Assessment of the Report "Development of a Proposed Procedure for Determining the Equivalency of Alternative Inspection and Maintenance Programs."

This report has been prepared under PO # 8A-0323-NA9X for The United States Environmental Protection Agency-National Vehicle and Fuel Emissions Laboratory.

Edward D. Rothman
Professor of Statistics,
Director of the Center for Statistical Consultation and Research
University of Michigan,
Ann Arbor, Michigan 48109

Assessment of the Report "Development of a Proposed Procedure for Determining the Equivalency of Alternative Inspection and Maintenance Programs."

Introduction

The purpose of this study is to provide a review of the statistical studies performed by Sierra Research Inc. presented in a November 10, 1997 report prepared for the U.S. Environmental Protection Agency Regional and State Programs Division under contract No. 68-C4-0056, Work Assignment No. 2-03. This review focuses on the statistical aspects of their work and on the rationale for the recommendations that they reach.

My comments are based primarily on this Sierra Research report and data used to compute the correlation between various measures of emissions. I also examined several other reports focused on remote sensing. These included a set of slides titled "Remote Sensing Briefing" by Joel Schwartz, July 14, 1998 and a second document titled Agenda Item #111: Remote Sensing, July 7, 1998. There were also reports prepared by Tom Wenzel, Lawrence Berkeley National Laboratory which were read.

The primary question I asked was whether the procedure proposed by Sierra Research allows us to determine equivalency of alternative inspection and maintenance procedures. The normalization or standardization of the measurement system seems to be well supported by their report. Their analysis of sample sizes based on a log- normal distribution also seems reasonable. So the answer to this first question depends primarily on potential biases that may affect measurements differently in different I/M programs.

Furthermore, Sierra Research claims that estimates of the level of emissions of various types, resulting from an I/M program, are possible. Though the potential biases that could effect such an extrapolation may not exist, Sierra has not demonstrated that these biases are absent. Extrapolation of measures of emissions obtained in the IM240 program, or through an alternative procedure, to emissions of vehicles in actual use is at the heart of this concern.

My overall recommendation is that the proposed program should be an integral part of effectiveness of I/M programs. However, the scale of the measurement process should be changed, and I believe that a complementary measurement process based on remote sensing of randomly selected sampling sites should be developed and implemented.

This report details these findings and recommendations in the next section and presents the rationales below each point.

Summary of Findings

Listed below in summary form are the key points regarding my assessment of the Sierra report. Below each point I provide comments designed to clarify the heading.

1. The use of the IM240 system of measurement as a measurement standard seems appropriate.

The measurement systems range from a Federal Test Procedure (FTP), with a cost estimated at $1000 per vehicle that requires two days of measurement under carefully constrained conditions, to other lower-cost but potentially less accurate alternatives. Though there may be justifiable reasons provided for any of these alternatives, it is important that a reasonable benchmark be established. Though the quality of the agreement indicated by Sierra Research between the IM240 emissions readings and the corresponding FTP readings is likely higher than should be anticipated, the IM240 seems best for this purpose.

A summary table describing their findings is presented below:

TABLE 1

Test Type	Conditions	Results
FTP	*Cold-start *One-hour evaporative emissions test *Measures CO, NOX,HC *$1000 per vehicle and requires two days	Standard
IM240	4 minute subset of FTP *No cold start or warmup *$25 per vehicle but requires expensive equipment	HC: rsq=.89 NOx: rsq.=.78 CO: rsq=.66
ASM	*Dynamometer tests to load Engine and simulate acceleration *No cold-start or warm-up *$20-$40 per vehicle	HC: rsq=.77 NOx: rsq=.53 CO: rsq=.72
Idle	*No load test *Can't measure NOx emissions *NOx emissions are significant *Easiest test to falsify results *Inexpensive	HC: rsq=.64 CO: rsq=.26 NOx: rsq= N/A
2500 rpm	*No load *Usually paired with idle test *Inexpensive	HC: rsq=.59 CO: rsq=.66 NOx: rsq= N/A

I examined the correlations described by Sierra Research are affected to a great extent by just a few observations. Correlation reflects the extent that two vectors of data are linearly related. If indeed a linear relationship exists, the strength of the relationship should remain constant over the entire range of data. To this end, I examined the impact of looking at the smallest 90% of the observations.

The following table provides contrasts between these correlations, using the entire data set and with the top 10% of the data removed from the analysis.

Log-transformed plots of the ASM2525 test for hydrocarbons are omitted because it assigns negative values to some vehicles.

Correlation coefficients between FTP and IM240 on the original scale	
Carbon Monoxide	r = +0.817 (p < 0.0005)
Hydrocarbons	r = +0.942 (p < 0.0005)
Nitrous Oxides	r = +0.880 (p < 0.0005)
Correlation coefficients between FFP and IM240 on the log scale	
Carbon Monoxide	r = +0.854 (p < 0.0005)
Hydrocarbons	r = +0.852 (p < 0.0005)
Nitrous Oxides	r = +0.905 (p < 0.0005)
Omitting top 10% of FIP values, correlation coefficients between FTP and IM240 on the original scale	
Carbon Monoxide	r = +0.781 (p < 0.0005)
Hydrocarbons	r = +0.753 (p < 0.0005)
Nitrous Oxides	r = +0.869 (p < 0.0005)
Omitting top 10% of FTP values, correlation coefficients between FTP and IM240 on the log scale	
Carbon Monoxide	r = +0.723 (p < 0.0005)
Hydrocarbons	r = +0.803 (p < 0.0005)
Nitrous Oxides	r = +0.891 (p < 0.0005)

2. The IM240 measurements should be re-scaled by taking the logarithm for each of the various emissions recorded.

Scales for measurements are selected for a variety of reasons. These include constancy of variance, and normality of the distribution. In both cases, the issue is simplicity of the model or frame of reference. Constancy of variance allows us to attach the same level of accuracy to a low emissions measure as we would to a high emissions measure. The use of a normal distribution provides a common useful frame from which we can evaluate what is within the system and what falls outside. In repeated measurements, under approximately the same circumstances, we would want to know whether a measurement was generated by a special mechanism. With a normal distribution, over 99% of the recorded observations should fall within three standard deviations of the average. Ninety-five percent are within two standard deviations of the mean. And additional results can be obtained for any range of values.

