3rd ed.), Wiley, 1977. The corrections for finite populations are neglected in these equations because the sample size is always expected to be less than 1% of the vehicle fleet.

In proportional sampling, the proportion of vehicles in each group in the sample is the same as the proportion of the group in the fleet (i.e., $f_g = F_g$ for proportional sampling). However, the variance of the stratified sample will be minimized, for a fixed total sample size, n, if the sample from each group is determined from the following equation:

$$f_g = \frac{F_g \, s_g}{\sum_g F_g \, s_g} \qquad \text{[A-58]}$$

This equation biases the sample so that vehicle groups with a greater variance will be over-represented in the sample.* The net effect of this selection process is to provide an overall sample that has a lower standard deviation. This fraction will depend on the specific pollutant.

Stratified Sampling Single Emission Measurements - The sample size required for a single emission test, using stratified samples, is found from equation [A-4] for data that follow the normal distribution or equation [A-34] for data that follow a lognormal distribution. In applying these equations to stratified sampling, the values of $\bar{x}$ and s must be computed from equations [A-56] and [A-57], respectively. This means that some estimate of the standard deviation and mean for each group must be known in addition to the fractions of vehicles in the fleet and vehicles in the sample. The fleet fraction is expected to come from data in EPA's MOBILE model or from locally generated data on vehicle distributions. The sample fraction is usually found from equation [A-58] to provide an optimum sample.

Using Other Emission Tests to Determine Fleet-Average Emissions - When short emission tests are used to determine FTP or IM240 emissions in stratified sampling, it is usually more accurate to have a separate regression equation for each stratum. However, the overall concern is not with the accuracy of each stratum, but with the overall accuracy of the fleet regression equations. Thus, a single regression statistic should be computed for the overall combination of regression equations. A regression technique that accomplishes this is outlined below.

The individual regression equations obtain predicted values of some variable, y, as a function of another variable, x, for different model year (MY) ranges. This can be written as follows:

*Note that although the sample is biased, the computation of the mean and variance will use the appropriate weighting so that the weighted results are representative of the actual fleet in any given year.

$$y = \begin{cases} a_1 + b_1 x & MY_0 < MY \le MY_1 \\ a_2 + b_2 x & MY_1 < MY \le MY_2 \\ \bullet & \\ a_n + b_n x & MY_{n-1} < MY \le MY_n \end{cases} \qquad \text{[A-59]}$$

If the data are sorted and individual regressions are fitted, the individual regression statistics can be determined for each equation, but the statistics found for the individual regressions do not apply to the entire data set.

An analysis of the entire data set, preserving model year groupings, can be developed using the delta variables defined below.

$$\delta_i = \begin{cases} 1 & MY_{i-1} < MY \le MY_i \\ 0 & otherwise \end{cases} \qquad \text{[A-60]}$$

With this definition, the set of equations represented by [A-59] can be replaced by the following single equation:

$$y = a_1 \delta_1 + b_1 \delta_1 x + a_2 \delta_2 + b_2 \delta_2 x + \bullet \bullet \bullet + a_n \delta_n + b_n \delta_n x \qquad \text{[A-61]}$$

The delta variables select the appropriate regression constants for the model-year range specified. Equation [A-61] can be written as a simple, no-intercept, linear regression model.

$$y = \sum_{i=1}^{2n} c_i z_i \qquad \text{[A-62]}$$

where

$$c_i = \begin{cases} a_{(i+1)/2} & odd\ \ i \\ b_{i/2} & even\ \ i \end{cases} \qquad \text{[A-63]}$$

and

$$z_i = \begin{cases} \delta_{(i+1)/2} & odd\ \ i \\ \delta_{i/2}\ x & even\ \ i \end{cases} \qquad \text{[A-64]}$$

Performing a single regression with equation [A-62], using data for all model years, will give the same regression coefficients as individual regressions for individual model year groups. However, the single regression would provide a single set of regression statistics giving the standard error, $s_{y|x}$, and R^2 value for the single regression.

