NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

NCHRP

Web-Only Document 217:

Precision and Bias Statements for AASHTO Standard Methods of Test TP 98 and TP 99

Elliott Dick
Tim Casey
HDR Engineering, Inc.
Minneapolis, MN

May Raad
HDR Corporation
Ottawa, ON

Christopher Nachtsheim
University of Minnesota, Twin Cities
Minneapolis, MN

Contractor's Final Report for NCHRP Project 10-88
Submitted October 2015

TRB TRANSPORTATION RESEARCH BOARD
The National Academies of
SCIENCES · ENGINEERING · MEDICINE

ACKNOWLEDGMENT

This work was sponsored by the American Association of State Highway and Transportation Officials (AASHTO), in cooperation with the Federal Highway Administration, and was conducted in the National Cooperative Highway Research Program (NCHRP), which is administered by the Transportation Research Board (TRB) of the National Academies of Sciences, Engineering, and Medicine.

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, non-governmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Ralph J. Cicerone is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.national-academies.org**.

The **Transportation Research Board** is one of seven major programs of the National Academies of Sciences, Engineering, and Medicine. The mission of the Transportation Research Board is to increase the benefits that transportation contributes to society by providing leadership in transportation innovation and progress through research and information exchange, conducted within a setting that is objective, interdisciplinary, and multimodal. The Board's varied activities annually engage about 7,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia, all of whom contribute their expertise in the public interest. The program is supported by state transportation departments, federal agencies including the component administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation.

Learn more about the Transportation Research Board at **www.TRB.org**.

CONTENTS

LIST OF FIGURES AND TABLES

List of Figures

List of Tables

AUTHOR ACKNOWLEDGMENTS

The research reported herein was performed under NCHRP Project 10-88 by HDR Engineering, Inc. (HDR).

Elliott Dick, Acoustical Scientist at HDR, was the principal author of this report and technical lead for the project (approach, measurement planning and execution, data management and processing, and reporting activities). May Raad, Principal Statistician at HDR, led the statistical approach and the statistical analyses, and co-authored this report. Dr. Christopher Nachtsheim, Professor at the Carlson School of Management, University of Minnesota, Twin Cities, provided invaluable guidance in the statistical approach and analysis. Gina Jarta, Acoustical Specialist at HDR, performed the literature review. Tim Casey, Acoustic Program Manager at HDR, was the principal investigator (HDR Project Manager).

ABSTRACT

This report documents and presents results of a study of precision and bias associated with two AASHTO standards used to measure tire-pavement noise. Those standards are:

- TP 98, Standard Method of Test for Determining the Influence of Road Surfaces on Vehicle Noise Using the Statistical Isolated Pass-By Method (SIP), and
- TP 99, Standard Method of Test for Determining the Influence of Road Surfaces on Traffic Noise Using the Continuous-Flow Traffic Time-Integrated method (CTIM).

This study incorporated innovative techniques to perform the measurements. The research team digitally recorded audio and video of traffic on segments of highways that were flat, had certain mixes of cars and heavy trucks, and met other criteria explained later in this report. Test operators on the research team were able to perform measurements in accordance with TP 98 and TP 99 in a controlled environment remote from the roadways in order to accomplish a Repeatability and Reproducibility study. Robust statistical analyses were performed on the test operator results. Statistical analysis results suggest that: the TP 98 measurement method for automobiles and heavy trucks is not precise, and the TP 99 measurement method has excellent precision. The modeling and normalization along with the outlier elimination process are likely contributors to the excellent precision of the TP 99 measurement process.

SUMMARY

Developing Precision and Bias Statements for AASHTO Standard Methods of Test TP 98 and TP 99

This analysis of test method TP 98 for both vehicle classes of automobiles and heavy trucks determined that this method is not precise since (1) the percent Repeatability & Reproducibility (R&R) value is greater than 30 percent and (2) the upper bound of the signal-to-noise ratio (SNR) confidence interval is less than 3. This finding suggests that the measurement process needs reexamination to increase the precision of the process.

With respect to test method TP 99, the R&R analysis did find that this test method is precise since (1) the R&R value is substantially smaller than 30 percent and (2) the lower bound of the SNR's value is substantially greater than 3. This method is precise and researchers can have confidence in the precision of their measurements provided they follow the protocols for test method TP 99.

The present assessment of the component of bias related to the acoustic environment in which sound recordings are conducted according to test methods TP 98 and TP 99 showed that the amount of bias is negligible.

According to the scope and purpose of TP 98, it is intended to be broadly applied and used to compare one pavement to another, either at different roadways or at different times in the lifespan of one roadway. However, this research indicates the precision of TP 98 doesn't necessarily allow pavements to be compared, unless the difference is substantial. There may be ways to improve the precision of the measurement process, including, most likely, larger sample sizes and perhaps multiple sampling periods in different conditions.

In contrast, the scope and purpose of TP 99 is intended to be narrowly applied to a single measurement site and compare different pavements or different times in the lifespan of one pavement. The precision of TP 99 is excellent, and users of this measurement process can be confident that differences in measurements reflect differences in sound levels. Variations in traffic, which could affect sound levels independently of effects from tire-pavement noise, are implicit in the measurement process. Speculating on all the components of this measurement procedure, the modeling and normalization along with the outlier elimination process are likely contributors to the excellent precision of the TP 98 measurement process. Identifying these particular factors in the measurement process is outside the scope of the research.

Results of this study determined that for both TP 98 and TP 99, both methods technically have no bias because the test result is defined only in terms of these test methods. Nonetheless, an evaluation of a component of bias found the 95 percent confidence interval for the true bias contains zero. This was determined using a Reference Sound Source (RSS) on the road which evaluated bias due to

environmental factors which could affect the propagation of sound along the ray from the roadway to the microphone. The interval can range from -0.194 dBA to 1.440 dBA at the 95 percent level of confidence. Overall, the impact on the accuracy related to this component of bias is negligible. However a ruggedness study showed that air temperature was a marginally statistically significant factor. Therefore, in lieu of further research, any comparisons between test results should also compare the temperatures during each of the tests.

CHAPTER 1

Introduction

Background

Some of the most pervasive environmental noise sources come from transportation systems. Highway traffic noise is a dominant noise source for both urban and rural environments (FHWA 2013). In the past land use planning, speed restrictions, vehicle prohibitions (including prohibitions on engine compression brakes), the use of quieter vehicles, noise walls, and in very limited instances receiver-based modification have been the primary methods of mitigating roadway traffic noise. Research in the U.S. and abroad has shown that tire-pavement noise can be a large contributor to total vehicular noise. As such, tire pavement noise has become an increasingly important consideration in the mitigation of traffic noise.

Currently noise walls are the most commonly used form of traffic noise abatement in the U.S. There are many instances in which noise barriers are not feasible from an engineering perspective, reasonable from a cost-effectiveness perspective, or accepted by community members. Additionally the acoustical benefits of a noise wall noise apply only to land uses located in the acoustical shadow zone behind the wall. In contrast, quieter pavements have the potential to help reduce highway traffic noise by reducing the noise generated at the tire/pavement interface and through absorption of propagating sound (Rochat et al. 2010). As more information becomes available concerning the benefit of quiet pavements and the acoustic longevity of such technologies, several states have shown interest in using quieter pavements as a form of traffic noise mitigation.

In the U.S. many states have implemented tire-pavement noise research programs to study if particular roadway surface mixtures or textures can be used as effective traffic noise mitigation measures (CDOT 2012; Donovan 2012; Herman et al. 2005; Harvey et al. 2011). Additionally, ongoing research focuses on the acoustic longevity of roadway surface treatments and the implementation of tire-pavement effects into traffic prediction models. The FHWA conducted two tire-pavement noise strategic planning workshops in 2004 and 2006, which led to the creation of an Expert Task Group (ETG) on tire-pavement noise. Uniform tire-pavement noise measurement methods were developed to facilitate a comparison of tire-pavement noise throughout the U.S. (Rochat 2009; Rochat et al. 2012). These methods determine the influence of roadway surface on traffic and vehicular noise at various distances alongside a roadway. The American Association of State Highway and Transportation Officials (AASHTO) published the standardized measurement methods:

- TP 98, Standard Method of Test for Determining the Influence of Road Surfaces on Vehicle Noise Using the Statistical Isolated Pass-By Method (SIP), and
- TP 99, Standard Method of Test for Determining the Influence of Road Surfaces on Traffic Noise Using the Continuous-Flow Traffic Time-Integrated method (CTIM).

Both of these standards were revised twice during the course of this research; the work described herein conformed to the most recent revision in 2013 (TP 98-13 and TP 99-13).

Research Objective

The objective of this research is to develop precision and bias statements for AASHTO Standard Methods of Test TP 98 and TP 99. Adoption of the precision and bias statements is governed by AASHTO bylaws and the subcommittees responsible for development and revision of the TP 98 and TP 99 standards documents. The goal of this research is to provide precision and bias statements based on a well-prepared and well-executed experimental design to facilitate review, acceptance, and implementation by AASHTO.

Organization of Report

This report summarizes the approach and the findings of this research project. This chapter discusses the background and research objective. Chapter 2 summarizes high-level findings from the literature review. Chapter 3 presents the approach and experiment design. Chapter 4 discusses the data gathering and measurements. Chapter 5 presents the statistical analysis towards the precision and bias of the measurements. Chapter 6 provides findings and recommendations for future research. Appendices provide further elaboration on the research.

CHAPTER 2

Tire-Pavement Noise

Description of Tire-pavement Noise

Documented research of tire-pavement noise generation and control has been intensive from the 1970s to the present. As vehicular traffic and highway travel speed increased in the 1950s and 1960s, interest in tire-pavement noise rose. The growing interest in tire-pavement noise was amplified by the development of lower noise powertrain units (Sandberg and Ejsmont 2002).

Along with power-train noise and aerodynamic noise, tire-pavement noise is one of the three general components that comprise overall vehicle pass-by noise. Generally speaking, at speeds above 30 mph, tire-pavement noise is the dominant of the three components of pass-by noise. Roadway material properties such as porosity (how large the voids in the roadway surface are), layer thickness (how thick the pavement surface is), and stiffness (how stiff or flexible the pavement surface is) also play a role in tire-pavement noise generation and propagation.

There are many factors which also affect tire-pavement noise. Environmental factors including air temperature, pavement temperature, humidity, wind, and precipitation have all been shown to have differing influences on tire-pavement noise. Test site factors that influence tire-pavement noise measurement results include ground absorption, reflective surfaces, alignments (straight, level, vs. curved and not level), and the pavement specimen itself. Vehicle fleet variability (the mix of vehicle types, the state of maintenance or disrepair, tire type, speed, etc.) also affects tire-pavement noise.

Today many countries, including the U.S., continue to study various aspects of tire-pavement noise abatement. Ongoing studies generally include developing pavement design parameters to decrease tire-pavement noise, developing standardized noise measurement methodologies, quantifying the influence of road surface on traffic and vehicle noise for various roadway types and textures, documenting the acoustic longevity of various roadway designs, and incorporating pavement-noise reduction into traffic noise prediction models. Appendix A contains results of a literature review on tire-pavement noise research.

Measuring Tire-pavement Noise with TP 98 (SIP)

AASHTO Standard Test Method TP 98 describes procedures for measuring the influence of the pavement surface on traffic noise and provides a quantitative measure of the sound pressure level at locations adjacent to a roadway. This test method allows comparison of roadways of varying surfaces. Due to its ability to compare results across test sections and study areas, both standards state that the SIP method is preferred over CTIM, where the SIP method is possible.

Testing Procedures for TP 98

The test site must be carefully selected to encourage free-flowing traffic without acceleration or deceleration. This means that the roadway must be straight and level. Additionally, the test site should be free from barriers or acoustically reflecting surfaces which could potentially interfere with the measurement of sound from the vehicles. The measurement microphones are positioned 25 and 50 feet from the roadway.

Each vehicle pass-by is evaluated to very specific criteria in order to minimize interference from other noise sources; measurements are discarded when they contain potential interference from other vehicles on the roadway, other ambient sounds, or atypical vehicle sounds (such as faulty mechanics or auxiliary equipment on the vehicle). The traveling speed of each vehicle is observed and vehicles are categorized as automobiles, medium trucks, heavy trucks, busses and motorcycles, according to commonly-used definitions by FHWA. The TP 98 test procedure identifies the automobile and heavy truck categories as the minimum vehicle categories for evaluation, and the medium trucks, busses and motorcycle categories are identified as optional.

The sample set for each vehicle category includes the data pair of measured sound level and vehicle speed. A simple linear regression analysis of the maximum pass-by sound pressure level (response or dependent variable) on the common logarithm of the speed (predictor, explanatory or independent variable) will yield an estimated sound level verses speed. There is a simple, unsophisticated procedure described by TP 98 for outlier elimination.

Final Measurement Result for TP 98

The final measurement result from the TP 98 test method is the *statistical isolated pass-by index* or the SIPI result, expressed in A-weighted decibels (dBA). The SIPI result is the difference between the predicted vehicle sound level at a designated speed (L_{veh}) and a reference vehicle sound level at the same designated speed ($L_{veh,ref}$). The designated speed is the mean predictor variable, the average of the common logarithms of the speeds. The reference curve is fixed and is distinct for each vehicle type. The reference curves are based upon a simplified linear model from FHWA standardized models developed in the early ’90s. The data were collected from thousands of vehicles over several general categories of pavement, aggregated to an average pavement.

Measuring Tire-pavement Noise with TP 99 (CTIM)

AASHTO Standard Test Method TP 99 describes procedures for measuring the influence of pavement surface on traffic noise at a specific site adjacent to a roadway. The CTIM measurements are intended for use on high-traffic roadways to capture sound from all vehicle types traveling on all roadway lanes during the measurement period. Measurement techniques aimed at measuring traffic noise, versus isolated pass-byes, are generally applied to roadways where measuring isolated pass-by events would not be possible due to the traffic density. The method currently allows for comparison of varying or aging pavement surfaces on a single-roadway. Extensions of the normalization process to include site variation and allow for site-to-site comparisons are being examined (Rochat et al. 2012).

Testing Procedures for TP 99

Similar to test method TP 98, the test site for TP 99 must be carefully selected to encourage free-flowing traffic without acceleration or deceleration. This means that the roadway must be straight and level. Additionally, the test site should be free from barriers or acoustically reflecting surfaces which could potentially interfere with the measurement of sound from the vehicles. Measurements are performed at a distance of 50-100 feet from the roadway; therefore, measured sound levels include propagation effects over the road surface and ground between the traffic and the measurement locations.

The equivalent-average sound pressure level is measured in regular time-blocks, called analysis time-blocks. The analysis time block duration is from 5 to 15 minutes according to TP 99. Vehicle counts, categories, speeds and any interfering sounds are observed and recorded for each analysis time-block. Then the traffic data are used to build a noise model using conventional traffic noise modeling techniques. The relative differences in the modeled L_{eq} for the time blocks are used to normalize the measured L_{eq} for each analysis time block; this should reduce the effects of different traffic counts, mixes and speeds during each analysis time block. Analysis time blocks which contain interfering sounds are discarded from the data set. A simple outlier elimination procedure is given in TP 99.