The distribution of emissions values for each of the three variables suggest a skewed distribution. And though there are many distributions-skewed to the right-not all provide an adequate fit to the observations. We find, in agreement with Sierra Research, that the log-normal distribution provides an adequate fit to the data.

The log-normal distribution is found to be an appropriate model when the logarithm of a measurement, rather than the original measurement has a normal distribution. Sierra Research studied the usefulness of the log-normal distribution in the appendix of their report, but use chi-square measures rather than a graphical display to support their findings.

Even when a model provides the appropriate approximation to the proportions of cases we should expect in a given interval, the actual number of cases observed could become quite different as the sample size is increased. With sample sizes as large as are used, the chi-square measure almost always leads to rejection of the model. I prefer, as common statistical practice, to look at quantile plots.

The data provided by IM240 measurement system were obtained from vehicles of various ages. Even when we adjust for the different ages of these vehicles, it appears that the measurements are a mixture of at least two processes. The right tail of the distribution appears to have been generated by a separate process. When these few observations are eliminated, however, the remaining observations appear to have a log-normal distribution.

To illustrate this fact, we can plot the ordered values for a particular emissions level against the quantiles we would expect for a normal distribution (Q-Q plot). An additional Q-Q plot, obtained by first taking the logarithm of the emissions values and then proceeding as before, supports the log-normal model.

The charts below provide some support for this recommendation. There are additional studies of age adjusted data which could be added. However, I would want to understand the nature of the admixture indicated in histograms of the emissions readings, and the nature of the sampling process before this recommendation would be final.

Figure 1: Normal Q-Q Plot of Log FTP CO (All Data)

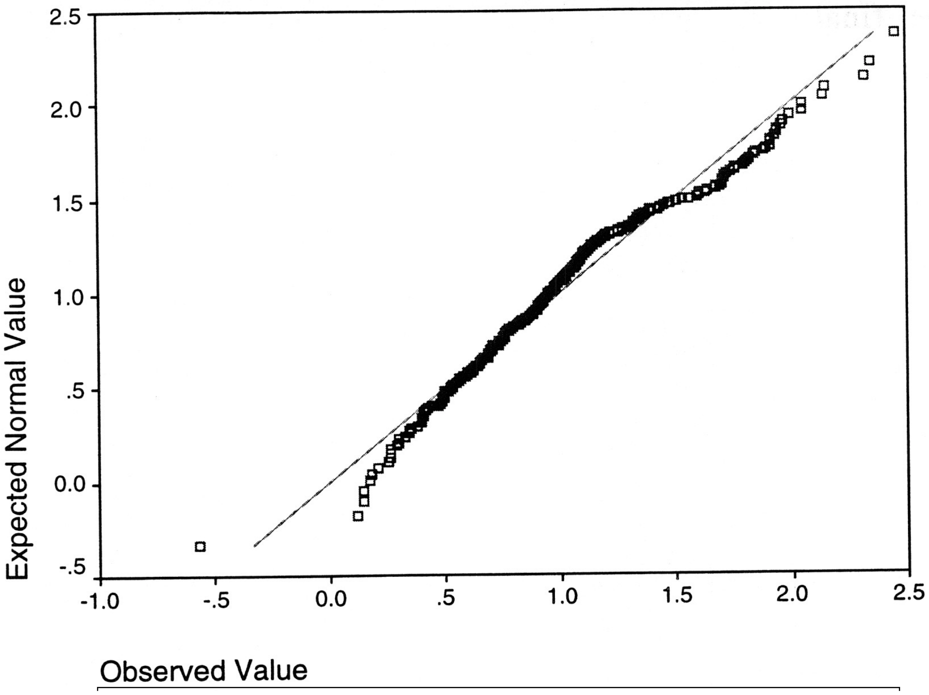


Figure 2: Normal Q-Q Plot of Log FTP CO (Excluding Top 10% of Data)

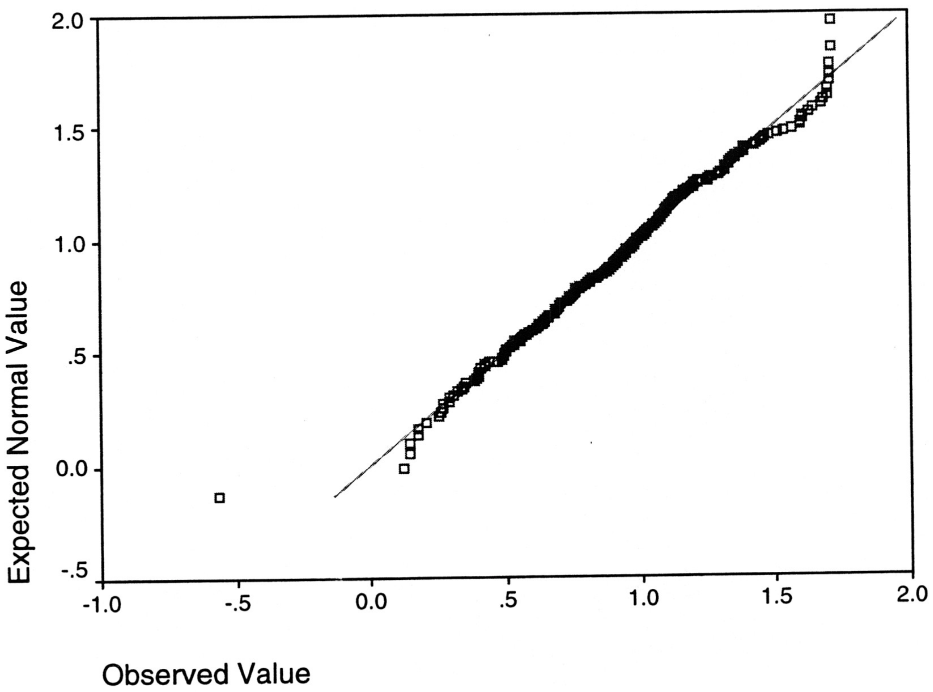


Figure 3: Normal Q-Q Plot of Log FTP HC (All Data)