As an example of this, the data from the 1994 California Air Resources Board pilot project on various vehicle inspection and maintenance tests were used to obtain regressions of acceleration simulation mode (ASM) tests with IM240 results. Data from a total of 801 vehicle tests were available, including some tests on the same vehicle before and after repairs. These data were used to obtain the following regressions for hydrocarbon emissions:*

IM240 = 0.42214 + 0.012796 ASM2525	$R^2 = 0.66$	$s_{y\|x} = 2.63$ (all model years)
IM240 = 2.19142 + 0.012233 ASM2525	$R^2 = 0.86$	$s_{y\|x} = 2.20$ (pre-1975)
IM240 = 1.182423 + 0.015278 ASM2525	$R^2 = 0.66$	$s_{y\|x} = 5.87$ (1975-1980)
IM240 = 0.36072 + 0.00779 ASM2525	$R^2 = 0.71$	$s_{y\|x} = 0.895$ (1981 and later)

A single regression with all data using equation [A-62] provided the same coefficients as the regressions for individual model year groups, with an overall R^2 of 0.73 and standard error, $s_{y|x} = 2.36$. This gives some improvement in R^2 over the value of 0.66 obtained with only two regression coefficients for all model years. However, the standard error for the combined analysis is actually higher than the individual analysis for the 1981 and later model year group. Confidence interval equations that use $s_{y|x}$ should be based on the values found from regressions on individual model year groups when considering only applications to a specific model year group.

Sample Calculations Using Stratified Sampling - The data that were used to obtain the mean values shown in Tables A-1 and A-2 were stratified into three model year groups: (1) 1974 and earlier vehicles, representing noncatalyst technology; (2) 1975-1980 model years, representing oxidation catalyst technology; and (3) 1981 and later model years, representing three way catalyst technology. The means and standard deviations for these model year groups are shown in Table A-3. This table also shows the fleet fractions and the sample fractions that should be used to provide an optimum sample. These sample fractions are found from equation [A-58].

*The regression is based on the ASM test, which simulates operation at 25 mph and 25% of the maximum load on the normal Federal test procedure (FTP). This is known as the ASM2525 test. IM240 data and $s_{y|x}$ values are in units of grams/mile; ASM2525 data are in ppm. Regressions with ASM data sometimes use the product of the ASM concentration and the vehicle weight. For the data set used here, this approach did not provide any significant improvement in the regression results.

Table A-3 Data for Stratified Sampling Sample Size Using 90% Confidence Level and 10% Relative Error							
Group and Fleet Fraction	Data Entry	Original Before Repair FTP Data			Logarithms of FTP Data		
		HC	CO	NOx	Ln(HC)	Ln(CO)	Ln(NOx)
1974-and-earlier model years	Mean	9.082	66.711	2.859	1.923	4.041	0.889
	Std Dev	8.764	44.5	1.604	0.689	0.559	0.605
0.0071	Sample Fraction	0.0258	0.0126	0.0114	0.0052	0.0047	0.0059
1975-1979 model years	Mean	7.463	59.482	2.772	1.221	3.72	0.771
	Std Dev	17.452	52.829	2.051	1.048	0.897	0.746
0.0325	Sample Fraction	0.2351	0.0686	0.0668	0.0359	0.0346	0.0332
1981-and-later model years	Mean	0.94	15.163	1.121	-0.601	2.237	-0.164
	Std Dev	1.857	23.93	0.958	0.948	0.843	0.731
0.9604	Sample Fraction	0.739	0.919	0.922	0.959	0.961	0.961
Overall Results for Stratified Sample	Mean	1.21	16.969	1.187	-0.524	2.298	-0.126
	Std Dev	2.413	25.015	0.999	0.949	0.843	0.73
	Sample size	1076	588	192	754	544	370

The final row of Table A-3 gives the sample size as computed from equation [A-4] for the raw data (assuming a normal distribution) and from equation [A-34] for the logarithmic data (assuming a lognormal distribution). The use of stratified sampling alone is seen to reduce the required sample size. The original estimate of the sample size using equation [A-4] without a stratified sample was 4,074 vehicles. With stratification, this is reduced to 1,076 vehicles. The original estimate using the lognormal distribution was 1,660 vehicles; with stratified sampling, this is reduced to 754 vehicles.