Final Measurement Result for TP 99

The results of the *analysis time-blocks* are combined to form a number of 15-minute-long *reporting time blocks* – sound levels are combined using the energy-average L_{eq} of valid analysis time blocks. Naturally, where the analysis time blocks are 15 minutes in duration, the levels are already representative of a 15-minute-long reporting time block. The final measurement from the TP 99 test method is the *average normalized Leq*, expressed in A-weighted decibels (dBA). This is the arithmetic mean of at least three reporting time blocks. For simplicity, the average normalized Leq will be referred to as the CTIM result.

Precision and Bias

ASTM E177-10 defines precision as the closeness of agreement between independent test results, whereas bias is the difference between expected test results and an accepted reference value. Precision and bias are products of measurement variability due to different factors.

While extensive research has been performed concerning the individual factors which influence tire-pavement noise and wayside measurement of traffic noise, few studies addressing the precision and bias of the CTIM and SIP test methods have been performed. The limited studies of precision and bias for statistical pass-by measurement of tire-pavement noise utilized alternate international standards which vary somewhat from the current AASHTO test procedures. The differences between TP 98 and other methods are likely to influence the precision and bias somewhat.

CHAPTER 3

Approach and Experiment Design

Measuring Precision and Bias of a Measurement Process

To appreciate the importance of having both a precise and unbiased measurement process, one needs to understand what is precision and what is bias and the relationship between the two concepts. ASTM E177-10 defines precision as the closeness of agreement between independent test results obtained under prescribed conditions. It is often expressed as a range of numbers at a given confidence level in the same units as the measurement units. The more precise a measurement process is, the closer the values are from the same process measuring the same samples. ASTM E177-10 defines bias as the difference between expected test results and an accepted reference value. An estimate of bias taken from a sample of measurements is the difference between the mean of those measurements and that of the reference or accepted value. The accepted reference value may be a theoretical value based upon scientific principles, an assigned value from some authority, or a consensus value of the scientific community. In the case of the TP 98 and TP 99 test methods, the results are defined only in terms of the respective test methods; therefore, the full extent of bias cannot be known. However, the research team developed a test method such that a component of the bias found in the measurements taken using TP 98 and TP 99 can be estimated. This information is included as part of the bias statement.

Figure 3-1 demonstrates the general notions of precision and bias.

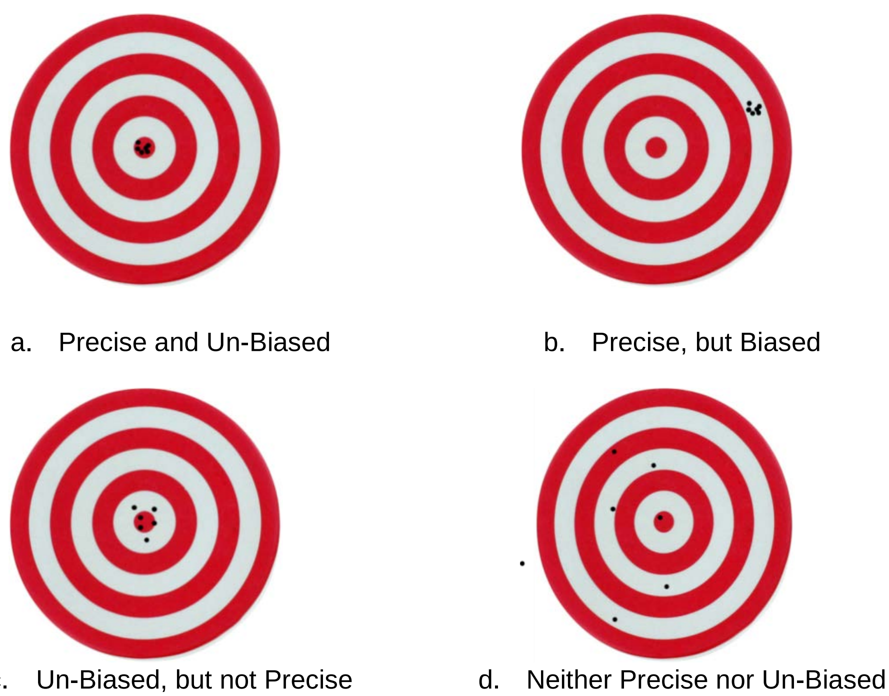


Figure 3-1. Visualization of Differences between Precision and Bias Concepts

Ideally, one wants to design a measurement process which is both precise and un-biased as shown in Figure 3-1(a). The samples as represented by the dots in Figure 3-1(a) have values that are close together and they are also close to the ‘mark’ or reference value which is the area in the center of the target. The measurement process in Figure 3-1(b) is precise since the values from all the samples are close together; however, the process has missed its mark. Figure 3-1(c) shows a process which is un-biased since the values from the samples are near the mark; however, it is not precise since the values from the samples are spread apart. Finally, Figure 3-1(d) represents a process that is highly unreliable as the values from the samples are not consistently near the mark nor are they close together: it is neither precise nor unbiased.

Approach to Measuring Precision of a Process

The manner in which the precision of a measuring process can be numerically quantified is through the variance of the sample’s observations. Variance is a statistic which quantifies how far apart the observations’ measurements are from that of the overall mean. Technically speaking, variance is the average of the squared differences between each observation and the mean of the sample. If every measurement from a sample were identical, then the sum of squared differences between each of the observations and the sample mean would be zero. In other words, the measuring process would have no variance and would have maximum precision. In the real world, this scenario is highly unlikely, as it is rare that a sample of observations randomly selected from a population would be all identical. For example, consider the height of individuals in a classroom. The chances that their heights would be the same are essentially nil.

When considering the variability across the samples from a measurement process, the differences in the samples are not solely due to the differences in the unique properties of those samples. What if the operator who takes the measurements interprets the results differently from other operators while still adhering to the protocols? Returning to the example of the heights of individuals in a classroom, consider the additional variability entered into the sample’s measurements if the person who was taking the measurements did not consistently place the bottom of the ruler on the floor before recording a person’s height. If the methodology standard for measuring people’s heights did not specify that the end of the ruler had to touch the floor, results across different operators and even repeated measurements by the same operator could vary significantly. The variability which can be attributed to the differences in measurements due to the operators and measuring equipment across the samples relative to the overall sample’s variability is what forms the precision statement of the measurement process.

Explanation of Repeatability and Reproducibility

The total variance of a measurement process is the average of the sum of squared differences between each sample’s measured value and the overall sample mean. We know that the variability in the measured values is due to the physical differences in the samples, plus the variability which arises from differences in how operators record the measurements across the samples. This is represented by the following equation.

$$\sigma_T^2 = \sigma_P^2 + \sigma_M^2 \tag{3-1}$$

Where:

σ_T^2 = total variance of a measurement process
σ_P^2 = variance of actual process (physical differences in samples)
σ_M^2 = variance of measurement methodology (operator differences)

The variance component σ_M^2 introduced into the measurement process by the operators can be broken down into repeatability and reproducibility components, as shown in equation (3-2).

$$\sigma_M^2 = \sigma_R^2 + \sigma_r^2 \tag{3-2}$$

Where:

σ_R^2 = reproducibility variance
σ_r^2 = repeatability variance

Repeatability and reproducibility are two key concepts which determine the overall precision of a measurement process:

- Repeatability: The variation observed when the same operator measures the same sample repeatedly with the same methodology.
- Reproducibility: The variation observed when different operators measure the same sample using the same methodology.

The variance in the measurements when different operators measure the same sample using the same methodology is referred to as the reproducibility of the measurement process. If the measurement process is truly precise, then even if different operators measure the same sample, each operator should produce the same measurement.

The repeatability of the total process variation represents the level of variation introduced into the process due to differences in the same sample's measurements when the same operator repeatedly measures it. If a measurement process is truly precise, then each time the same operator measures the sample, he or she should produce the same result. But in reality, this is not always the case.

The present research utilizes a *Repeatability and Reproducibility Study* (R&R study). The purpose of an R&R study, also called a *Gauge Study* (Burdick et al. 2005), is to determine whether or not the measurement procedure variation is small relative to the variation that can be attributed to the test objects or samples. The variation arising from the differences in the samples is called process variation. If measurement variation is relatively large, then the procedure must be improved before it can be useful for monitoring purposes.

For practical purposes, researchers compare the standard deviation of each of the components to the total sample's standard deviation. The standard deviation of a sample of measurements is the square root of the sample's variance. The advantage of using the standard deviation instead of the variance is that the units of the standard deviation are identical to the units of the actual measurements. Precision in terms of deviations in units of decibels versus deviations in units of square-decibels is more easily understood.

R&R studies have two key performance indicators which researchers evaluate to determine if their process is precise or not: the percent R&R value, and the statistical signal-to-noise ratio (SNR). The percent R&R value is the ratio of the standard deviation from the R&R component to that of the total sample's standard deviation multiplied by 100 as shown in equation (3-4).

$$\text{Percent R\&R Value} = \frac{\sqrt{\sigma_R^2 + \sigma_r^2}}{\sqrt{\sigma_P^2}} \tag{3-4}$$

If this value is greater than 30 percent, the measurement process is deemed inadequate and not precise (AGIG 2010, p. 78).

The SNR is the square root of the ratio of the actual process variance component to that of the R&R variance component. This is a statistical metric, distinct from the acoustical signal-to-noise ratio. The statistical SNR is calculated according to equation (3-5).

$$\text{SNR} = \sqrt{\frac{2\sigma_P^2}{\sigma_R^2 + \sigma_r^2}} \tag{3-5}$$

Values of 3 or less for this ratio indicate that the process is inadequate. A value greater than 3 is required for the measurement process to be considered reliable (AIAG 2010, p. 171).

As described in ASTM E177-10, the standard deviations of repeatability and reproducibility when multiplied by 2 provide an approximate 95 percent confidence interval for the sample means of future tests conducted under similar reproducibility or repeatability conditions. These same standard deviations of repeatability and reproducibility when multiplied by 2.8 provide a 95 percent confidence interval for the difference of means from pairs of tests also conducted under similar reproducibility or repeatability conditions. Consider the following examples.

- If one operator performs many tests on the same sample, then the average of the SIPI measurements would fall within the range ($2 \times \sigma_r$) at the 95 percent confidence level.
- If many different operators perform tests on the same sample, then the average SIPI measurements would fall within the range ($2 \times \sigma_R$) at the 95 percent confidence level.
- If an operator randomly runs two tests on the same sample, then 95 percent of the time the difference between those two test results would not be more than ($2.8 \times \sigma_r$).
- If many operators ran pairs of tests, then 95 percent of the time the difference between those two test results would not be more than ($2.8 \times \sigma_R$).

Experimental Design for Repeatability and Reproducibility Study

R&R studies are ideal methods to measure precision and bias (provided a reference value is known) because they decompose the variability across the major sources of variability which issue from the differences in the parts, the operators, and the measurement process itself. With a large enough sample of measurements, the R&R study can be used to set a precision statement at a given confidence level so that future users of the process can determine if their measurements are adequate. By specifying the number of samples, operators and replicates, the total size of an R&R study can be planned. For example, if one has six different samples and six different operators measure all six samples twice (two replicates), the total samples size is 72.

This research started with a limited pilot study to determine the optimum number of samples and operators within the constraints of the project. The number of replicates was set at two, because with any more replicates the operators are likely to “learn” the samples and memory of the previous measurements could influence the results. The limited pilot study is presented in greater detail in an appendix and includes a rigorous statistical analysis including statistical signal-to-noise and power evaluations. Analysis of the limited pilot R&R study resulted in the following experimental design for the full R&R study, identical for both TP 98 and TP 99:

- five operators qualified to perform acoustic measurements;
- four process samples which are calibrated audio and synchronized video to present the operator with a repeatable measurement opportunity of a particular pavement during a period with repeatable environmental conditions and vehicle fleet;
- two replicates – each operator will read each sample twice; the order that samples will be processed will be randomized and will be unknown to the operator.

The limited pilot R&R study indicated that this combination was the largest that could be accommodated by the research project budget and still result in the highest predicted power and statistical SNR for the full R&R study.

The purpose of these tests is to measure the variance of the test procedure itself. Therefore, these tests exist within a larger repeatability and reproducibility study. For purpose of this study, the actual loudness or quietness of any particular pavement is less relevant than the differences between multitudes of test results. In other words, this research is not evaluating tire-pavement noise, but rather it is evaluating the measurement itself. In the course of executing the measurement to evaluate, the research just happens to also be generating tire-pavement noise data.

Operator Presentation

An important design consideration is replication of the experiment and how to create conditions that are consistent among replicates. Since noise is generated by passing traffic, once an operator conducts his or her measurements on that sample of traffic, it is no longer available for re-sampling. Traffic and environmental conditions are continuously changing. To evaluate repeatability where each operator is presented with the same measurement subjects, this research opted for presenting operators with calibrated audio and synchronized video of two live traffic sessions for analysis.

The premise of the test method evaluation is to use recordings of calibrated audio and synchronized video to control the measurement subject, the measurement environment, and the otherwise random vehicle mix. Each recording presents the operator with a particular pavement specimen under reproducible measurement conditions. The measurement conditions for each recording comprise the environmental conditions present during rerecording and common vehicle sample sets which drove over the pavement specimen. The replicates are used in this context to determine how much variance is attributable to operators when presented with the same measurement opportunity – a situation where all other factors are unchanged, including identical meteorological conditions and an identical vehicle fleet.

Approach to Measuring Bias of a Process

As previously stated, the TP 98 and TP 99 test methods technically have no bias because the test results are defined only in terms of the respective test methods. A theoretical accepted reference value based upon scientific principles is not available, and is impractical. This would need to account for everything that goes into the measurement, including the pavement itself as well as the tires, drive-train, and aerodynamic characteristics from each vehicle category of the vehicle fleet to determine a theoretical accepted reference value.

In lieu of a theoretical approach to address the issue that neither the TP 98 nor TP 99 test methods have known accepted reference values, the research team estimated some *limited components* of bias. One component of a measurement bias is the effect of the measurement environment. The measurement environment includes ground cover, ground profile, reflecting objects, and the atmospheric characteristics such as temperature, wind, and humidity. The environment is not entirely controllable from location to location, from day to day, or even over the course of a single same-day measurement. The research team conceived a means of measuring the effect of the environment at particular measurement sites, borrowing from other standardized acoustical test methods, as described below.

To characterize the effect of the acoustic environment on a measurement, several standardized test methods employ a Reference Sound Source (RSS). The RSS has known acoustic emission characteristics which are usually established using measurement standards ANSI S12.5 or ISO 6926 or ASTM E1179, or one of the ISO 3740 series of sound power measurements depending upon the application. The research team measured the acoustic emission characteristics of a RSS in an acoustical laboratory. The difference between the RSS measurement in the field and the calculated sound pressure level based on the known RSS acoustic emission characteristics is a component of bias. It represents the effect that the environment has on the known noise level of the reference sound and the operator’s measurement of it.

Identifying Influential Factors

The research team also conducted a ruggedness test of the AASHTO standard test methods TP 98 and TP 99 to determine if certain environmental or physical factors can influence measurement values. A ruggedness test is type of experimental design that allows researchers to determine if factors of interest have a statistically significant impact on the outcomes of a measurement process. For the purpose of this study, we are interested in testing if certain factors are influential in the final measurements produced using the test methods TP 98 and TP 99.