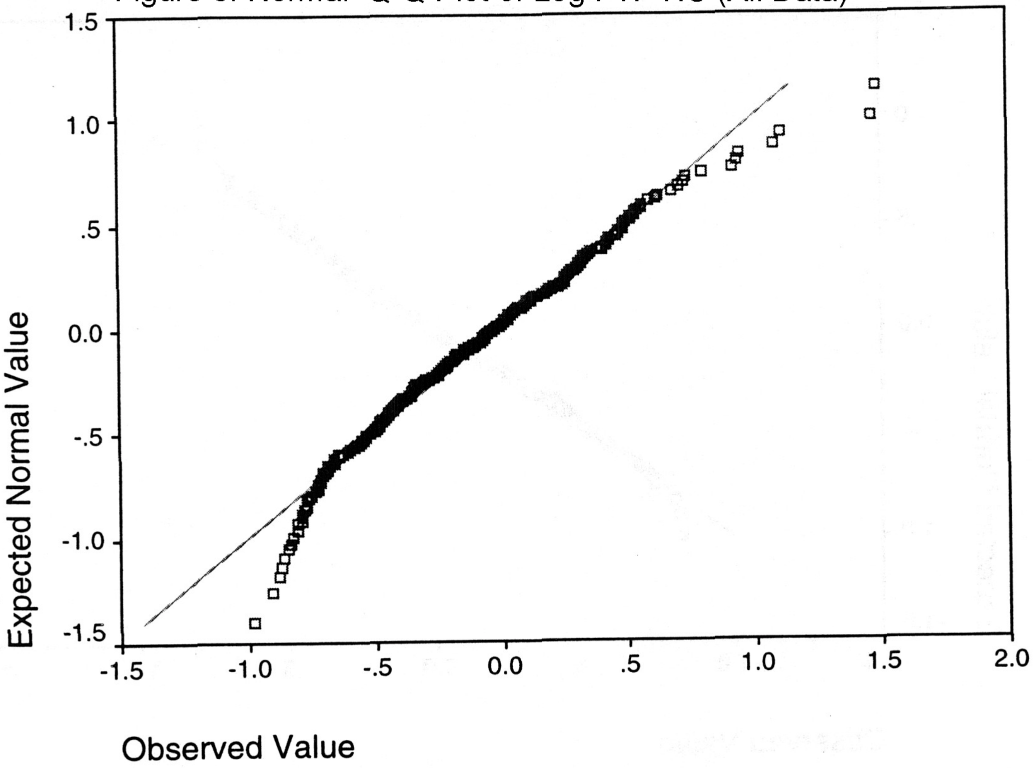


Figure 4; Normal Q-Q Plot of Log FTP HC (Excluding Top 10% of Data)

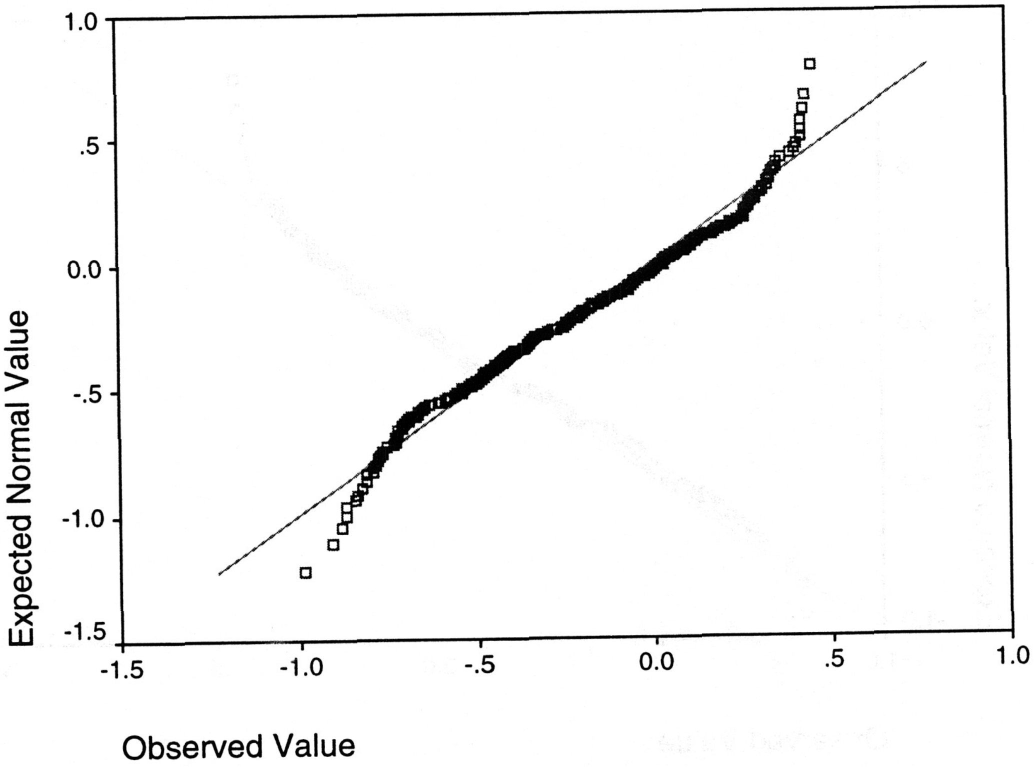


Figure 5: Normal Q-Q Plot of Log FTP NOx (All Data)