Data Tables

Table A-4 gives the required sample size for data following the lognormal distribution. Separate tables are available for user-selected confidence levels of 99.9%, 99.5%, 99%, 95%, 90% and 80%. Each table gives the sample size as a function of the expected standard deviation, s, of the logarithmic data and the desired relative error in the original data.

Table A-4
Sample Size Table for Variables Following a Lognormal Distribution
(Standard Deviations below are the Standard Deviations of the Logarithm of the Original Data.)

Relative Error	Sample size for various standard deviations and relative errors and a confidence level of 99.9%									
	s = 0.1	s = 0.2	s = 0.3	s = 0.5	s = 0.8	s = 1.0	s = 1.5	s = 2.0	s = 2.5	s = 3.0
1.0%	1,260	5,707	14,469	50,098	144,071	318,707	1,044,848	2,549,594	5,235,641	9,588,954
2.0%	323	1,446	3,659	12,656	36,383	80,476	263,815	643,734	1,321,982	2,421,000
3.0%	149	653	1,647	5,685	16,335	36,125	118,412	288,927	593,341	1,086,625
5.0%	59	245	610	2,093	6,002	13,266	43,470	106,055	217,786	398,841
7.5%	30	115	282	957	2,737	6,044	19,791	48,277	99,130	181,533
10.0%	20	69	166	555	1,581	3,485	11,400	27,802	57,082	104,527
12.5%	15	48	111	367	1,039	2,286	7,469	18,209	37,382	68,450
15.0%	12	36	81	263	741	1,627	5,309	12,936	26,553	48,618
20.0%	10	24	51	158	439	960	3,125	7,607	15,609	28,576
Relative Error	Sample size for various standard deviations and relative errors and a confidence level of 99.5%									
	s = 0.1	s = 0.2	s = 0.3	s = 0.5	s = 0.8	s = 1.0	s = 1.5	s = 2.0	s = 2.5	s = 3.0
1.0%	917	4,153	10,530	36,458	104,844	231,931	760,362	1,855,384	3,810,328	6,978,093
2.0%	235	1,053	2,663	9,210	26,477	58,564	191,984	468,456	962,045	1,761,827
3.0%	109	476	1,199	4,138	11,888	26,290	86,172	210,259	431,788	790,759
5.0%	43	178	444	1,523	4,368	9,655	31,634	77,179	158,489	290,248
7.5%	23	84	205	697	1,992	4,399	14,403	35,132	72,140	132,107
10.0%	15	51	121	404	1,151	2,536	8,297	20,233	41,540	76,068
12.5%	12	35	81	267	756	1,664	5,436	13,252	27,205	49,814
15.0%	10	27	59	192	539	1,184	3,864	9,415	19,324	35,381
20.0%	7	18	37	116	320	699	2,274	5,537	11,360	20,796
Relative Error	Sample size for various standard deviations and relative errors and a confidence level of 99%									
	s = 0.1	s = 0.2	s = 0.3	s = 0.5	s = 0.8	s = 1.0	s = 1.5	s = 2.0	s = 2.5	s = 3.0
1.0%	773	3,497	8,867	30,700	88,284	195,298	640,273	1,562,331	3,208,557	5,875,819
2.0%	198	887	2,243	7,756	22,295	49,315	161,662	394,464	810,106	1,483,593
3.0%	91	401	1,010	3,484	10,010	22,138	72,561	177,050	363,589	665,860
5.0%	36	150	374	1,283	3,679	8,130	26,638	64,989	133,456	244,404
7.5%	19	71	173	587	1,678	3,704	12,128	29,584	60,746	111,242
10.0%	12	43	102	341	969	2,136	6,987	17,037	34,979	64,053

Table A-4
Sample Size Table for Variables Following a Lognormal Distribution
(Standard Deviations below are the Standard Deviations of the Logarithm of the Original Data.)