Through the application of statistical tests, conclusions are made regarding whether the factors impact or do not impact the measurement values in a statistically significant manner. If the certain factors are found to be statistically significant, then these factors may potentially influence the ability to accurately and precisely measure traffic sound levels. Factors which can be assessed in a ruggedness test include, but are not limited to, the type of vehicle, mix of vehicle types, pavement type, tire type and condition, condition of the vehicle, time of day, temperature, humidity, surrounding vegetation, presence or absence of reflecting objects, and composition of any such objects. The research team aimed for seven factors, the maximum number of factors for an 8-run ruggedness test; however, the field data collection did not support two of the factors in the experimental design. The field data supported working with the following five factors for ruggedness testing:

- Factor A - Pavement Age: (newer or older);
- Factor B - Ground Impedance: (harder or softer);
- Factor C - Speed Regime: (highway or interstate);
- Factor D - Measurement Option (shorter or longer); and,
- Factor E - Air Temperature (warmer or cooler).

The actual experimental design of the ruggedness test is referred to as the Plackett-Burman design. The purpose of this design is to allow researchers to quickly evaluate many factors (with the smallest number of tests) to see which ones are important (or not important). This is possible since Plackett-Burman designs are orthogonal, meaning that the factors behave independently of each other (i.e., changes in the levels of each factor are not related to changes in the levels of the other factors). Plackett-Burman designs are also referred to as "screening designs" because they help researchers screen out unimportant factors.

Generally these tests require that the number of levels per factor is equal to two, typically indicated as (-1) and (+1) in the design stage. The selected two levels for any factor represent relatively opposing values such as warmer versus cooler or slower versus faster. The statistical test evaluates if the differences in the measurements can be attributed to the differences in those two levels. If the differences are not statistically significant, then the observed differences are a result of random fluctuations. Given the constraints of the research project (cost and schedule), the research team conducted eight test runs, each involving a different treatment combination, or setting, of the five factor levels (see Table 3-1). A single operator was used to help control within-test variation as much as possible.

Table 3-1. Design for Five Factor, Eight Run Plackett-Burman Screening Design

	Factor A	Factor B	Factor C	Factor D	Factor E
Run 1	+1	+1	+1	-1	+1
Run 2	-1	+1	+1	+1	-1
Run 3	-1	-1	+1	+1	+1
Run 4	+1	-1	-1	+1	+1
Run 5	-1	+1	-1	-1	+1
Run 6	+1	-1	+1	-1	-1
Run 7	+1	+1	-1	+1	-1
Run 8	-1	-1	-1	-1	-1

Note: +1 = Level 1; -1 = Level 2.
Source: ASTM E1169-07

After substituting the test levels for the +1 and -1 values, the final Plackett-Burman design for both test methods TP 98 and TP 99 is shown in Table 3-2 below. The definitions for all factors' levels with the exception of Factor D, Measurement Option, were identical. With respect to the measurement option factor, TP 98's definition of shorter and longer are "microphone 25 ft. from road" and "microphone 50 feet from road", respectively, while TP 99's definition is "5 minute blocks" and "15 minute blocks", respectively. With respect to the air temperature factor, runs identified as cooler were measured in air temperatures between 57°F and 70°F; warmer runs had air temperatures greater than 75°F and less than 85°F. (See Chapter 4 and Chapter 5 of this report for the full TP 98 and TP 99 ruggedness test methodology and measurements).

Table 3-2. Test Levels for Five Factors, Eight Run Plackett-Burman Design

	Factor A Age	Factor B Ground	Factor C Speed	Factor D Meas. Option	Factor E Air Temp.
Run 1	Older (+1)	Softer (+1)	Interstate (+1)	Shorter (-1)	Cooler (+1)
Run 2	Newer (-1)	Softer (+1)	Interstate (+1)	Longer (+1)	Warmer (-1)
Run 3	Newer (-1)	Harder (-1)	Interstate (+1)	Longer (+1)	Cooler (+1)
Run 4	Older (+1)	Harder (-1)	Highway (-1)	Longer (+1)	Cooler (+1)
Run 5	Newer (-1)	Softer (+1)	Highway (-1)	Shorter (-1)	Cooler (+1)
Run 6	Older (+1)	Harder (-1)	Interstate (+1)	Shorter (-1)	Warmer (-1)
Run 7	Older (+1)	Softer (+1)	Highway (-1)	Longer (+1)	Warmer (-1)
Run 8	Newer (-1)	Harder (-1)	Highway (-1)	Shorter (-1)	Warmer (-1)

Approach to Influential Factors Testing

The research team used regression analysis to test if the selected factors influenced the values for either the SIPI or CTIM measurement collected under the protocols specified by test methods TP 98 and TP 99, respectively. A regression analysis summarizes the statistical efficacy of a proposed regression model, which relates mathematically the values of a dependent variable to those of a set of independent variables. Often the dependent variable is represented using a Y and the independent variables are represented as a series of factors as shown in equation (3-6)

$$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 D + \beta_5 E + \varepsilon \tag{3-6}$$

Where:

Y = measured value for either the SIPI or the CTIM measurement for a given run
$\mathrm{A, B, C, D, E}$ = observed conditions of factors for the given run
β_0 = constant representing mean Y when all factors have value of 0
$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ = regression coefficients representing effect of factor conditions
ε = experimental error

Estimated regression coefficients are output from the regression analysis. The regression coefficients represent the impact that each of their respective factors has on either the SIPI or CTIM measurements. If a coefficient has a positive value, it means that as one goes from a lower factor level to a higher factor level as defined in Table 3-2, the expected value of the measured response increases. If the coefficient is negative, then the effect of changing levels is to reduce the value.

Each coefficient has a test statistic calculated to determine if the value of the coefficient is statistically different from zero. The p-value of the test statistic, which ranges in values from 0 to 1, is compared to a statistical concept called the significance level. The value of the significance level is used to establish the criterion of either accepting that there is no impact due to the factor or there is an impact. Often, researchers use a 0.05 significance level as the acceptance or rejection criterion. If the p-value of the test statistic is less than 0.05, it means that the probability of obtaining an estimated effect as large as the one obtained from the regression analysis, when the true factor effect is zero, is less than 5 percent. In other words, the observed effect is statistically significant at the 5 percent significance level. With such a low probability, researchers have confidence that the effect they have observed is real and not as a result of random fluctuation. Note that the smaller the p-value is for the test statistic of a given coefficient, the stronger the evidence is that the coefficient captures a real impact.

For the Plackett-Burman design used here, there will be five tests conducted, each at the 0.05 level of significance. Since each of the five tests has a 0.05 chance of an error, the error rate for the analysis can be as high as 25 percent. To keep the overall (experiment-wide) error rate at 0.05 percent, it is standard practice to conduct each test at the 0.05/5 = 0.01 level. (This is called the Bonferroni adjustment and is described in Kutner, et al., 2005.) Thus, in what follows, a factor will be considered statistically significant if its p-value is less than 0.01.

One issue that must be addressed in the analysis is the fact that only one replicate of the experiment was employed. This means that there is only one observation for each of the eight treatment combinations in the experiment. This is nearly always the case when small screening designs, such as a Plackett-Burman design, are used for ruggedness testing. As a result, there is very little information about experimental error (i.e., repeatability). In the current study, the eight runs must be used to estimate the six regression coefficients, leaving only two degrees of freedom for the estimation of experimental error. For this reason, the tests for significance of the experimental factors have relatively low statistical power. (That is, they may conclude that the factor effect is zero, when it is nonzero.) In summary, the tests may not be as reliable as desired, due to the small sample size.

An analysis procedure which can overcome this problem is to use a model selection technique. Model selection techniques generally take one of two forms: stepwise model selection, or best-subsets model selection. In stepwise model selection, the regression equation starts out empty and the factors are added one at a time, depending on the importance of the added factor. The next factor to be added to the model is the factor that, when added, leads to the model having the most information about the response variable. The amount of information provided by the model can be measured using a variety of criteria. Popular alternatives include adjusted R-square, the mean square error (MSE), or the p-value of the added term. Akaike's (corrected) Information Criterion (AICc) is particularly well suited to analyses of data sets having small sample sizes, as is the case here (Hurvich 1989). In best-subsets model selection (sometimes called all-possible-regressions selection), all possible regression models are evaluated by the information criterion and the best model is chosen. The stepwise model section method with the AICc (Stepwise/AICc selection method) is simpler to use in determining if any of the five factors are influential since either of the model selection methods, Stepwise/AICc or best-subsets/AICc selection, is generally effective when used in conjunction with orthogonal designs such as the Plackett-Burman design.

For each of the standard test methods TP 98 (automobiles and heavy trucks) and TP 99, the research team ran a full model regression analysis, produced p-values for the coefficients and produced the AICc value for each step of the Stepwise/AICc selection method. The results of such tests per test method are described in the following three sections. The complete output from the statistical analysis of the data from the ruggedness tests for standard test methods TP 98 (automobile and heavy trucks) and TP 99 is in Appendix D of this report.

CHAPTER 4

Data Gathering and Measurements

Measurements for Repeatability and Reproducibility Study

In adherence to the approach for the Repeatability and Reproducibility Study, the samples were presented to the operators as recordings of calibrated audio and synchronized video of the traffic. The research team gathered field recordings with instrumentation microphones connected to a digital data acquisition unit, while a video camera simultaneously recorded all the traffic at the measurement position. The digital data acquisition unit captured the microphone signal in the field – the road traffic noise as well as calibration tones – for later playback to the test operators. The recording and playback systems have been qualified using procedures which conform to and expand on the 2005 version of ANSI S1.13, §8.7, to ensure that the microphone signal playback in the office precisely replicates the original acoustic signal in the field.

Sample Recordings

Due to the volume of heavy trucking in the region, the favorable climate and project scheduling, measurement sites were selected primarily in Arizona. Sites were initially screened using an assortment of information sources, including aerial photography, coarse topography, traffic data from Arizona Department of Transportation (ADOT), and other imagery publicly available through Internet sources. These were examined to find portions of roadway which meet the criteria outlined in the measurement standards. Searches were limited to four ADOT Engineering Districts which appeared to be the most likely to offer suitable candidate measurement sites. Districts to the northeast were discarded because the area is generally very mountainous. The team obtained encroachment permits for segments of candidate roadways. Field staff scouted the roadway segment for suitable sites which conform to the measurement standards, and identified the best location to set up equipment as the work area. Once the work area is identified, then the temporary traffic control (TTC) devices ("shoulder work ahead" signs) are deployed upstream of the work area as required by the ADOT encroachment permit, and then the equipment setup and measurements commence.

Sites were generally selected irrespective of the pavement type, in accordance with the ability of the R&R study to partition the variance due to the actual process (physical differences in samples) from the variance due to the measurement methodology (operator differences). As noted in the section on the experimental design, the R&R study is not evaluating the actual tire-pavement noise, but rather it is evaluating the measurement itself.

Table 4-1 shows the measurement sites selected for the R&R study.

Table 4-1. Measurement Sites for Repeatability and Reproducibility Study.

Site	Route	Side of Road	Milepost	Offset	Test Method	
E	SR 86	South (EB)	162	+ 3550 ft	TP 98	TP 99
G	I-8	South (EB)	163	+ 3530 ft	TP 98	
H	I-8	South (EB)	108	+ 2900 ft	TP 98	TP 99
K	I-10	North (WB)	27	+ 3430 ft		TP 99
M	I-40	North (WB)	134	+ 3750 ft	TP 98	TP 99

Pavements which exhibit joints, seams, or other imperfections were not excluded from the site selection because while they may not offer ideal conditions for measuring the pavement surface, it is likely that a before-and-after comparison would require measuring a pavement surface which exhibits such imperfections. Therefore they were not excluded in the possible selection of candidate sites.

Field Data Acquisition

Two microphones were deployed at the measurement location. The microphones were positioned according to the AASTHO standards, one at a distance of 25 feet from the centerline of the outermost lane of travel and the other at a distance of 50 feet from the centerline. They were positioned 5 feet and 12 feet, respectively, above the surface of the pavement (the pavement elevation at the centerline of the outermost lane of travel). Positions are illustrated in Figure 4-1. Microphones conformed to the Type 1 precision and their directional responses were random-incidence type, as required by Section 6 of TP 98 and TP 99. The microphone and preamplifier assemblies are oriented vertically. Consequently, the incident road traffic noise is at a grazing angle on the microphone diaphragm.

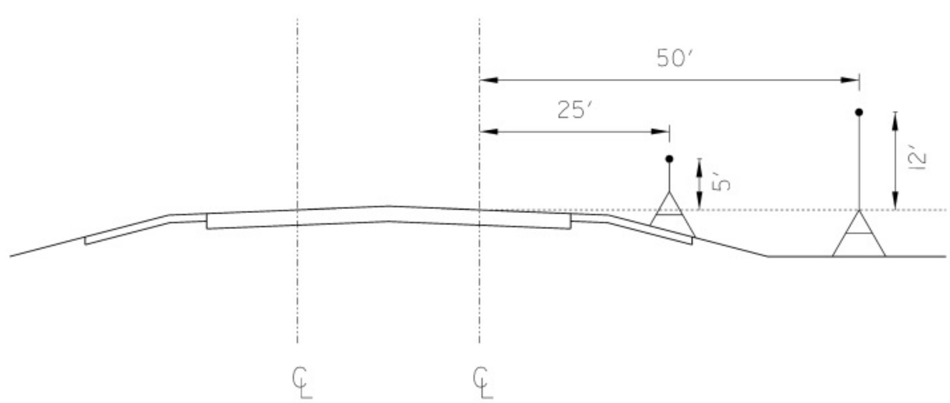


Figure 4-1. Microphone Positions

A wind sensor gathered continuous wind speed and direction data near the microphone locations. An observation vehicle was parked beyond the shoulder and at least 150 feet from the microphone position in the upstream traffic direction. The digital audio signal recorder, the video camera, and the traffic sensor were all located at this observer position. Each of the two microphones was connected to a microphone splitter; one side of each splitter was connected to a microphone power supply and a supplemental sound level meter (SLM) and data-logger. The other side of each microphone splitter was connected to a two-hundred-foot cable and a digital signal recorder, located at the observer position. The observer would listen with headphones to monitor for audibly detectable interfering signals. To ensure accurate reproduction from the recordings, gain settings in the recording system were adjusted prior to recording the calibration signal to ensure a good SNR ratio and avoid peak-clipping of the audio signal. After recording the calibration signal, the gain was not adjusted for the remainder of the measurement. Calibration signals were recorded at approximately hourly intervals and at the end of the measurement using a handheld acoustic calibrator meeting ANSI and IEC Type 1 tolerances, traceable to National Institute of Standards and Technology (NIST) reference quantities.

The video camera was a compact, wide-angle, high-resolution camera, intended to be used as an action/sports camera. It was mounted stationary and pointed toward the roadway in front of the microphone positions; this gives the view from the observer position for playback observations. The traffic sensor was a microwave vehicle detector which logged individual vehicle pass-by times, lanes and speeds. This was used for the speed of individual pass-by events for the SIP method, and also to model the normalization factors for the CTIM method. It was mounted on a tall, stable mast (approx. 20-25 feet above the road surface), and the software was configured for the lane distances and angles at the site, adjusting thresholds as necessary.

Upon completion of the apparatus setup, the field test operator commenced recording audio, video, traffic and meteorological data. For the duration of the measurement the field test operator was in the vehicle to observe traffic and monitor the recording devices. Both test methods require at least hourly logging of local meteorological readings, including wind speed and direction, air temperature, relative humidity and pavement surface temperature. Wind speed and direction was continuously logged. Air temperature, relative humidity, and pavement temperature were manually logged on an hourly basis from a handheld sensor with air probe and infrared sensors.