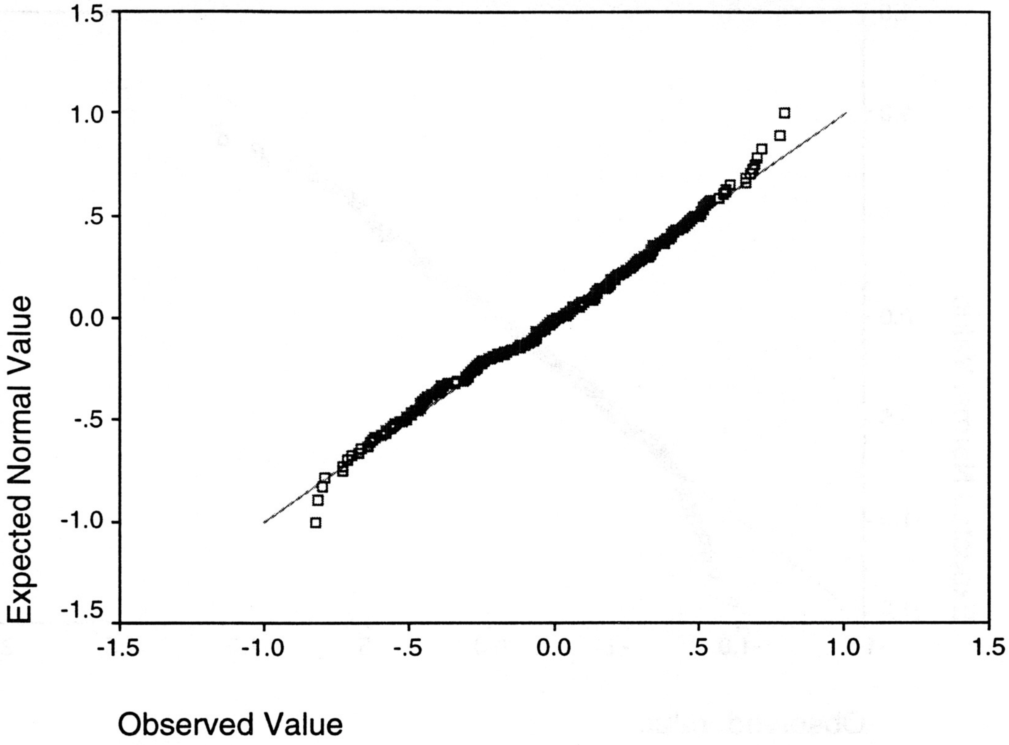


Figure 6: Normal Q-Q Plot of Log FTP NOx (Excluding Top 10% of Data)