12.5%	10	30	69	225	637	1,401	4,578	11,159	22,908	41,946
15.0%	8	23	50	162	454	997	3,254	7,928	16,272	29,793
20.0%	6	15	31	97	270	589	1,915	4,662	9,566	17,511
Relative Error	Sample size for various standard deviations and relative errors and a confidence level of 95%									
	s = 0.1	s = 0.2	s = 0.3	s = 0.5	s = 0.8	s = 1.0	s = 1.5	s = 2.0	s = 2.5	s = 3.0
1.0%	448	2,025	5,134	17,775	51,116	113,074	370,696	904,566	1,857,627	3,402,028
2.0%	115	514	1,299	4,491	12,909	28,553	93,599	228,388	469,024	858,951
3.0%	53	232	585	2,018	5,796	12,818	42,012	102,508	210,509	385,524
5.0%	21	87	217	743	2,130	4,708	15,424	37,628	77,270	141,505
7.5%	12	41	101	340	972	2,145	7,023	17,129	35,172	64,407
10.0%	7	25	59	198	562	1,237	4,046	9,865	20,253	37,086
12.5%	6	18	40	131	369	812	2,651	6,462	13,264	24,287
15.0%	5	14	30	94	264	578	1,884	4,591	9,422	17,251
20.0%	4	9	19	57	157	342	1,110	2,700	5,539	10,140
Relative Error	Sample size for various standard deviations and relative errors and a confidence level of 90%									
	s = 0.1	s = 0.2	s = 0.3	s = 0.5	s = 0.8	s = 1.0	s = 1.5	s = 2.0	s = 2.5	s = 3.0
1.0%	315	1,427	3,616	12,519	36,001	79,638	261,087	637,080	1,308,395	2,396,063
2.0%	81	362	915	3,163	9,092	20,110	65,922	160,855	330,330	604,957
3.0%	38	164	412	1,421	4,083	9,028	29,590	72,197	148,266	271,526
5.0%	15	62	153	524	1,501	3,316	10,863	26,502	54,422	99,662
7.5%	8	29	71	240	685	1,511	4,947	12,065	24,772	45,363
10.0%	6	18	42	140	396	872	2,850	6,948	14,265	26,120
12.5%	5	13	29	92	260	572	1,868	4,551	9,342	17,106
15.0%	4	10	21	67	187	408	1,328	3,234	6,637	12,150
20.0%	3	7	14	40	111	241	782	1,902	3,902	7,142
Relative Error	Sample size for various standard deviations and relative errors and a confidence level of 80%									
	s = 0.1	s = 0.2	s = 0.3	s = 0.5	s = 0.8	s = 1.0	s = 1.5	s = 2.0	s = 2.5	s = 3.0
1.0%	192	866	2,196	7,600	21,855	48,344	158,489	386,728	794,218	1,454,555
2.0%	50	220	556	1,921	5,520	12,208	40,018	97,646	200,525	367,241
3.0%	23	100	251	863	2,479	5,481	17,963	43,827	90,003	164,828

Table A-4
Sample Size Table for Variables Following a Lognormal Distribution
(Standard Deviations below are the Standard Deviations of the Logarithm of the Original Data.)

5.0%	10	38	94	318	912	2,014	6,595	16,089	33,037	60,500
7.5%	6	18	44	146	416	918	3,003	7,324	15,038	27,538
10.0%	4	12	26	85	241	530	1,731	4,219	8,660	15,857
12.5%	3	8	18	57	159	348	1,134	2,764	5,672	10,385
15.0%	3	6	13	41	113	248	807	1,964	4,029	7,376
20.0%	3	5	9	25	68	147	475	1,155	2,369	4,336

Appendix B
Goodness-of-Fit Tests for Various Distributions

A χ^2 test was used to provide a quantitative measure of the goodness of fit of the various distributions.* In this test the observations are ranked and grouped into k bins. The number of observations in the i^{th} bin, n_i, is determined and compared to the expected number of observations in the bin. The expected number in the i^{th} bin is the product of the total number of observations, n, multiplied by the probability, p_i, that a random observation from the postulated distribution will fall in this bin. The value of p_i is found from the cumulative distribution, F(x); i.e., F(x) is the probability that a random observation, y, is less than equal to x. If the upper and lower bounds of bin i are denoted as x_i and x_{i+1}, $p_i = F(x_{i+1}) - F(x_i)$. The χ^2 test is based on the following statistic.