Calibration of Recordings

For normal calibration of precision acoustic instrumentation in the field, the SLM takes measurements in the field using an attached measurement microphone. A handheld calibrator produces a reference signal, typically a 1 kHz sine-wave with a sound pressure of either 1 Pa or 10 Pa (respectively 94 dB or 114 dB re. 20 µPa). Typically the calibrator is fitted over the microphone cartridge, and the SLM is adjusted manually or automatically to read the correct level from the reference signal. Conceptually, this same process occurred for the R&R measurements; however, the electrical signal from the microphone to the SLM was interrupted by a recording and a playback system. All recordings gathered in the field included hourly recording of calibration tones generated by a handheld acoustical calibrator. During preparation for the R&R study, all calibration tone signals were reviewed for deviation or calibration drift. The largest change in calibration readings for a single microphone over the duration of recordings at a single site was 0.22 dB; there were seven microphone recordings with maximum changes between 0.10 and 0.16 dB; and all others were below 0.10 dB. Calibration drift on this order is not uncommon and is usually attributable to environmental changes, especially changes in ambient air temperature, the temperature of the microphone components, or static air pressure.

For operators performing each measurement, the output of the playback system replicated the electrical signal that the microphone produced, and the sound level meter measured the signal as if a microphone were connected. To calibrate the recorded audio in the office, the recorded calibration tone was played and the sound level meter was adjusted to the calibration tone. Subsequent measurements in the office were then accurate to the original acoustic signal. The playback systems exhibited very little calibration drift, generally not more than a couple hundredths of a decibel.

All the recording and playback equipment were qualified and calibrated to ensure the acoustic observations were unaffected by the introduction of the recordings. Neither the recording system nor the playback system employed automatic gain control, audio limiting or audio compression. The system recorded the entire audio signal without any data compression and stored the signal in a .wav file format.

Test Operator Procedures

Each measurement consisted of the output of the digital audio recording system connected to the input of a sound level meter. The output of the recording system mimicked the electrical signal that the microphone produced, and the sound level meter measured the signal as if a microphone were connected. The calibration tone was played first, and the sound level meter was adjusted to the calibration tone. Subsequent measurements were then accurate to the original acoustic signal. Each operator used the same type of playback system for all sampled sites. All instrumentation used in the office to measure the recorded acoustic signals met Type 1 specifications and underwent NIST-traceable maintenance calibration regularly. The video recording of passing traffic was synchronized to the digital audio for playback and allowed for vehicle counting and classification.

The order of presentation was randomized using a sequence of random numbers, and each operator was assigned a unique random order to perform these measurements.

Procedures for TP 98 Measurements

Each operator measured the noise level of each pass-by event, and evaluated the pass-by measurement, and evaluated the audio and video playback. Evaluating the measurement, the audio, and the video, each operator categorized the vehicle and checked that each vehicle pass-by event met all criteria for inclusion in the measurement sample set as described by TP 98. The approach to appraising the vehicle pass-by events was up to each operator. All operators were familiar with the acoustical and non-acoustical requirements for a valid pass-by according to the standard. They chose independently how to determine validity using any tools at their disposal which also met the requirements. If the pass-by met all qualifications, then the operator included the measured sound level and the vehicle speed data pair in the sample set. Data pairs were separated into their respective vehicle classes. The resulting count of vehicle observations varied between operators for each measurement and each vehicle class. The operator performed analysis of each measurement as described by TP 98, and produced a full set of results; the final result for determining the precision and bias of TP 98 is the SIPI. Each operator/recording/replicate combination produced two final SIPI results for TP 98, one each for the two vehicle types: automobiles and heavy trucks.

Procedures for TP 99 Measurements

The calibrated audio signal and synchronized video were used to measure the nose level of analysis time-blocks, count and categorize traffic, and verify that each analysis time block met all criteria for inclusion in the measurement sample set. During playback, each operator started measuring time-blocks; the duration of measurement time-blocks was left to each operator (from 5 to 15 minutes according to TP 99), however all operators opted for 5-minute time-blocks. During each time block, the operator measured sound level as well as traffic volume (count of cars and trucks) and traffic speed, and made note of any interfering sounds. Time-blocks with interfering sounds would be discarded from the data set.

Each operator performed analysis of each measurement as described by TP 99. Operators normalized each valid analysis time-block, eliminated outliers, determined analysis time-blocks and then found the final CTIM result, the average normalized L_{eq}. Each operator/recording/replicate combination produced a CTIM result for TP 99.

Quality Control of Measurements

All operators received training on the measurement process described in the TP 98 and TP 99 test standards, as well as instructions to correctly calibrate and measure from the digital recordings. Careful attention was given to instructing the standards only; operators were given copies of TP 98 and TP 99 and were allowed free interpretation where the standard didn't prohibit it. For TP 98, operators were instructed to measure the two vehicle categories recommended as a minimum by the standard: automobiles and heavy trucks.

All measurements were reviewed to ensure conformity to the measurement standards, and that the operators didn't (unknowingly or intentionally) violate the written requirements of either TP 98 or TP 99. The first measurement of each standard from each operator was reviewed in order to avoid perpetuating an error through an entire operator's measurement set. Upon completion, the whole of the measurements were reviewed for correct computation of the results, based upon the operators' observations.

Measurements for Bias Study

The bias study is intended only to characterize a particular component of bias, specifically the measurement environment. The study utilized an RSS placed on the pavement directly in front of the microphone positions, and measuring the sound pressure level at the microphones. These measurements were performed during the R&R site recordings at the hourly recording breaks, at the same time as the calibration checks. This produced two values for the RSS: (1) a measured sound pressure level of the RSS observed from the field data, and (2) a calibration value of the RSS based upon precision laboratory measurements and established acoustic principals. A small number of RSS measurements were discarded due to interfering noises.

Reference Sound Source

Commercially available aerodynamic RSSs generate broadband random noise through a motor and a blower (the blower is often an unshielded "squirrel cage" fan). However, the motors generally draw too much current to be practical in a remote location (they would blow the fuse of the observation vehicle). The RSS utilized for the present research was a battery-operated motor with a blower. The battery-operated motor had significantly lower power than for other commercial RSSs, which means it also generated lower sound levels. However, this RSS provided ample signal-to-noise in the mid- and high-frequency range, as well as sufficient signal-to-noise of A-weighted sound pressure levels, as measured in the field. The sound power of the RSS was measured in an accredited laboratory using test method ISO 3741, a precision measurement method in a reverberation chamber, combined with the directivity index measured in one direction of interest (in the direction of the microphones for TP 98 and TP 99) using an abbreviated version of ISO 3745, a precision measurement in a hemi-anechoic chamber.

Bias Calibration Value

The calibration value was determined by the precision laboratory measurement described above, along with computing the expected sound level at each microphone distance. This was simply the hemispherical geometric divergence – the spreading of sound over a reflecting plane, the roadway – along the diagonal distance to each microphone from the position of the RSS on the roadway.

Measurements for Influential Factors Study

There were eight runs for the influential factors study which resulted in eight test results each of TP 98 automobiles, TP 98 heavy trucks, and TP 99. These measurements were executed similarly to the R&R measurements, utilizing calibrated audio and synchronized video. This allowed the operator to evaluate the data quality back in the office. There was one operator each to perform the test methods.

Measurement Planning and Site Selection

Given the specific requirements of each measurement site and difficult-to-predict weather, the research team obtained encroachment permits for long segments of highways and interstate freeways in southern Nevada and western Arizona. The segments were scouted for the pavement age and the ground type alongside the roads to identify candidate sights which meet factors A, B and C. Weather forecasts were monitored for appropriate measurement times to meet factor E. Factor D was simply controlled with the execution of the test method for that particular run. This resulted in the following measurements for the influential factors study.

Table 4-2. Measurements for Influential Factors Study.

Run	Site	Route	Age	Ground	Speeds	TP 98 Mic	TP 99 Blocks	Air Temp.
1	K	NV I-15	Older	Softer	Interstate	25 ft.	5 min.	Cooler
2	P	AZ I-8	Newer	Softer	Interstate	50 ft.	15 min.	Warmer
3	Q	NV I-15	Newer	Harder	Interstate	50 ft.	15 min.	Cooler
4	R	AZ US-93	Older	Harder	Highway	50 ft.	15 min.	Cooler
5	S	NV US-95	Newer	Softer	Highway	25 ft.	5 min.	Cooler
6	T	AZ I-10	Older	Harder	Interstate	25 ft.	5 min.	Warmer
7	U	AZ SR-68	Older	Softer	Highway	50 ft.	15 min.	Warmer
8	V	AZ SR-85	Newer	Harder	Highway	25 ft.	5 min.	Warmer

Recall that for the R&R study, all operators uniformly performed TP 98 using only the 25-foot microphone position, and TP 99 using only the 5-minute analysis time-blocks. In contrast, this influential factors study investigates whether changing these respective parameters would produce a different level of variation. Therefore, this set of measurements includes both microphone positions for TP 98, and two different time-blocks for TP 99.

CHAPTER 5

Statistical Analysis

Development of Precision and Bias Statements

The following sections describe the statistical results of the present research. The research team designed and implemented an experimental design known as Repeatability and Reproducibility (R&R) Study to measure the expected limits on the precision of noise measurements determined using either test methods TP 98 and TP 99. With respect to measuring bias, the research team designed and implemented another experiment to measure a component of bias since a reference value for any one measurement from either TP 98 or TP 98 is not known.

Repeatability and Reproducibility Study Results

The research team implemented an R&R study to break down the variability from a sample of noise measurements collected and analyzed using measurement protocols specified by test methods TP 98 and TP 99. R&R studies are ideal methods to measure precision and bias (provided a reference value is known) because they distinguish between the major sources of variability due to the differences in the samples, the operators, and the measurement process itself.

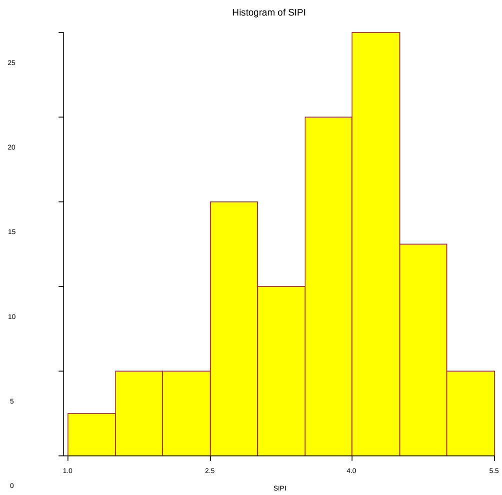
With a large enough sample of measurements, the R&R study can be used to set a precision statement at a given confidence level so that future users of the process can determine if their measurements are adequate. By specifying the number of samples, operators, and replicates, the total size of an R&R study can be planned. For example, if one has six different samples and six different operators measure all six samples twice (two replicates), the total samples size is 72. For this research, the experiment design for both TP 98 and TP 99 test methods is identical: four samples, five operators, and two replicates. This combination was the largest that could be accommodated by the research project budget and still result in the highest predicted power and statistical SNR. This particular design generated a sample of 40 different measurements per test method. The following sections describe the measurement outcomes and results of the R&R study for test methods TP 98 automobile, TP 98 heavy truck, and TP 99. The test results are formatted in a fixed structure appropriate for the estimation procedure. The research team used NCSS software (Hintze 2009) to generate estimates of the variance components and their levels of precision. Precision results for test method TP 98 were separated by class of vehicle since each of the two vehicle classes have unique acoustic properties which warrant separate analyses.

Statistical Analysis of Test Method TP 98 Repeatability and Reproducibility

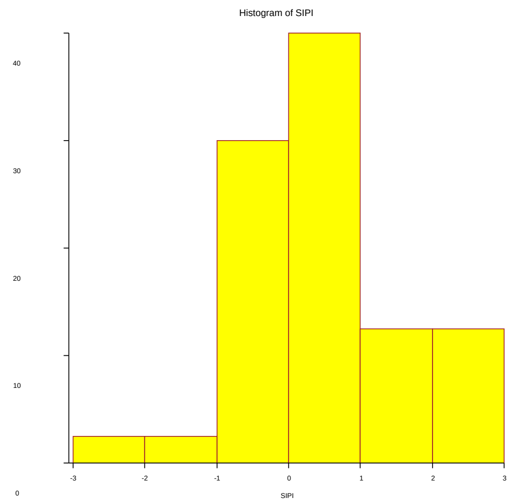
Each sample from the four sites E, G, H and M is processed by five different operators and each operator processes each sample twice. Hence, there is a total of 40 tests in this design. The test measurement outcome for TP 98 is referred to as the Statistical Isolated Pass-by Index or SIPI, and is presented as separate SIPI results for each vehicle category measured.

Figure 5-1 shows histograms of SIPI results for TP 98, one histogram each for automobiles and heavy trucks. While the protocols to collect the SIPI from heavy-truck-traffic sound levels are identical to that for automobiles, the SIPI for the two vehicle categories have a very different distribution. The average

SIPI for automobiles measured across the 40 tests is 3.64 dBA with the lowest value being 1.39 dBA and the highest value being 5.40 dBA. The average SIPI for heavy trucks measured across the 40 tests is 0.43 dBA with the lowest value being -2.69 dBA and the highest value being 2.60 dBA. The statistical results of the R&R study are presented separately for the vehicle categories: first automobiles, then heavy trucks.



a. Histogram of SIPI for Automobiles

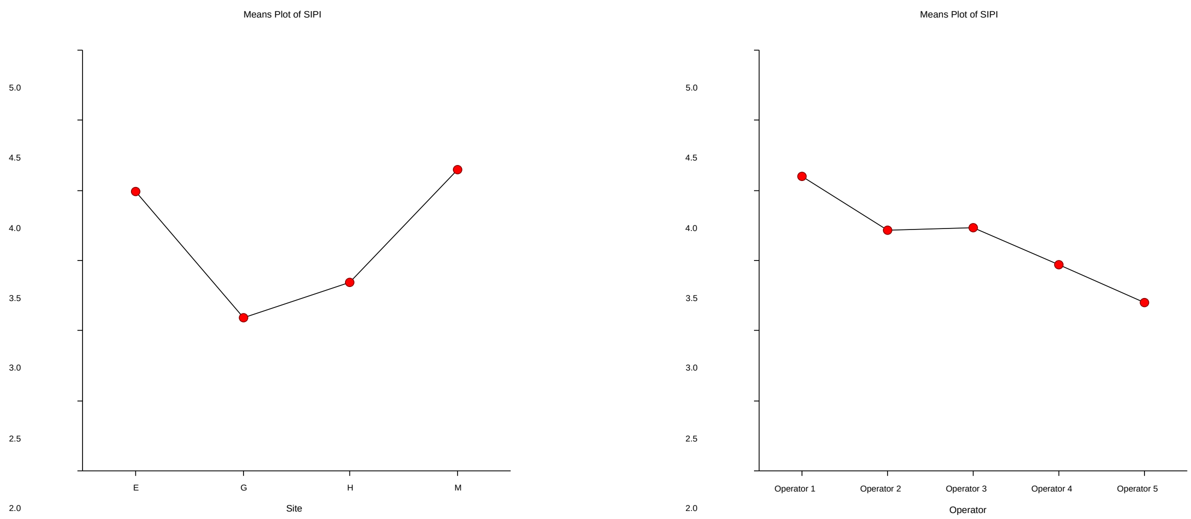


b. Histogram of SIPI for Heavy Trucks

Figure 5-1. Test Method TP 98 (Automobiles & Heavy Trucks) Histogram of SIPI Measurements

Statistical Results of the R&R Study for TP 98 (Automobiles)

Plots of the average SIPI measurement per sample site and then per operator in Figure 5-2 (a) and (b) below show that the range (difference between the highest value and the lowest value) of the average SIPI for both figures is about 1 dBA. These graphs represent the mean of all sites, per operator. The data values used for the plots are in Appendix C, along with additional data plots. In a precise system, the range in averages across the operators would be much smaller than that found in the range of averages across the sites. Imagine two measurements of different pavements by different operators; if those measurements differed by 1 dBA, that difference could be attributable either to the differences in the tire-pavement noise, or the differences between the two operators.