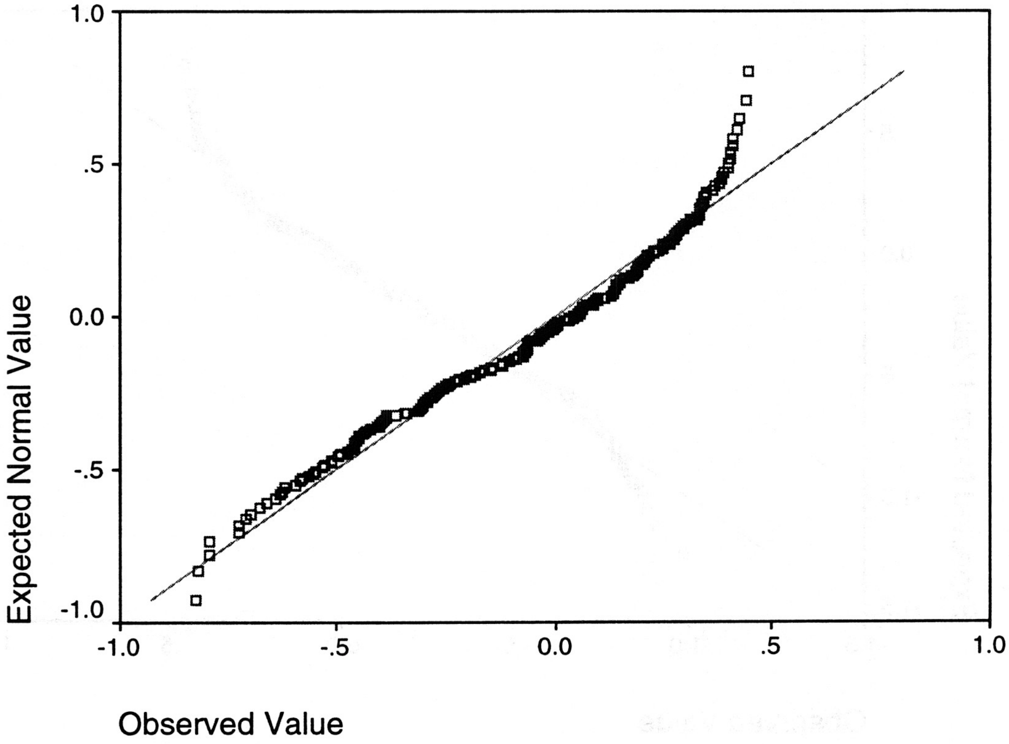


Figure 7: Normal Q-Q Plot Untransformed FTP CO

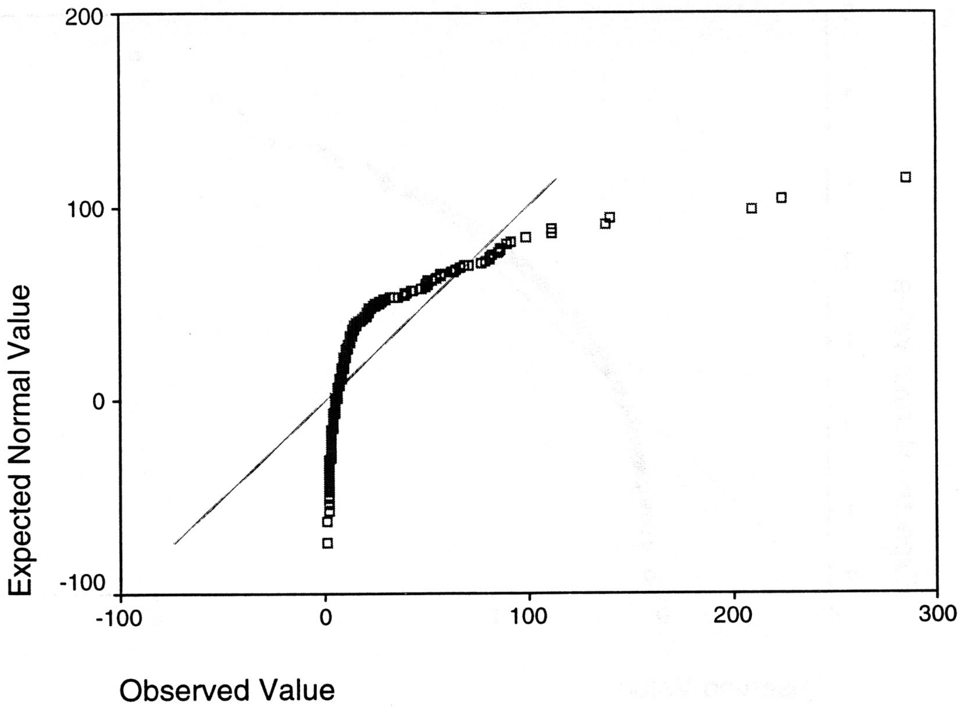


Figure 8: Normal Q-Q Plot Untransformed FTP HC

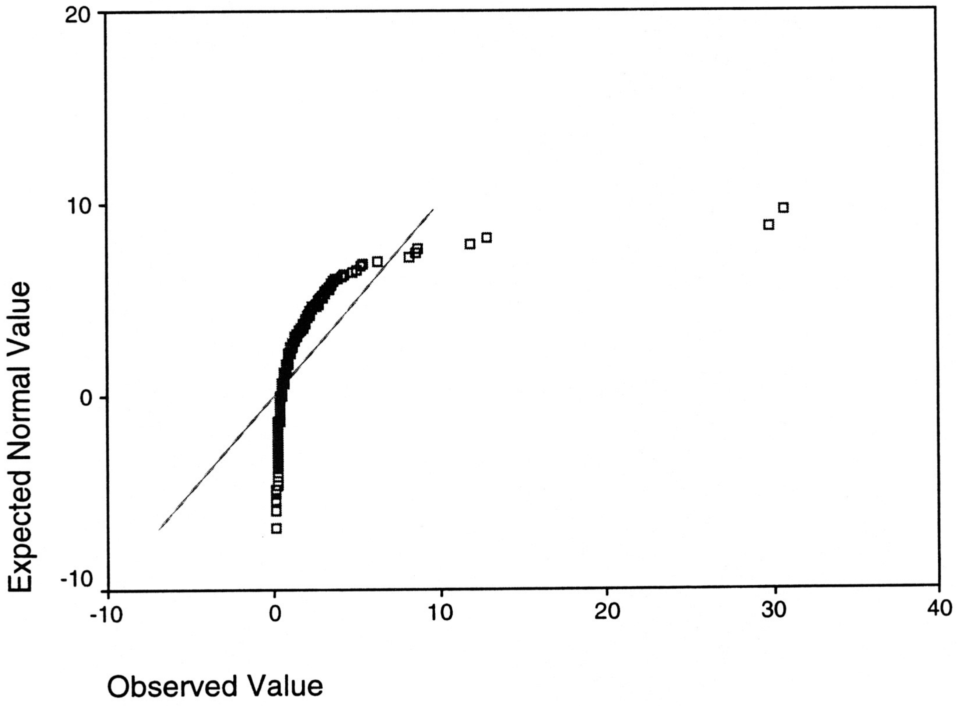


Figure 9: Normal Q-Q Plot Untransformed FTP NOx

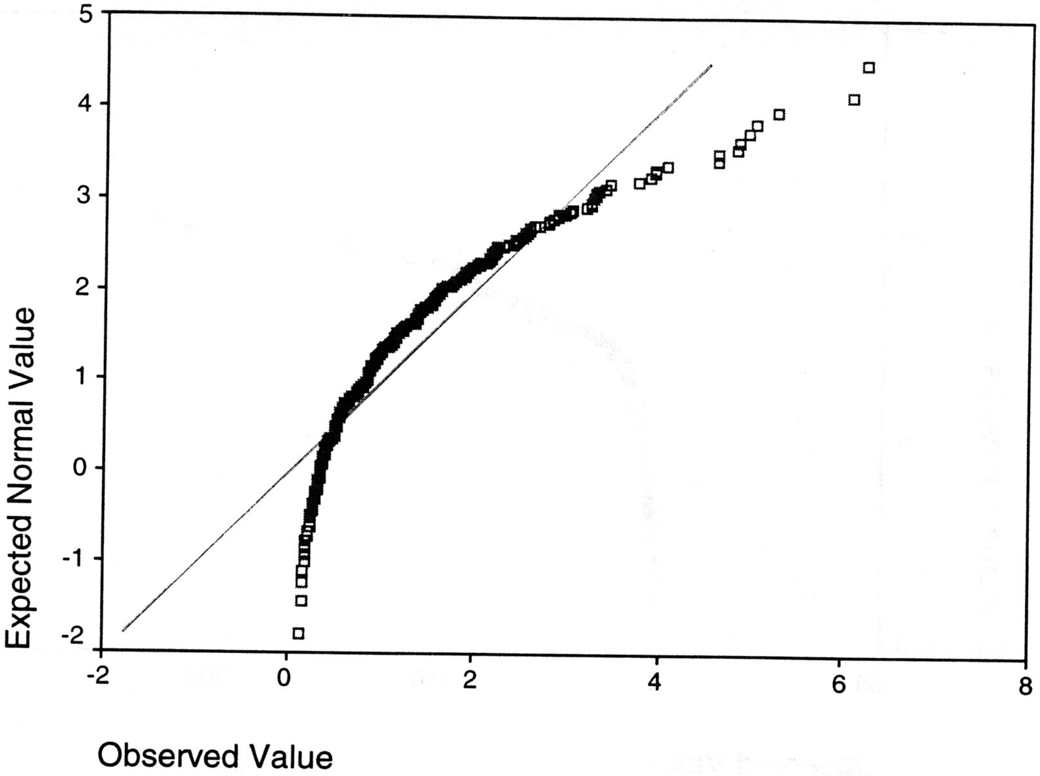


Figure 12: Normal Q-Q Plot of Log IM240 HC
(All Data)