$$C = \sum_{i=1}^{k} \frac{(n_i - n p_i)^2}{n p_i} \qquad [B\text{-}1]$$

This statistic has a χ^2 distribution for sufficiently large values of $n p_i$.** The degrees of freedom associated with this statistic are equal to the number of bins minus one minus the number of parameters in the distribution that have been estimated from the data. The normal, log-normal and gamma distribution are all two parameter distributions, thus the degrees of freedom are the number of bins minus three.

For all the calculations reported here 20 bins were used. The sizes of the bins were set so that each bin would contain 5% of the total observations. With a total of 20 bins, the χ^2 test had 17 degrees of freedom for the two-parameters distributions used here. The results of the χ^2 tests for various I/M data sets from the CARB pilot project data are shown in Table B-1.

The result of the χ^2 test can be expressed as the probability that the calculated value of C (or a higher value) would be observed if the experimental data in the sample came from the hypothetical distribution. This probability, called the p value, is found from the

---

*The introductory material may be found in many standard statistical tests. The reference used here was Richard J. Larsen and Morris L. Marx, *An Introduction to Mathematical Statistics and Its Applications* (second edition), Prentice-Hall, 1986, Chapter 9.

**To apply the χ^2 test, the minimum value of np_i should be greater than or equal to five. In all the comparisons made here, the minimum value of np_i was just above 30.

cumulative χ^2 distribution, $F(\chi^2,df)$ as follows: p value = 1 - F(C,df).*

Table B-1
χ^2 Tests for Selected Data Sets from CARB Enhanced I/M Pilot Program

Test	Species	Distribution	Test Statistic, C	p Value(See Note)
IM240	HC	Normal	2,753.05	-
		Lognormal	22.13	0.1797
		Gamma	155.62	1.9×10^{-24}
	CO	Normal	1363.29	8.2×10^{-277}
		Lognormal	38.03	0.0024
		Gamma	129.26	2.9×10^{-19}
	NOx	Normal	398.09	4.6×10^{-74}
		Lognormal	24.71	0.1013
		Gamma	27.18	0.0555
ASM2525	HC	Normal	2,413.70	-
		Lognormal	41.79	0.0007
		Gamma	217.76	7.4×10^{-37}

Note: A dash (-) in the p-Value column indicates a numerical underflow in the calculation via the Excel function, chiinv.

For the four cases examined here the lognormal distribution provides the highest p values showing the best agreement between the experimental data and the postulated distribution. The p values are not high enough to provide confidence that the lognormal

*Calculating the p value is an alternative to hypothesis testing where a significance level is picked in advance. In a hypothesis test the null hypothesis is that the data do come from the postulated distribution. This null hypothesis is rejected if the value of the test statistic C exceeds the ordinate of the χ^2 distribution for the selected significance level (and the degrees of freedom for the data set.) For seventeen degrees of freedom (df = 17) used here the critical values of the χ^2 distribution at various levels of significance are shown below.

Significance Level, α	0.2	0.1	0.05	0.01	0.005	0.001
Confidence Level, 1 - α	80.0%	90.0%	95.0%	99.0%	99.5%	99.9%
χ^2 Ordinate for df = 17	21.615	24.769	27.587	33.409	35.718	40.791

distribution gives a correct representation of the data in all cases. The poorest agreement is for the case of the hydrocarbon data from the ASM2525 test. In this case the p value is only 0.0007; in a hypothesis test the null hypothesis that the data fit a lognormal distribution would be rejected at a 99.9% confidence level. However, even in this case the lognormal distribution still provides a much better representation of the data than the other distributions. As expected, the symmetrical normal distribution always provides the worst representation of the data. The gamma distribution is qualitatively similar to the lognormal distribution in that it has a long tail. However, this distribution provides a poor fit in the three of the four cases shown here. Even in the one case where it provides a reasonable fit – the IM240 NOx data – it still does not perform as well as the lognormal distribution.