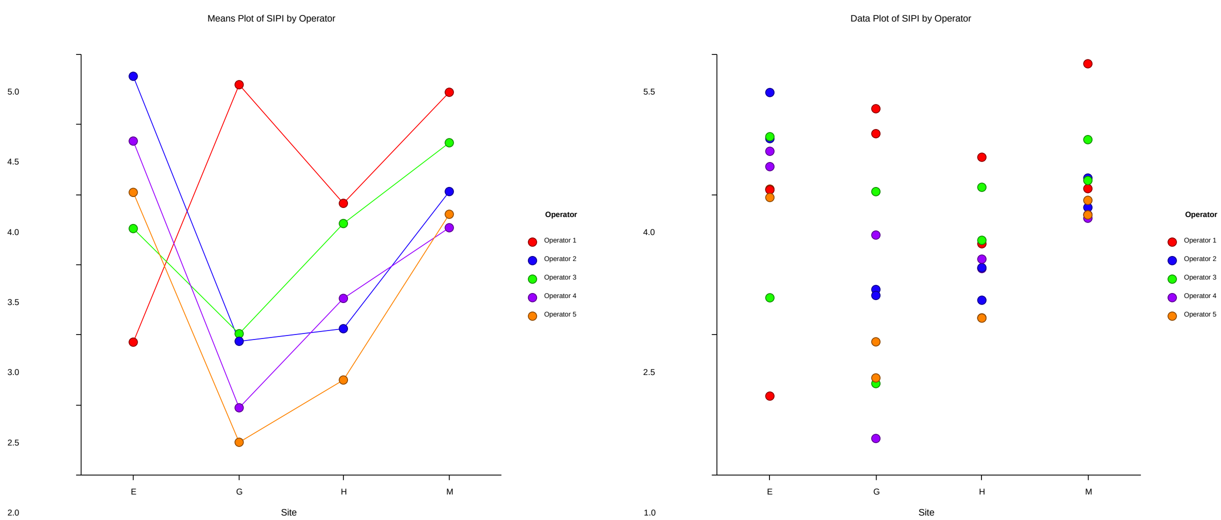


a. Plot of the Average SIPI Measurements per Site

b. Plot of the Average SIPI Measurements per Operator

Figure 5-2. Test Method TP 98 (Automobiles) Plots of Average SIPI per Site and per Operator

Figure 5-3 demonstrates the level of variability across the measurements by operator per sample site. Depending on the operator at a given site, the spread of measurements can be relatively large and contrary. For example, operator 1's measured values from site E and G are significantly lower and higher, respectively, than those measured and recorded by the remaining operators.



a. Plot of Average SIPI Measurements per Operator per Site

b. Plot of Individual SIPI Measurements per Site

Figure 5-3. Test Method TP 98 (Automobiles) Variability Trends across Sites and Operators

Close inspection of Figure 5-3b allows the reader to evaluate the difference between replicates from each sample for each operator. One message from Figure 5-3b is that some operators came closer than others in replicating their own measurement results. The results from the R&R study are summarized in Table 5-1 for the test method TP 98 (automobiles); complete results are included in Appendix C.

Table 5-1. Test Method TP 98 (Automobiles) Precision Results

Unit of Analysis	SIPI (dBA)
	0.950
R&R Precision	0.862
90% Confidence Interval	0.740 – 1.22
95% Confidence Interval	0.722 – 1.350
Percent R&R Value	90.831%
Signal-to-Noise Ratio	0.651
90% Confidence Interval	0 – 2.319
95% Confidence Interval	0 – 2.973
Reproducibility	
Std. dev. (σ_{R})	0.476
$R = 2 \times \sigma_{\mathrm{R}}$	0.952
$R = 2.8 \times \sigma_{\mathrm{R}}$	1.332
Repeatability	
Std. dev. (σ_{r})	0.719
$r = 2 \times \sigma_{\mathrm{r}}$	1.439
$r = 2.8 \times \sigma_{\mathrm{r}}$	2.014

The results from the R&R study for the Test Method TP 98 (Automobiles) indicate that the total sample's standard deviation is 0.9495 dBA. The reproducibility component's percent standard deviation is 0.4758 dBA and the repeatability component's is 0.7193 dBA. The expected R&R precision level (i.e., the standard deviation of the R&R component) is 0.8624 dBA; its actual precision ranges between 0.7397 and 1.2196 dBA at the 90 percent confidence level and between 0.7220 to 1.3496 dBA at the 95 percent confidence level.

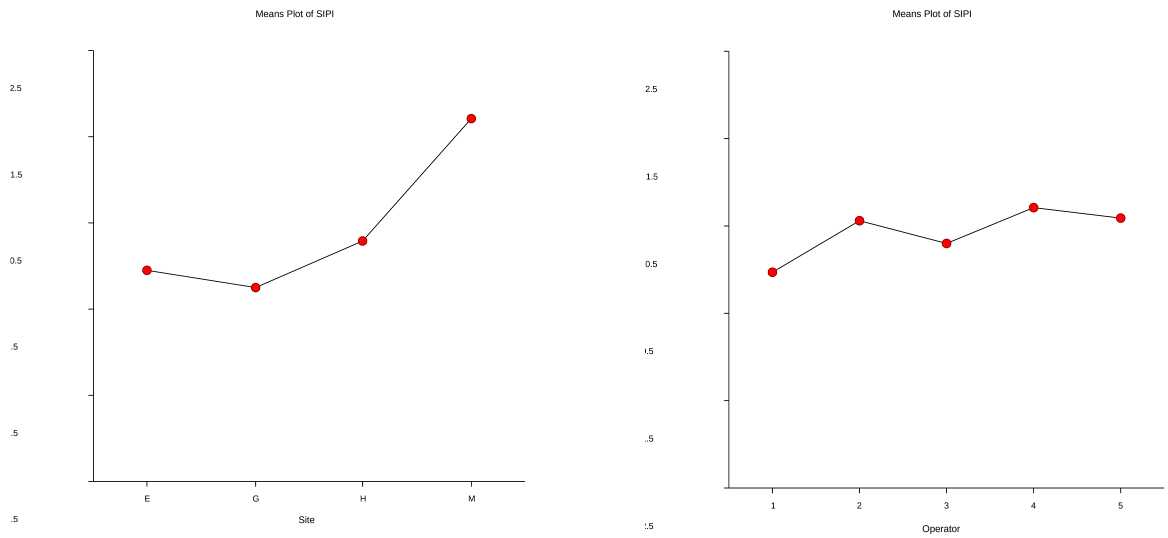
The R&R component's percent of the total standard deviation is 91 percent, which is significantly higher than the criterion of a maximum of 30 percent of total the standard deviation. This suggests that the process is not precise.

The SNR based on the study's results is 0.6513, which does not meet the minimum SNR criterion of 3. The 90 percent confidence interval of the SNR ranges from 0 to 2.3185, and the 95 percent confidence interval is 0 to 2.9726. Since even the upper limit of the 90 percent confidence interval for the SNR is less than 3, this suggests that process is not adequate and hence not precise.

The reason for the R&R component's high percentage is due to the measurement protocols that define the standard test method TP 98. The test method TP 98's specifications when followed do not allow for precise measurements by trained operators. This method may not have the required precision to compare noise measurements among different road samples. While the team has produced precision values for the reproducibility and the repeatability of a measurement process, the high level of variability which enters the process due to the combined operator and sample effect warrants extreme care when referring to these precision guidelines. Future measurements may fall into the reproducibility and repeatability ranges; however, these outcomes may not guarantee that the measurements were conducted precisely. Because of the imprecision of the methodology with respect to measuring noise from automobiles, differences in the SIPI measurements cannot be fully attributed to differences in the tire-pavement noise of the road surface and vehicle fleet at the time of measurement.

Statistical Results of the R&R Study for TP 98 (Heavy Trucks)

Figure 5-4 (a) and (b) show the variability in the average SIPI measurements from the perspective of the sample sites and the operators, respectively. Appendix C contains the operators' results as well as average values by sample site and by operators. The range in the average SIPI measurements across the sites (range of 1.96 dBA) is slightly larger than that observed for the range of the average SIPI measurements across the operators (range of 0.740 dBA). This result may suggest a more precise measuring process for measuring traffic sound levels from heavy trucks than that found for automobiles.

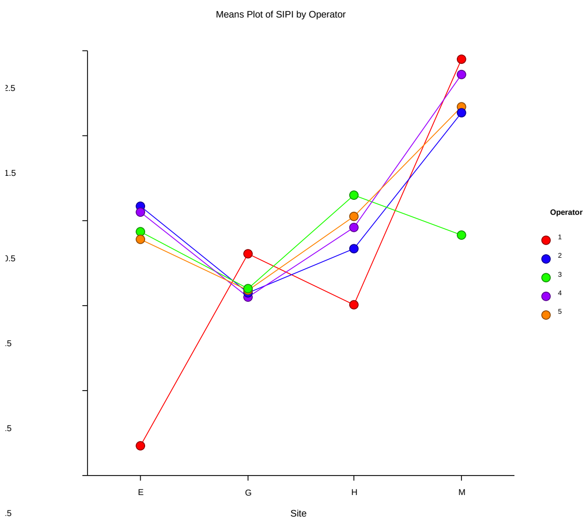


a. Plot of the Average SIPI Measurements per Site

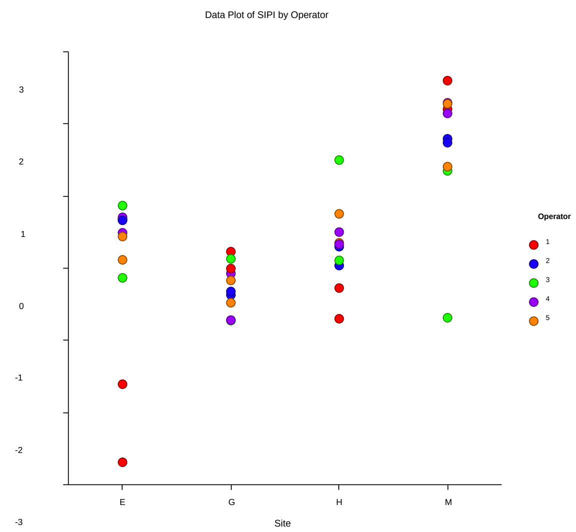
a. Plot of the Average SIPI Measurements per Operator

Figure 5-4. Test Method TP 98 (Heavy Trucks) Plots of Average SIPI per Site and per Operator

The variability patterns across operators and sites as shown in Figure 5-5 demonstrate an improvement relative to that observed from the test method for automobiles (see Figure 5-3); however, at least one operator per site has visibly different measurements than the majority of the other operators. For example, operator 1's mean of the SIPI replicates from site E is significantly lower than the other means while operator 3's mean of the SIPI replicates is the lowest of all the means collected from site M.



a. Plot of Mean SIPI Measurements per Operator per Sample Site



b. Plot of Individual SIPI Measurements per Sample Site

Figure 5-5. Test Method TP 98 (Heavy Trucks) Variability Trends across Sites and Operators

The results from the R&R study are summarized in Table 5-2 for the test method TP 98 (heavy trucks); complete results are included in Appendix C.

Table 5-2. Test Methods TP 98 (Heavy Trucks) Precision Results

Unit of Analysis	SIPI (dBA)
Total Standard Deviation	1.174
R&R Standard Deviation	0.853
90% Confidence Interval	0.707 – 1.225
95% Confidence Interval	0.686 – 1.346
Percent R&R Value	72.634%
Signal-to-Noise Ratio	1.338
90% Confidence Interval	0.544 – 4.182
95% Confidence Interval	0.378 – 5.342
Reproducibility	
Std. dev. (σ_R)	0.671
$R = 2 \times \sigma_R$	1.342
$R = 2.8 \times \sigma_R$	1.879
Repeatability	
Std. dev. (σ_r)	0.526
$r = 2 \times \sigma_r$	1.053
$r = 2.8 \times \sigma_r$	1.474

The results from the R&R study for the Test Method TP 98 (Heavy Trucks) indicate the total sample's standard deviation is 1.1741 dBA. The reproducibility component's standard deviation is 0.6709 dBA and the repeatability component's is 0.5264 dBA. The expected R&R precision level (i.e., the standard deviation of the R&R component) is 0.8528 dBA; its actual precision ranges between 0.7073 and 1.2254 dBA at the 90 percent confidence level and between 0.6860 to 1.3455 dBA at the 95 percent confidence level.

The R&R component's percent of the total standard deviation is 73 percent, which is significantly higher than the criterion of a maximum of 30 percent of total standard deviation. These R&R results suggest that the test method TP 98 measurement of heavy trucks is not precise.

The SNR from the R&R results is 1.3383 which does not meet the minimum SNR criterion of 3. The 90 percent confidence interval ranges from 0.540 to 4.1824, and the 95 percent confidence interval ranges from 0.3781 to 5.3420. These SNR results suggest that this process cannot distinguish across the samples and hence is not precise. Of note is that TP 98 Heavy Trucks' percent R&R value is better (meaning closer to the accepted maximum of 30 percent) than that observed for TP 98 Automobiles (91 percent).

In summary, this method may not have the required precision to compare noise measurements among different samples. While the team has produced precision values for the reproducibility (R) and the repeatability (r) of a measurement process, the high level of variability which enters the process due to the combined operator and sample effect warrants extreme care when referring to these precision guidelines. Future measurements may fall into the ranges provided by r and R; however, this outcome may not guarantee that the measurements were conducted precisely. Because of the imprecision of the methodology with respect to measuring noise from heavy trucks, differences in SIPI measurements cannot be fully attributed to differences in tire-pavement noise of the road surface and vehicle fleet at the time of measurement.

Statistical Analysis of Test Method TP 99 Repeatability and Reproducibility

As with the R&R study design for Test Method TP 98, the R&R study design for Test Method TP 99 is that each of the four samples is processed by five different operators and each operator processes each sample twice, yielding a total sample size of 40 tests. Three of the sample sites were reused from the R&R study design of Test Method TP 98, with the addition of one other sample site for the TP 99 R&R study. The test measurement outcome is referred to as the average normalized L_{eq}. For the purpose of this report and to simplify the statistical language, the TP 99 test result will be referred to as the CTIM measurement.

Figure 5-6 shows a histogram of the CTIM measurements for TP 99. The average of the CTIM measurements across the 40 tests is 72.94 dBA with the lowest value being 69.88 dBA and the highest value being 77.40 dBA.