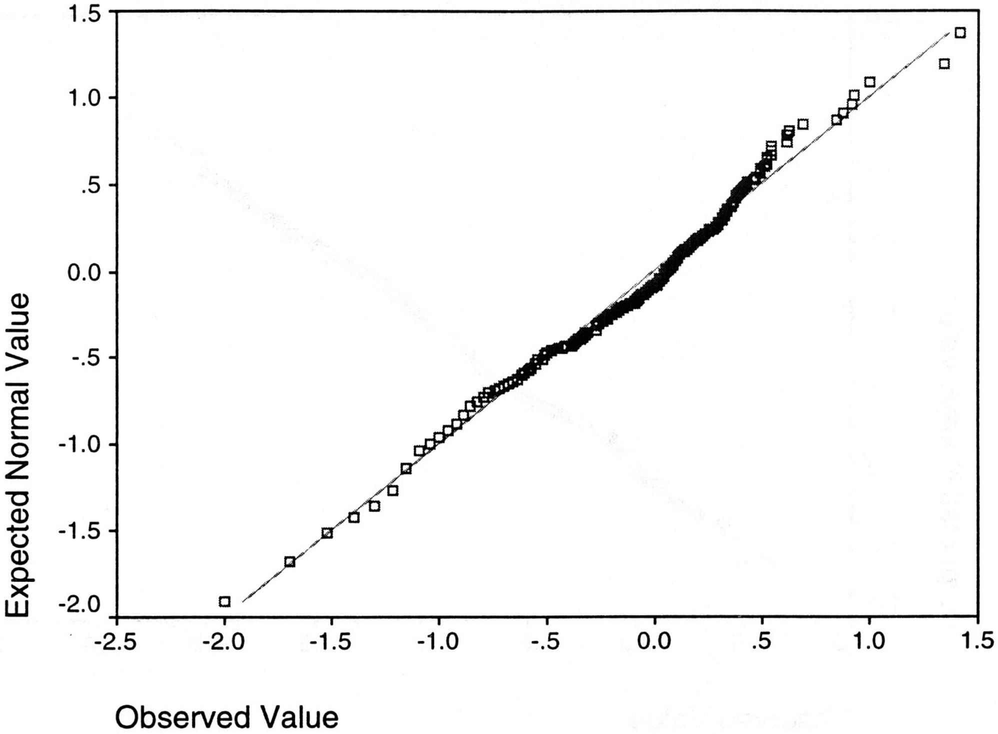


Figure 13: Normal Q-Q Plot of Log IM240 HC
(Excluding Top 10% of Data)

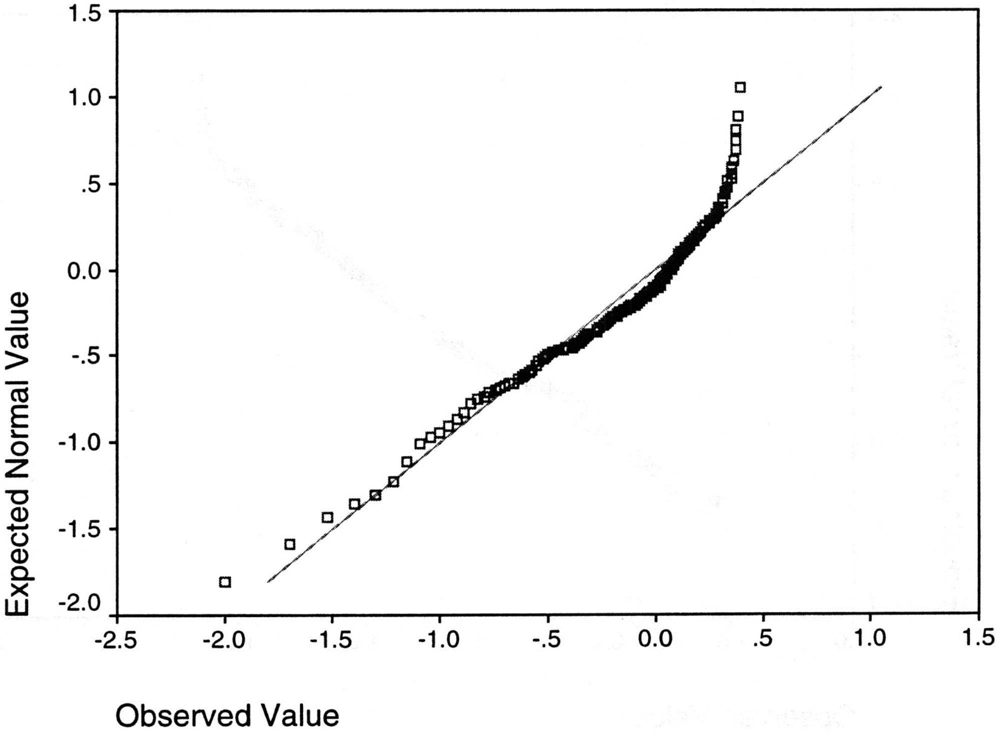


Figure 14: Normal Q-Q Plot of Log IM240 NOx
(All Data)