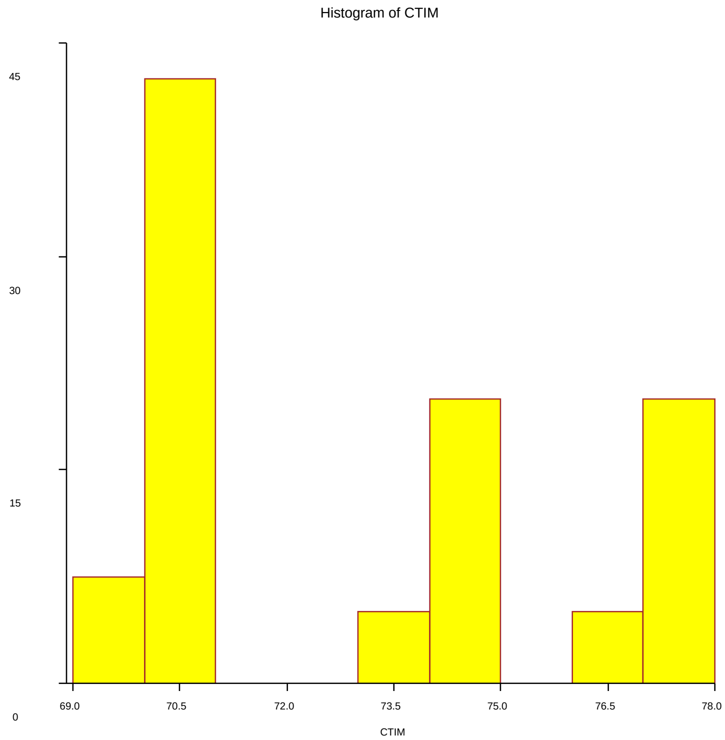
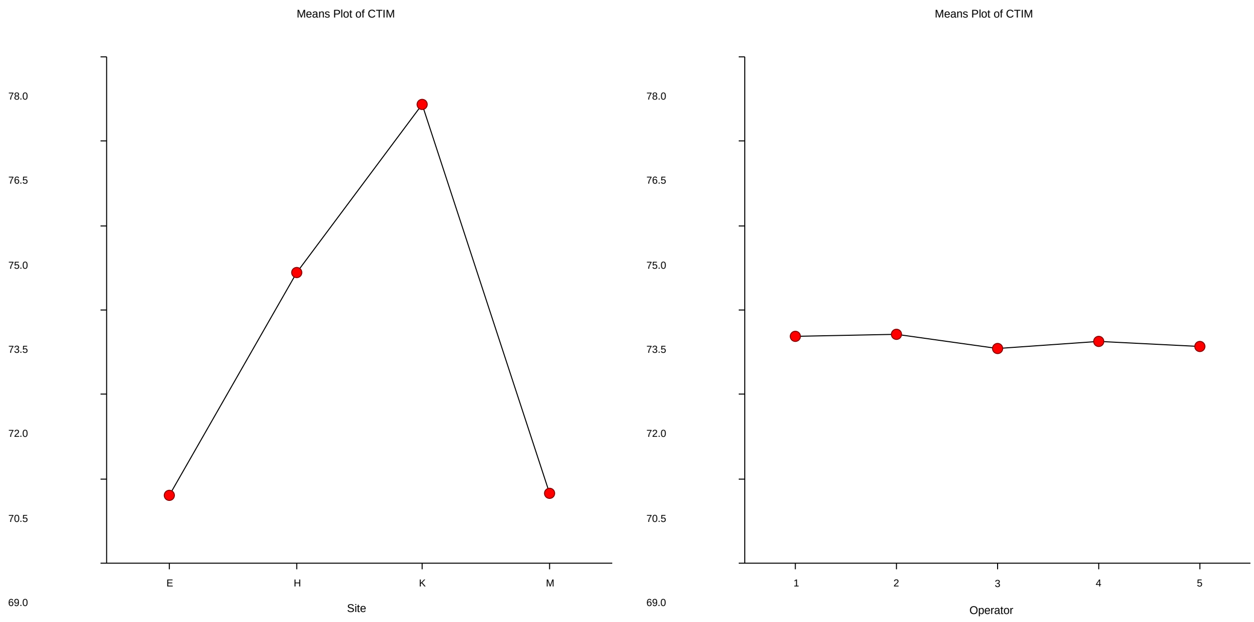


Figure 5-6. Test Method TP 99 Histogram of CTIM Measurements

Statistical Results of the R&R Study for TP 99

Plots of the average of the CTIM measurements per sample site and operator show that the range of the average values per sample site in Figure 5-7(a) is dramatically larger than range of the average of the CTIM measurements by operator in Figure 5-7(b). The averages by sample site and by operator are also shown in Appendix C. The range from Figure 5-7(a) is 6.9 dBA, which is 27 times larger than the range of 0.3 dBA in Figure 5-7(b). This very large differential is highly suggestive that the measurement process is precise as the operators consistently produce similar measurements regardless of the sample site. The R&R analysis of the actual data can confirm if this finding from the graphical analysis supports the assumption that this test method is precise.

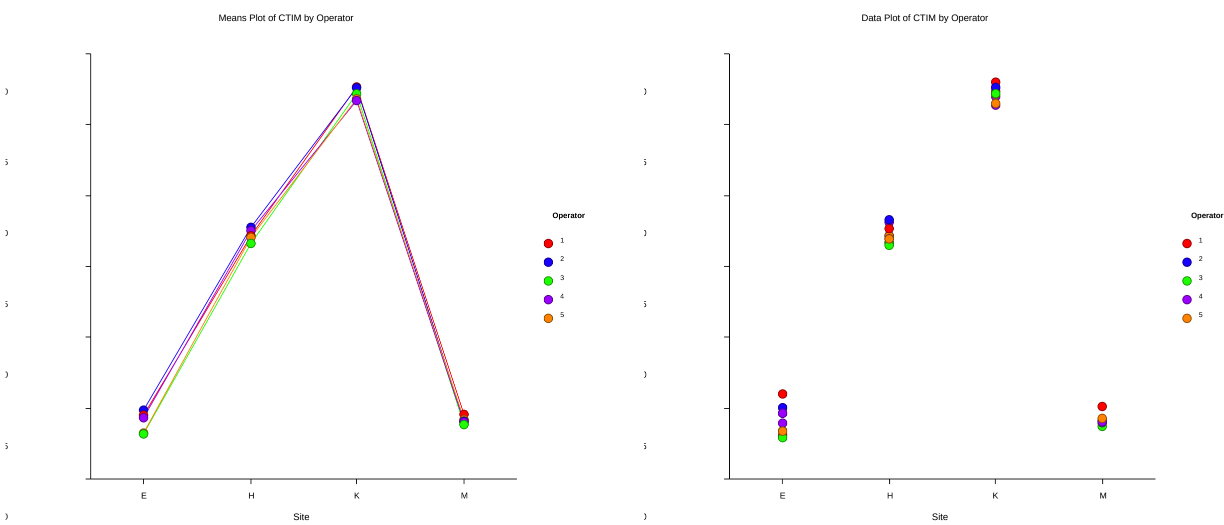


a. Plot of the Average CTIM Measurements per Site

b. Plot of the Average CTIM Measurements per Operator

Figure 5-7. Test Method TP 99 Plots of Average CTIM Measurements per Site and per Operator

Figure 5-8 demonstrates the level of variability across the measurements by operator per sample site. Unlike the comparable plots from the R&R study for test methods TP 98 automobile and heavy trucks, these plots show that all operators using the same measurement process are highly consistent in their final measurements, regardless of whether the measurement is a replicate or a different sample.



a. Plot of Average CTIM Measurements per Operator per Site

b. Plot of Average CTIM Measurements per Site

Figure 5-8. Test Method TP 99 Variability Trends across Sites and Operators

The results from the R&R study are summarized in Table 5-3 for the test method TP 98 (heavy trucks); complete results are included in Appendix C.

Table 5-3. Test Method TP 99 Precision Results Summary

Unit of Analysis	Average normalized L_{eq} (dBA)
Total Standard Deviation	3.377
R&R Standard Deviation	0.205
90% Confidence Interval	0.176 – 0.324
95% Confidence Interval	0.172 – 0.374
R&R % Standard Deviation	6.062%
Signal-to-Noise Ratio	23.287
90% Confidence Interval	12.741 – 66.258
95% Confidence Interval	10.960 – 84.139
Reproducibility	
Std. dev. (σ_{R})	0.071
$R = 2 \times \sigma_{\mathrm{R}}$	0.143
$R = 2.8 \times \sigma_{\mathrm{R}}$	0.200
Repeatability	
Std. dev. (σ_{r})	0.192
$r = 2 \times \sigma_{\mathrm{r}}$	0.384
$r = 2.8 \times \sigma_{\mathrm{r}}$	0.537

The results from the R&R study for the Test Method TP 99 indicate the total sample's standard deviation is 3.3767 dBA. The reproducibility component's standard deviation is 0.0713 dBA and the repeatability component's is 0.1919 dBA. The expected R&R precision level (i.e., the standard deviation of R&R component) precision is 0.2047 dBA; its actual precision ranges between 0.1757 and 0.3243 dBA at the 90 percent confidence level and between 0.1718 to 0.3735 dBA at the 95 percent confidence level.

The R&R component's percent of the total standard deviation is 6 percent, which is significantly lower than the criterion of a maximum of 30 percent of total standard deviation for a measurement process to be considered adequate. The overwhelming majority of the variation in the overall sample can be explained by differences in the properties of the samples across the sites. These differences constitute nearly 100 percent of the overall standard deviation. The SNR from the R&R results is a healthy index of 23.2865, which is significantly larger than the minimum cut-off of 3. The 90 percent confidence interval for the SNR ranges from 12.7411 to 66.2589, and the 95 percent confidence interval ranges from 10.9595 to 84.1385. These R&R results suggest that the test method TP 99 is an excellent measurement process and can be relied upon to produce precise measurements.

The protocols contained within the Test Method TP 99 produce consistent values for the same samples regardless of whether they are measured by the same operator or different operators. The R and r precision values in Table 5-3 can be used as precision guides by other laboratories if they conduct tests under similar reproducibility and repeatability conditions to those described in this report. This method, due to its high level of precision, can be used to compare the traffic sound levels emitted across different road surfaces and to determine if the differences in the sound levels can be attributed to tire-pavement noise of the road surface and vehicle fleet at the time of measurement.

Statistical Analysis of Bias Study Results

The approach to measuring bias is a relatively straightforward task provided a known reference value is available. The average of the differences between each measurement and its appropriate reference value represents the bias. The bias value indicates how well the measurements compare to the reference value. A positive bias indicates that the measurement process overestimates. A negative bias indicates that the measurement process underestimates. As explained earlier (see Chapter 3), there are no known reference values for the SIPI and CTIM measurements from the test methods TP 98 and TP 99, respectively, on which to base a bias statement. In lieu of this shortcoming, the present research estimated a component of bias (see Chapter 4 of this report for the measurement details).

Data Structure for Bias Study

The targeted measurement from this study is the RSS recorded in units of A-weighted decibels. The sound recordings were taken at the same microphone positions used for TP 98 and TP 99: 25 feet and 50 feet from the nearest lane of travel. These measurements were performed during the R&R site recordings at the hourly recording breaks, at the same time as the calibration checks. Measurements were taken at sites E, G, H, K and M. Depending on the site and microphone location, one to four separate measurements were made. The calibrated values of the RSS assigned to the 25- and 50-foot distances are 68.531 dBA and 61.217 dBA, respectively. These values represent the fixed reference values for the purpose of estimating a component of bias. In fact, the reference value is actually a predicted value based upon precision laboratory measurements and established acoustic principals. The precision level for these measurements is extremely small or 0.5 dBA, a value which the human ear cannot discern. Given the calibration's small value and that it would add only a very small increment to the confidence interval around the average bias, the research team considers this value as fixed in every practical sense. The experimental outcomes from the 25- and 50-feet settings are recorded in Table 5-4 and Table 5-5 below. Bias is equal to the field RSS measurement minus the predicted RSS measurement.

Table 5-4. Bias Dataset for RSS Measurements take at 25 feet

Site	Field RSS Measurement	Predicted RSS Measurement	Bias
E	69.670	68.531	1.139
E	70.070	68.531	1.539
E	69.544	68.531	1.013
E	70.364	68.531	1.833
G	67.858	68.531	-0.673
G	68.149	68.531	-0.382
G	68.129	68.531	-0.402
G	68.490	68.531	-0.042
H	65.240	68.531	-3.291
H	65.124	68.531	-3.408
K	71.196	68.531	2.665
K	70.712	68.531	2.180
K	67.869	68.531	-0.663
M	68.964	68.531	0.433
M	69.070	68.531	0.539

Table 5-5. Bias Dataset for RSS Measurements take at 50 feet

Site	Field RSS Measurement	Predicted RSS Measurement	Bias
E	61.858	61.217	0.642
E	62.314	61.217	1.097
E	61.482	61.217	0.266
E	61.629	61.217	0.412
G	61.412	61.217	0.195
G	61.358	61.217	0.141
G	61.090	61.217	-0.127
G	61.528	61.217	0.311
H	60.690	61.217	-0.527
H	59.265	61.217	-1.951
K	67.573	61.217	6.356
K	68.836	61.217	7.619
K	62.477	61.217	1.260
M	61.183	61.217	-0.033
M	61.767	61.217	0.551

Statistical Results for the Bias Study

The bias in the RSS measurements under the two different distance scenarios is positive. The average bias when the RSS is measured from a distance of 25 feet is 0.165 dBA and the average bias is 1.081 dBA as shown in Table 5-6. Overall, the sample mean value is 0.623 dBA. The median values per distance scenario and overall are also positive. The table also shows 95 percent confidence intervals for the averages. All intervals contain the value zero which suggests that the observed average for each of the distance scenarios and overall is not different from zero at the 0.05 level of statistical significance. In other words, the bias effect is negligible and constant over the two types of distance scenarios.

Table 5-6. Statistical Summary of the RSS Bias by Distance

Distance	Sample Size	Mean	Median	Std Dev	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
25ft	15	0.165	0.433	1.765	-.812	1.143
50ft	15	1.081	0.311	2.520	-.315	2.477
Total	30	0.623	0.362	2.188	-.194	1.440

Figure 5-9 demonstrates the box-plots of the bias values by distance. A box-plot is useful for visually studying whether the observations across the different groups are similar or not. If there are no outliers in the sample, then the line below the box represents the minimum observation and the line above the box represents the maximum observation. However, if there are outliers on either extremes of the sample, then the lines represent outlier thresholds. The outlier values are displayed depending on their values either below the bottom line or above the top line. Often these lines are referred to as ‘whiskers’. The lower edge of the box, the middle of the box and the top of the box represent the 25th, 50th (median) and 75th percentiles of the observations within that group. While there are 3 outliers (-1.951, 6.356, 7.619) from the sample of bias values from the RSS measurements taken at 50 feet, overall, the distribution of the bias values are not different from those collected from RSS measurements taken at 25 feet.