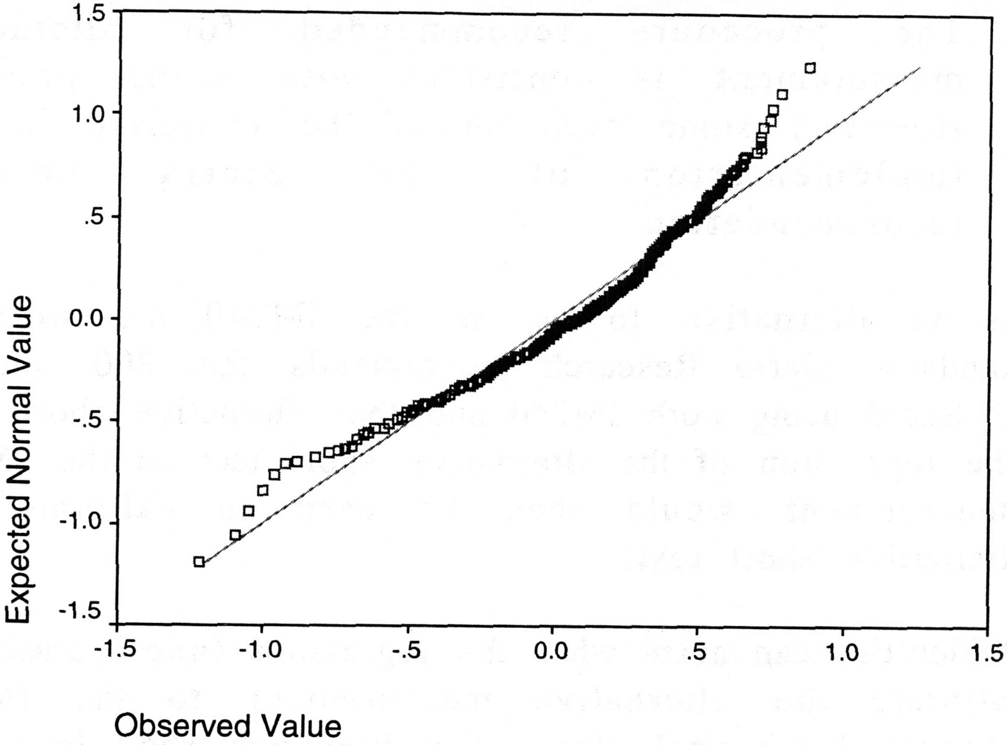
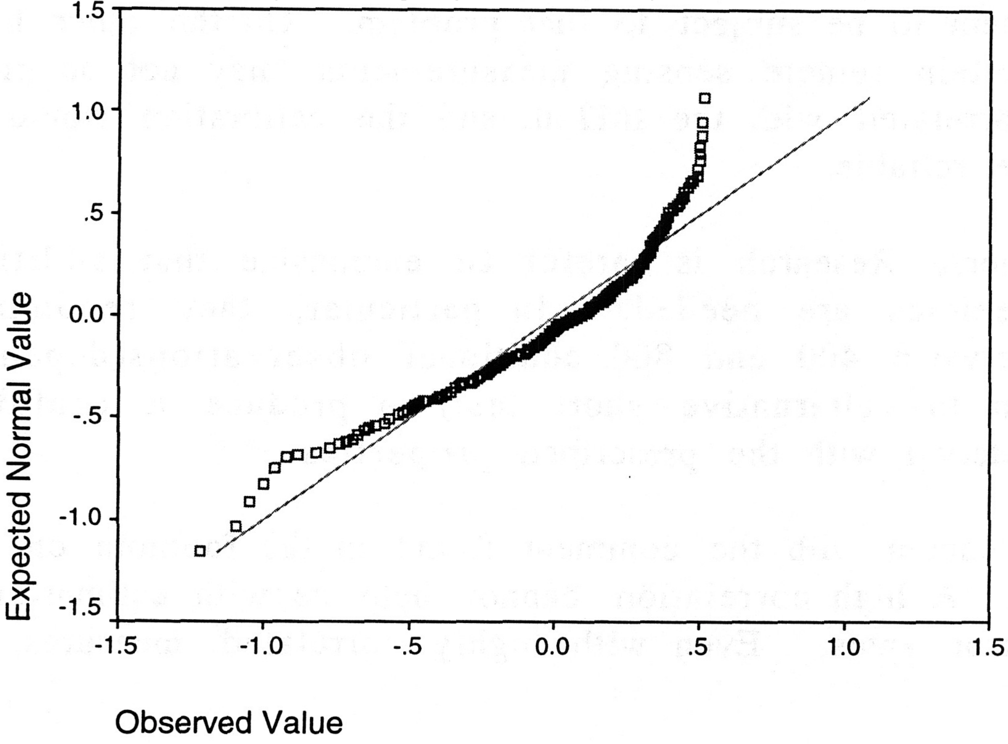


Figure 15: Normal Q-Q Plot of Log IM240 NOx
(Excluding Top 10% of Data)



3. The procedure recommended for alternative measurement is consistent with sound practice. However, some care should be exercised in the implementation of the Sierra Research recommendation.

As an alternative to use of the IM240 measurement standard, Sierra Research recommends that 800 vehicles be tested using both IM240 and the alternative short test. The regression of the alternative short test on the IM240 measurement would then be used to calibrate the alternative short test.

Difficulties can arise when the regression function used to calibrate the alternative measurement to the IM240 standard has a small slope. For alternative short tests that exhibit such a pattern, this process may lead to rather poor estimates. The short tests described here do not seem to be subject to that problem. On the other hand, certain remote sensing measurements may not be highly correlated with the IM240, and the calibration would not be reliable.

Sierra Research is careful to emphasize that additional vehicles are needed. In particular, they recommend between 400 and 800 additional observations(depending on the alternative short test) to produce a confidence interval with the prescribed properties.

I concur with the comment found in the footnote on page 7. A high correlation cannot help us with estimation of error rates. Even with highly correlated measures, the proportion of false positives and false negatives can be quite different.

4. Sierra Research does not describe how the relationship between the various measurements can and should be used. In particular, there is a strong correlation between HC and CO measurements.

An examination of the data on emissions obtained on the same vehicles indicates a high correlation between the HC readings and CO. Since the responses are related to each other, an opportunity to use this connection can lead to better understanding of the measurements. In particular, if an I/M program was found to have improved in reducing emissions by certain levels for HC and for CO, then other similar programs should exhibit a similar improvement in each and the relative improvement should also be similar.

I recommend that the relative improvement be measured. In particular, the relative magnitude of changes in hydrocarbons, the oxides of nitrogen, carbon monoxides, and gas cap emissions should be compared. I anticipate that the relative level of any one feature should be related to the magnitude of another emissions feature. Substantial differences in the relationship could be cause for investigation.

Though I do not have a specific recommendation as to how to proceed with individual vehicles, there is an opportunity here too. The idea is that higher levels of CO should generally be associated with higher levels of HC. If we find unusually high readings of one and not the other, then an additional investigation might be reasonable.