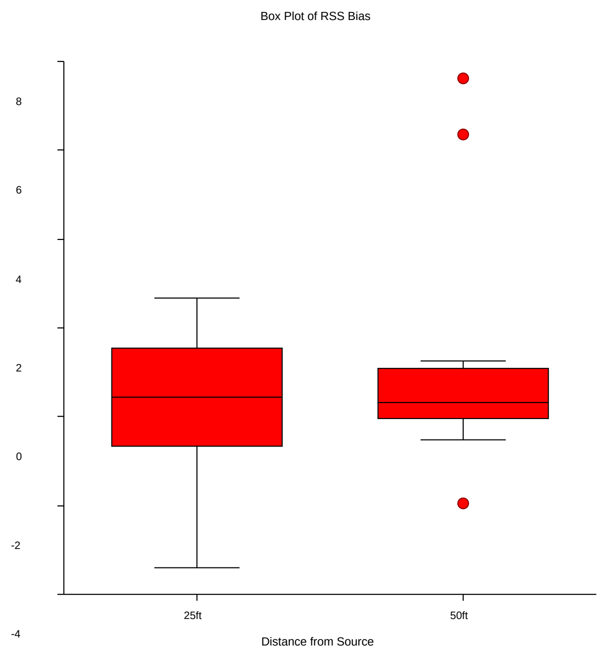


Figure 5-9. Comparison of the RSS Bias Amounts by Distance

Bias Statement

Because the 95 percent confidence interval for the true bias contains zero, we conclude that there is no measurable bias. The interval tells us that the bias for the SIPI and CTIM measurements can range from -0.194 dBA to 1.440 dBA at the 95 percent level of confidence. Overall, the impact on the accuracy of the SIPI and CTIM measurements related to this component of bias is negligible.

Influential Factors Study Results

For each of the standard test methods TP 98 (automobiles and heavy trucks) and TP 99, the research team ran a full model regression analysis, produced p-values for its coefficients, and produced the AICc value for each run of the forward selection method. The results of such tests per test method are described in the following three sections. The complete output from the statistical analysis of the data from the ruggedness tests for standard test methods TP 98 (automobile and heavy trucks) and TP 99 can be found in Appendix C of this report.

Ruggedness Test Results for Test Method TP 98 (Automobiles)

As described in the approach, each measurement was sited and scheduled to fit the specific criteria for the prescribed runs of the ruggedness test, following a Plackett-Burman design. The purpose of this screening-level experimental design is to determine whether the measurement factors influence the measurement results, including the different characteristics of the sites, the environment, and the measurement options. Each run of the Plackett-Burman design for test method TP 98 generated eight different SIPI values, as shown in Table 5-7. The average SIPI is 0.111 dBA with a minimum observed value of -6.209 dBA from run 5 to a maximum observed value of 5.993 dBA in run 6. By averaging the SIPI measurements separately for the levels per factor, one can have a preliminary estimate of the impact that the differing levels have on the SIPI measurements. For example, the differences of the two averages between newer and older aged pavement surfaces yields a value of 4.295 (2.258 – (-2.037) = 4.295). The positive value suggests that older pavement generates higher traffic sound levels than newer pavement. However, without statistically validating if this observed difference is meaningful, one cannot say that it is an influencing factor or not.

Table 5-7. Test Method TP 98 (Automobiles) Ruggedness SIPI Measurements

	Factor A Age	Factor B Ground	Factor C Speeds	Factor D Mic Position	Factor E Air Temp.	SIPI
Run 1	Older (+1)	Softer (+1)	Interstate (+1)	25 ft (-1)	Cooler (+1)	-1.911
Run 2	Newer (-1)	Softer (+1)	Interstate (+1)	50 ft (+1)	Warmer (-1)	2.219
Run 3	Newer (-1)	Harder (-1)	Interstate (+1)	50 ft (+1)	Cooler (+1)	-2.799
Run 4	Older (+1)	Harder (-1)	Highway (-1)	50 ft (+1)	Cooler (+1)	1.375
Run 5	Newer (-1)	Softer (+1)	Highway (-1)	25 ft (-1)	Cooler (+1)	-6.209
Run 6	Older (+1)	Harder (-1)	Interstate (+1)	25 ft (-1)	Warmer (-1)	5.993
Run 7	Older (+1)	Softer (+1)	Highway (-1)	50 ft (+1)	Warmer (-1)	3.576
Run 8	Newer (-1)	Harder (-1)	Highway (-1)	25 ft (-1)	Warmer (-1)	-1.360
Avg +	2.258	-0.581	0.876	1.093	-2.386	
Avg -	-2.037	0.802	-0.654	-0.872	2.607	
Diff	4.295	-1.383	1.530	1.964	-4.993	

Statistical Results of the Ruggedness Test for Test Method TP 98 (Automobiles)

Of the five factors tested for statistical significance using regression analysis of all five factors, none have a p-value that is less than 0.01. Thus, we conclude that there is insufficient evidence that any of the factors have an impact on the final SIPI measurements. This means that two otherwise identical measurements are unlikely to exhibit much difference even if they differ in hard versus soft ground, highway versus interstate speeds, etc. (outside of the variation due to repeatability and reproducibility). We note that the p-value for air temperature, at 0.041, is marginally (or potentially) significant. If so, the impact of this effect is that as the temperature moves from warmer to cooler, the SIPI measurement decreases in value, holding all other factors constant.

Table 5-8. Test Method TP 98 (Automobiles) Ruggedness Regression Output

Coefficient Name	Coefficient Value	P-Value
Constant	0.111	0.851
Pavement Age	2.148	0.054
Ground Impedance	-0.692	0.314
Speed Regime	0.765	0.279
Measurement Option	0.982	0.199
Air Temperature	-2.496	0.041

Table 5-9 below shows the results of assessing the relationship between the factors and the SIPI measurement using Stepwise/AICc. The order in which the factors were entered is based on the resulting value of AICc. For example, the one-term model that leads to the lowest value of AICc is the model having only the air temperature factor. Given that air temperature is in the model, the two-term regression model that has the lowest value of AICc comprises air temperature and pavement age. Continuing in this fashion, the best three-, four-, and five-term models occur by adding the measurement option, speed regime, and ground impedance, respectively. However, the model that produced the lowest AICc value was the null model, the model that results when all of the factors are removed from the model. This implies that none of the factors have a statistically significant impact on the values of the SIPI measurements. This is consistent with the p-value analysis of the full model previously described.

Table 5-9. Test Method TP 98 (Automobiles) Ruggedness Stepwise/AICc Output

Step	Coefficient Name	P-Value	AICc
1	Air Temperature	0.063	50.475
2	Pavement Age	0.030	51.582
3	Measurement Option	0.196	66.480
4	Speed Regime	0.281	118.847
5	Ground Impedance	0.314	---
Best Model	Model with No Factors		49.873

Ruggedness Test Results for Test Method TP 98 (Heavy Trucks)

The average SIPI measurement from the test method TP 98 for heavy trucks is -2.138 dBA ranging from a low of -5.919 in run 5 to a high of 2.257 dBA in run 6 (See Table 5-10). Of interest is that these are the same runs from the automobile testing which generated that series' minimum and maximum, respectively. The signs of the differences between the average positive level SIPI and the negative level SIPI follow that observed from the automobile testing. For example, the difference of 0.690 dBA between the average positive SIPI and the average negative SIPI for pavement age is positive, albeit smaller than the equivalent difference from the automobile testing (4.295 dBA). The factor with the largest directional difference relative to the differences observed for the other factors is air temperature. The difference between the average of the SIPI collected under cooler temperatures and the average taken under warmer temperatures is -3.837 dBA. The negative value suggests that measurements taken during cooler temperatures have lower traffic sound levels than measurements taken during warmer temperatures. This screening-level experimental design is unable to offer statistical validation that the observed difference is meaningful; however, there is a plentiful body of research which indicates that temperature does affect tire-pavement noise.

Table 5-10. Test Method TP 98 (Heavy Trucks) Ruggedness SIPI Measurements

	Factor A Age	Factor B Ground	Factor C Speeds	Factor D Mic Position	Factor E Air Temp.	SIPI
Run 1	Older (+1)	Softer (+1)	Interstate (+1)	Shorter (-1)	Cooler (+1)	-4.680
Run 2	Newer (-1)	Softer (+1)	Interstate (+1)	25 ft (-1)	Warmer (-1)	-0.564
Run 3	Newer (-1)	Harder (-1)	Interstate (+1)	50 ft (+1)	Cooler (+1)	-2.224
Run 4	Older (+1)	Harder (-1)	Highway (-1)	50 ft (+1)	Cooler (+1)	-3.403
Run 5	Newer (-1)	Softer (+1)	Highway (-1)	50 ft (+1)	Cooler (+1)	-5.919
Run 6	Older (+1)	Harder (-1)	Interstate (+1)	25 ft (-1)	Warmer (-1)	2.257
Run 7	Older (+1)	Softer (+1)	Highway (-1)	25 ft (-1)	Warmer (-1)	-1.345
Run 8	Newer (-1)	Harder (-1)	Highway (-1)	50 ft (+1)	Warmer (-1)	-1.226
Avg +	-1.793	-3.127	-1.303	-1.884	-4.056	
Avg -	-2.483	-1.149	-2.973	-2.392	-0.220	
Diff	0.690	-1.978	1.670	0.508	-3.837	

Statistical Results of the Ruggedness Test for Test Method TP 98 (Heavy Trucks)

The results of the ruggedness test on the SIPI measurements from measuring heavy truck sound levels are similar to those found from the automobile ruggedness test. Based on the regression analysis of all five factors summarized in Table 5-11, none have a p-value that is less than 0.01. Thus, we again conclude that there is insufficient evidence to conclude that any of the factors have an impact on the final SIPI measurements. Interestingly, the coefficient for the air temperature factor is again nearly (or marginally) significant, having a p-value of 0.02. The direction of its impact mirrors that from the automobile results. None of the other factors showed any evidence of influencing the SIPI values.

Table 5-11. Test Method TP 98 (Heavy Trucks) Ruggedness Regression Output

Coefficient Name	Coefficient Value	P-Value
Constant	-2.138	0.017
Pavement Age	0.345	0.349
Ground Impedance	-0.989	0.074
Speed Regime	0.835	0.099
Measurement Option	0.254	0.467
Air Temperature	-1.918	0.022

The outcome from the Stepwise/AICc selection method did identify that air temperature does influence the SIPI measurements in an inverse fashion. In other words, as temperature goes from warmer to cooler, the SIPI values decline. The p-value from Table 5-12 for this factor is substantially lower than the p-value from the same factor measured in the automobile ruggedness test. There is enough evidence in the truck data based on the Stepwise/AICc selection method to confirm that the impact of air temperature on the SIPI measurements is statistically significant.

Table 5-12. Test Method TP 98 (Heavy Trucks) Ruggedness Stepwise/AICc Output

Step	Coefficient Name	P-Value	AICc
1	Air Temperature	0.016	40.336
2	Ground Impedance	0.083	44.378
3	Speed Regime	0.047	54.216
4	Pavement Age	0.298	106.836
5	Measurement Option	0.467	---
Best Model	Model with Air Temperature		40.336

While the forward selection method regression of the outcomes from the ruggedness test for the automobile test method TP 98 did not pick up statistical significance of any of the five factors, the fact that the same method did pick up statistical significance of the air temperature factor from the ruggedness test for the heavy truck test method TP 98 is suggestive that air temperature may be an influential factor for the overall test method TP 98. The contrary outcomes for the same test method TP 98, albeit for different vehicle classes, does support further research into the influence of air temperature on SIPI measurements.

Ruggedness Test Results for Test Method TP 99

As described in the approach, each measurement was sited and scheduled to fit the specific criteria for the prescribed runs of the ruggedness test, following a Plackett-Burman design. The purpose of this screening-level experimental design is to determine whether the measurement factors influence the measurement results, including the different characteristics of the sites, the environment, and the measurement options. The average CTIM measurement from the five factors, eight run Plackett-Burman design is 70.978 dBA (See Table 5-13). The minimum value of 64.394 came from run 5 while the maximum value of 77.300 came from run 6. As with both vehicle classes from the test method TP 98, these are the same runs which captured the minimum and maximum values, respectively. The differences between the group averages across the five factors are most prominent for pavement age, ground impedance, speed regime, and air temperature. The negative difference value from factor B, ground impedance, indicates sound levels are greater where microphones are over harder ground and sound levels are quieter over softer ground. The positive difference value from factor C, speed regime, suggests that sound levels from interstate roads are greater than sound levels from highways roads. To prove that these differences have an impact on CTIM measurements in a predictable manner, one has to conduct the appropriate statistical tests as shown in the following section.

Table 5-13. Test Method TP 99 Ruggedness CTIM Measurements

	Factor A Age	Factor B Ground	Factor C Speeds	Factor D Meas. Block	Factor E Air Temp.	SIPI
Run 1	Older (+1)	Softer (+1)	Interstate (+1)	5 min (-1)	Cooler (+1)	70.047
Run 2	Newer (-1)	Softer (+1)	Interstate (+1)	15 min (+1)	Warmer (-1)	71.614
Run 3	Newer (-1)	Harder (-1)	Interstate (+1)	15 min (+1)	Cooler (+1)	71.188
Run 4	Older (+1)	Harder (-1)	Highway (-1)	15 min (+1)	Cooler (+1)	70.044
Run 5	Newer (-1)	Softer (+1)	Highway (-1)	5 min (-1)	Cooler (+1)	64.394
Run 6	Older (+1)	Harder (-1)	Interstate (+1)	5 min (-1)	Warmer (-1)	77.300
Run 7	Older (+1)	Softer (+1)	Highway (-1)	15 min (+1)	Warmer (-1)	72.257
Run 8	Newer (-1)	Harder (-1)	Highway (-1)	5 min (-1)	Warmer (-1)	70.431
Avg +	72.412	69.578	72.537	71.276	68.918	
Avg -	69.407	72.241	69.282	70.543	72.901	
Diff	3.005	-2.663	3.255	0.733	-3.982	

Statistical Results of the Ruggedness Test for Test Method TP 99

The p-values for the regression coefficients in the ruggedness test for Test Method TP 99 are shown in Table 5-14. Again, since none of the p-values are less than 0.01, there is insufficient evidence to conclude that any of the factors have an impact on the final CTIM measurements. This means that two otherwise identical measurements, even if they differ in hard versus soft ground, highway versus interstate speeds, etc., are unlikely to exhibit much difference (outside of the variation due to repeatability and reproducibility).

Several factors appear to be potentially (or marginally) significant based on the p-values from the regression model results in Table 5-14. They are, ranked in the order of increasing p-values, air temperature, speed regime, pavement age and ground impedance. For example, the p-value for air temperature is 0.018, which is very close to the 0.01 threshold. As in the prior analyses, the sign of the regression coefficient suggests that sound levels in warmer air are greater than sound levels in colder air. The p-values for speed regime, pavement age, and ground impedance also close to the 0.01 threshold at 0.026, 0.031, and 0.039, respectively. Thus the data give some indication that (1) highway-speed sound levels are less than freeway-speed sound levels; (2) older pavements are noisier than newer pavements, and (3) sound levels at the microphone are greater over hard ground versus soft ground.

Table 5-14. Test Method TP 99 Ruggedness Regression Output

Coefficient Name	Coefficient Value	P-Value
Constant	70.909	0.000
Pavement Age	1.503	0.031
Ground Impedance	-1.331	0.039
Speed Regime	1.628	0.026
Measurement Option	0.366	0.308
Air Temperature	-1.991	0.018

When the Stepwise/AICc selection method is used to determine the most influential factors, none of the factors had explanatory power to model the variation in the CTIM measurements (see Table 5-15). This is in complete agreement with the p-value analysis described above, using the 0.01 (Bonferroni) threshold; this threshold was discussed in Chapter 3.