In general, I would study the connections between these measures of emissions. These relationships may not be as strong as found between measurements of the same quantity, but they may be useful. These relationships, if they exist in the population of vehicles of interest, can be used in a variety of contexts.

For example, consider a non-centralized program. Data are recorded for each vehicle, but the relationship between the different measures is quite similar in most test facilities but quite different in a few others. This could suggest the data have been "cooked."

5. The environmental impact of a program of inspection and maintenance depends on factors that are not reflected in the measurement system.

The focus of the Sierra Research report examines the relative effectiveness of an I/M program as found under special circumstances of the IM240 short test. However, if we want to know whether the observed improvement found under these circumstances impacts air quality, then there are some concerns.

The issue is bias. In both options, we are to begin with recruitment of a random sample of vehicles. This selection is described on pages 22 and 23 of the Sierra Research report. Their recommendations, though reasonable, could still be subject to a recruitment bias. And though this bias is perhaps more likely to arise in a decentralized program, it could affect the centralized program too. Indeed, if this were not a concern, then there would not be a need for a covert audit of vehicles.

Vehicles included in the study could, as a result of prior notification, have changes made to them.

Facilities could add measurements to a compilation to meet the needs of the region.

6. Measurements over time are essential to assess the impact of a program.

Time is an element of every system. What we seek are I/M programs that have an effect on a vehicle fleet and for which this effect will endure. Unless the results of the program of evaluation are provided at regular intervals of time, this objective cannot be studied.

As Sierra Research points out, we do not have a record of the pre-I/M program. However, the pattern of responses over time can lead us to better understand whether we have reached a stable point or the process continues to evolve.

A stable process is a process whose mechanism does not change in a systematic way. The output of such a process can be viewed as a sequence of random observations from a single collection. The work of W. Edwards Deming, "The New Economics," for instance, may provide a useful reference.

When the emissions reported over time by a region exhibit a random pattern, it would suggest that the program will not yield further improvements. I would also recommend, in this context, that the relationship between the various measures of emissions also be studied. A change in the relationship could signal a change in mechanism too.

7. I recommend that the compliance rate (e.g. the size of the vehicle fleet that has completed I/M program requirements relative to the overall size of the vehicle fleet) should be provided.

The effectiveness of an I/M program depends on enrollment as well as on the effectiveness of the program for enrolled vehicles. Inferences to an entire program based only on vehicles that have passed an inspection is problematic. The compliance rate in the population of vehicles is not part of the assessment proposed by Sierra Research.

8. The lack of agreement between IM240 measurements and remote sensing recorders should be investigated.

The general lack of agreement between measurements may, as indicated by Sierra Research, reflect the fact that the measurements are obtained under different circumstances. I don't doubt that this is a primary cause for the lack of agreement. The other side of the lack of agreement could reflect that the laboratory measurements' of an IM240 are unrelated to any real world emissions. It is also possible that there may be serious measurement problems with the remote sensing record.

Before we reflect on the potential bias, there may also be differences due to measurement error. These are errors that would move a measurement in either direction (e.g. above or below a 'true value'). Measurement error is likely to be greater when measurements are based on remote sensing.

Such errors may be important for individual measurements, but they are not expected to be important when averages of large numbers of readings are available. Since the cost of remote sensing measurements may not be too large, a substantial number of readings can be taken to reduce the standard error of a mean. The reduction in standard error of an estimate is inversely proportional to the square root of the sample size.

9. The focus of the Sierra Research recommended program should be included as a common objective of all studies.

I have indicated above that inference from the special circumstances proposed by Sierra Research to any real world situation may be subject to bias. On the other hand, the opportunity to observe a factor with little confounding from factors over which the region may have little control can be valuable.

There are still potential biases associated with the Sierra Research program not associated with extrapolation to effectiveness of the program at a single point in time. These biases could be due to the measurement system, the sample of vehicles selected for test, and possibly others. However, it may be reasonable to expect that these biases will continue from year to year. Changes in the measurement system could then reflect the incremental gain over past programs.

It must be noted that such incremental changes may be very small in programs that have had success in the past. The magnitude of the improvement would need to be gauged relative to past measures of performance.

In addition, even if we observed the remote sensing readings without variation in every situation, the contribution to this measurement from a variety of factors would be unknown. Drivers' behavior and other potential contributors to the air, for instance, can lead to changes in records that may not have anything to do with the I/M program.

10. I recommend that a direct measurement of emissions by random sampling of observed emissions for vehicles should accompany the process recommended by Sierra Research.

This recommendation is analogous to the use of a stress test for heart disease. The standard expensive and invasive look at the arteries feeding the heart does an effective job, I'm told, in the identification of blockage. However, unless the physician knows the potential impact of reduced flow, the importance to the individual may be unknown. Without a measurement system that focuses on impact, the improvements observed may not have a realizable effect.

To conduct such an evaluation, the sampling process would focus on use conditions. Sampling highways, including all lanes, exits and entries, and roads both major and minor would need to be included. And though the process may seem to represent too large a task, without such records it is difficult to know whether an improvement has affected the air quality.

The sampling process for vehicles would not necessarily represent a random sample of vehicles. When a vehicle is used more, it would contribute more to the total emissions. And if points on the highway are identified as acceleration points, then the recorded emissions may also contribute a different amount to the total.

11. The modeling provided by MOBILE needs evaluation.

There is little evidence provided that the extrapolation from a pressure emissions test to a contribution to the atmosphere is reasonable. The evaluation suggested by Sierra Research may well be adequate, but I have little information on which to base such an opinion.

12. Sample sizes are adequate.

A section on sample size needs some rewriting. For example, a formula is presented for the sample size as a function of a sample mean and a sample standard deviation. Such quantities are unknown prior to sampling!

A similar comment applies to the confidence interval section of their report. Though the confidence limits for the mean of a normal distribution are as presented in A-1, the formula suggests that the population mean is found by solving an equation. This isn't true. The formula gives the lower and upper limits of a confidence interval. If we use the process again and again, we expect the limits will contain the true mean in 1-alpha cases.

Similar comments apply to the subsequent formulas in Appendix A of the Sierra report.