Table 5-15. Test Method TP 99 Ruggedness Stepwise/AICc Output

Step	Coefficient Name	P-Value	AICc
1	Air Temperature	0.114	50.225
2	Speed Regime	0.140	55.725
3	Pavement Age	0.104	68.457
4	Ground Impedance	0.022	108.526
5	Measurement Option	0.308	.
Best Model	Model with No Factors		48.231

In summary, the statistical analysis of the ruggedness test data using the forward selection method regression with the AICc shows that the five factors do not influence the CTIM measurements.

CHAPTER 6

Findings and Recommendations

Precision and Bias Findings

The R&R Analysis of test method TP 98 for both vehicle classes of automobiles and heavy trucks determined that this method is not precise since (1) the percent R&R value is greater than 30 percent and (2) the upper bound of the SNR's confidence interval is less than 3. This finding suggests that the measurement process needs reexamination to increase the precision of the process.

With respect to test method TP 99, the R&R analysis did find that this test method is precise since (1) the R&R value is substantially smaller than 30 percent and (2) the lower bound of the SNR's value is greater than 3. This method is precise and researchers can have confidence in the precision of their measurements provided they follow the protocols for test method TP 99.

The present assessment of the component of bias related to the acoustic environment in which sound recordings are conducted according to test methods TP 98 and TP 99 showed that the amount of bias is negligible. Table 6-1 summarizes the results.

Table 6-1. Precision Results for Test Methods TP 98 and TP 99

Test Method	TP 98 (Automobiles)	TP 98 (Heavy Trucks)	TP 99
Unit of Analysis	SIPI (dBA)	SIPI (dBA)	average normalized L_{eq} (dBA)
R&R Precision	0.862	0.853	0.205
90% Confidence Interval	0.740 – 1.22	0.707 – 1.225	0.176 – 0.324
95% Confidence Interval	0.722 – 1.350	0.686 – 1.346	0.172 – 0.374
Percent R&R Value	90.831	72.634	6.062
Signal-to-Noise Ratio	0.651	1.338	23.287
90% Confidence Interval	0 – 2.319	0.544 – 4.182	12.741 – 66.258
95% Confidence Interval	0 – 2.973	0.378 – 5.342	10.960 – 84.139
Reproducibility			
Std. dev. (σ_R)	0.476	0.671	0.071
$R = 2 \times \sigma_R$	0.952	1.342	0.143
$R = 2.8 \times \sigma_R$	1.332	1.879	0.200
Repeatability			
Std. dev. (σ_r)	0.719	0.526	0.192
$r = 2 \times \sigma_r$	1.439	1.053	0.384
$r = 2.8 \times \sigma_r$	2.014	1.474	0.537
Precision Evaluation	Not Acceptable	Not Acceptable	Excellent

According to the scope and purpose of TP 98, it is intended to be broadly applied and used to compare one pavement to another, either at different roadways or at different times in the lifespan of one roadway. However this research indicates that the precision of TP 98 doesn't necessarily allow pavements to be compared, unless the difference is substantial. There may be ways to improve the precision of the measurement process, including, most likely, larger sample sizes and perhaps multiple sampling periods in different conditions.

In contrast, the scope and purpose of TP 99 is intended to be narrowly applied to a single measurement site and compare different pavements or different times in the lifespan of one pavement. The precision of TP 99 is excellent, and users of this measurement process can be confident that differences in measurements reflect differences in sound levels. Variations in traffic, which could affect sound levels independently of effects from tire-pavement noise, are implicit in the measurement process. Speculating on all the components of this measurement procedure, the modeling and normalization along with the outlier elimination process are likely contributors to the excellent precision of the TP 99 measurement process. Identifying these particular factors in the measurement process is outside the scope of the research.

Statement of Precision and Bias for TP98

The findings from the present research support the suggested precision and bias statement below for the TP 98 test method.

Precision

Automobile Vehicle Category:	
95% repeatability limit (test-to-test, same operator)	2.0140 dBA
95% reproducibility limit (operator-to-operator)	1.3322 dBA
Heavy Truck Vehicle Category:	
95% repeatability limit (test-to-test, same operator)	1.4739 dBA
95% reproducibility limit (operator-to-operator)	1.8785 dBA

The values shown above for the limits were determined from repeatability and reproducibility studies which included the automobile vehicle category and the heavy truck vehicle category. See results of the research project NCHRP 10-88 for details of the repeatability and reproducibility study. Octave-band or third-octave-band measurements can be expected to have greater limits than shown above. The respective standard deviations among test results may be obtained by dividing the above limit values by 2.8.

Bias – This method technically has no bias because the test result is defined only in terms of this test method. Nonetheless, an evaluation of a component of bias found the 95 percent confidence interval for the true bias contains zero. This was determined using an RSS on the road which evaluated bias due to environmental factors which could affect the propagation of sound along the ray from the roadway to the microphone. The interval can range from -0.194 dBA to 1.440 dBA at the 95 percent level of confidence. Overall, the impact on the accuracy related to this component of bias is negligible. However a ruggedness study showed that air temperature was marginally statistically significant factor. Therefore, in lieu of further research, any comparisons between test results should also compare the temperatures during each of the tests. See the results of the research project NCHRP 10-88 for details of the bias measurements and the ruggedness study.

Statement of Precision and Bias for TP99

The findings from the present research support the suggested precision and bias statement below for the TP 99 test method.

Precision

95% repeatability limit (test-to-test, same operator)	0.5373 dBA
95% reproducibility limit (operator-to-operator)	0.1996 dBA

The values shown above for the limits were determined from a repeatability and reproducibility study. See results of the research project NCHRP 10-88 for details of the repeatability and reproducibility study. The respective standard deviations among test results may be obtained by dividing the above limit values by 2.8.

Bias – This method technically has no bias because the test result is defined only in terms of this test method. An evaluation of a component of bias found the 95 percent confidence interval for the true bias contains zero. This was determined using an RSS on the road and this included the environmental factors which could affect the propagation of sound along the ray from the roadway to the microphone. The interval can range from -0.194 dBA to 1.440 dBA at the 95 percent level of confidence. Overall, the impact on the accuracy related to this component of bias is negligible. However a ruggedness study showed that air temperature was a marginally statistically significant factor. Therefore, in lieu of further research, any comparisons between test results should also compare the temperatures during each of the tests. See the results of the research project NCHRP 10-88 for details of the bias measurements and the ruggedness study.

Influential Factors Findings

Of the five factors tested to determine if any of them could possibly influence the measurements from the test method TP 98, only air temperature showed any statistical significance and even that factor was significant only when measuring traffic sound levels from heavy trucks (see Table 6-2). There is marginal evidence which indicates that cooler temperatures impact SIPI measurements for automobiles by lowering them. When attempting to compare a pavement earlier in its lifespan to later in its lifespan, the effect of temperature could possibly mask the effect of pavement aging; this counteracts the intent. Further testing is required to confirm if air temperature's effect on both vehicle classes is statistically significant and hence an influential factor.

Table 6-2. Statistical Analysis Results of Influential Factors Testing for Test Method TP 98

Factor/Description	Change of Levels	TP 98 (Automobiles)		TP 98 (Heavy Trucks)	
		Effect on Noise	Statistical Significance	Effect on Noise	Statistical Significance
A – Pavement Age	Newer → Older	↑	No	↑	No
B – Ground Impedance	Harder → Softer	↓	No	↓	No
C – Speed Regime	Highway → Interstate	↑	No	↑	No
D – Microphone Position	25 ft. → 50 ft.	↑	No	↑	No
E – Air Temperature	Warmer → Cooler	↓	No	↓	Yes

None of the five factors' perceived impacts on CTIM measurements collected using test method TP 99 were statistically significant (see Table 6-3). Given air temperature's marginal significance on the value of the CTIM measurement, further testing of this factor can be explored.

Table 6-3. Statistical Analysis Results of Influential Factors Testing for Test Method TP 99

Factor/Description	Change of Levels	TP - 99	
		Effect on Noise	Statistical Significance
A – Pavement Age	Newer → Older	↑	No
B – Ground Impedance	Harder → Softer	↓	No
C – Speed Regime	Highway → Interstate	↑	No
D – Analysis Blocks	5 min. → 15 min.	↑	No
E – Air Temperature	Warmer → Cooler	↓	No

Suggested Improvements to Published Methods

When reviewing and implementing the test standards TP 98 and TP 99, the research team and the test operators encountered some questions and ambiguities. Some of the explanations would benefit from clarification. When standardized procedures are difficult to read or follow, an operator is more likely to misunderstand or mistake part of the procedure.

One suggestion for both standards is to explicitly write out the formulas within the body of the text. This is common in standardized test method documents. This also gives an opportunity to explicitly identify a homogenous nomenclature – the mathematical symbol for a particular quantity would be consistently referred to throughout the entire document.

Suggested Improvements to TP 98

The TP 98 standard merely states that the designated speed is the "average speed." It is not explicit whether it is the mean of the raw speeds or the mean of the logarithmic-transformation of speeds. A number of indirect indicators within the text of the standard suggest the intent was the mean of the transformed speeds, especially the section describing the regression uncertainty at the designated speed. The formulas are consistent with more general confidence interval formulas when calculated at the mean predictor variable, which is the mean of the transformed speeds, and thus cancels out numerous terms to yield the equation presented in the standard.

An additional observation of this test method is a property of simple linear regression where the regression lines always pass through the mean response at the mean predictor – in other words, the point represented by the average x-value and the average y-value will always be on the linear regression line (this is represented by the formula, $\hat{y}\{x = \bar{X}\} = \bar{Y}$, where the symbols are generic statistics). Therefore, the same value for L_{veh} determined from the regression analysis could be easily and identically determined from the arithmetic average of the individual vehicle sound levels. Statisticians are always looking for the simplest model to explain a result, and using a regression rather than a simple arithmetic mean is certainly complicating the result.

The outlier elimination described in the standard can actually detract from the reliability of a simple linear regression analysis; if the data are all concentrated in a small cluster, the slope of the regression line becomes more sensitive to small changes. There are better methods for determining outliers available (for example, some common statistical tools to detect outliers and influential data points include leverage tests, Cook's distance, and Studentized residuals), However determining a suitable method must be reserved for future research.

Suggested Improvements to TP 99

The language discussing the averaging of analysis time-blocks and reporting time-blocks can be confusing; several highly qualified operators required several readings of these sections to ensure correct understanding. Explicitly defining the formulas as described above with a consistent nomenclature throughout would help clarify.

Recommendations for Further Research

The pilot R&R study indicated that six samples and six operators may be desired to achieve a minimum power for the full R&R study. The research plan for the full R&R study couldn't accommodate that much testing, so the combination which could be accommodated and resulted in the highest predicted power and statistical SNR was selected. The shortcoming to selecting fewer operators or samples is potentially overestimating the true precision of the measurement process. This requires further study and an expanded R&R experiment design in order to determine if the precision values were overestimated.

The question of temperature effects on tire-pavement noise has been addressed in a large body of research. However, the question of air temperature's effect on both TP 98 and TP 99 is statistically significant and hence an influential factor requires further testing specifically in the context of these two test methods. Whether the temperature effects may mask the test results of TP 98 and TP 99 would also need further research.

Test method TP 98 would benefit from a statistical analysis of a step-by-step breakdown of the procedure; this would help clarify where in the measurement process the precision suffers most. Additionally, further research should examine the minimum number of pass-by measurements per vehicle type to achieve significance and also more acceptable precision.

Given that the same measuring process is used regardless of the vehicle class, it is expected that the precision of the test method TP 98 when used to measure traffic sound levels of heavy trucks would be similar to that found from measuring automobile traffic sound levels. The fact that the percent R&R value of 73 percent for heavy trucks is slightly lower than that for automobiles at 91 percent can be attributed to the higher variance of the SIPI measurements from differences in the site samples relative to the variance from the operators. Nonetheless, a review could be done to see if there are gaps in the protocols which dictate how to isolate heavy trucks versus automobiles. If these gaps exist, then the standard test method TP 98 may need to be updated to address these gaps.

The present research resulted in a very large database of observations, especially of vehicle pass-by observations for the TP 98 test method which alone resulted in over two thousand observations. It is possible that these could be reused in part for some of the above suggestions. However, the question of the masking effect of air temperature would need additional data collection with more controlled selection of measurements temperatures.

References

AIAG. 2010. *Measurement Systems Analysis Reference Manual, 4th Edition*. The Automotive Industries Action Group, Troy, MI

Burdick, R. K., and G. A. Larsen. 1997. Confidence Intervals on Measures of Variability in R&R Studies. *Journal of Quality Technology*, Vol. 29, No. 3, pp. 261–273.

Burdick, R. K., C. M. Borror, and D. C. Montgomery. 2005. Design and Analysis of Gauge R&R Studies. Cambridge University Press.

[CDOT] Colorado Department of Transportation. 2012. DTD Applied Research and Innovation Branch. Report No. CDOT-2012-5. *Tire Pavement and Environmental Traffic Noise Research Study*. By R O. Rasmussen and R. C. Sohaney.

Donovan, P. R. 2012. SPR-577 Quiet Pavement Pilot Program: Progress Report 3. Rep. no. FHWA AZ-12-577, Phoenix, AZ: Arizona Department of Transportation.

[FHWA] U.S. Department of Transportation Federal Highway Administration. 4 Feb 2013. *FHWA-HEP-10-025Highway Traffic Noise: Analysis and Abatement Guidance*. Web. http://www.fhwa.dot.gov/environment/noise/regulations_and_guidance/analysis_and_abatement_guidance/revguidance.pdf.

Harvey, J. T. and E. Kohler. 2011. "Quieter Pavement Research: Concrete Pavement Tire Noise." Rep. no. CA12-1200 A.

Herman, L. A. and J. M. Withers. 2005. "Effectiveness of Tire/Road Noise Abatement Through Surface Retexturing by Diamond Grinding for Project SUM-76-15.40." Rep. no. FHWA A/OH-2005/009.

Hintze, J. 2009. Number Cruncher Statistical System (NCSS). NCSS, LLC, Kaysville, UT. http://www.ncss.com.

Hurvich, C. M. and C-L. Tsai. (1989). "Regression and time series model selection in small samples." *Biometrika*, 76, 297—307.

Kutner, M. H., C. J. Nachtscheim, J. Neter, and W. Li. 2005. *Applied Linear Statistical Models*. McGraw-Hill Irwin, Burr Ridge, IL.

Minitab. "Technical Support Document: Number of Distinct Categories and %Study." *Minitab: Knowlegebase ID 2393.* 2010. Accessed May 29, 2013: http://www.minitab.com/support/documentation/Answers/ndc%20and%20StudyVar.pdf.

Neter, J., M. Kutner, W. Wasserman, C. J. Nachtsheim. 1996. Applied Linear Statistical Models. Fourth Edition. Irwin, Chicago, IL. Pages 948–951.

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.r-project.org/.

Rochat, J. L. 2009. "Developing U.S. Wayside Methods for Measuring the Influence of Road Surfaces on Traffic Noise." Inter-noise 2009: Innovations in Practical Noise Control. Washington: Institute of Noise Control Engineering.

Rochat, J. L., D. R. Read, and G. G. Flemming. 2010. "Caltrans Thin Lift Study: Effects of Asphalt Pavements on Wayside Noise." Rep. no. CA 10-0146.

Rochat, J. L., D. Lodico, P. Donavan, and R. O. Rasmussen. 2012. "Overview and Application of the Continuous-Flow Traffic Time-Integrated Method (CTIM) for Determining the Influence of Road Surfaces on Traffic Noise." Internoise 2012: Quieting the World's Cities. New York City.

Sandberg, U. and J. A. Ejsmont. 2002. Tyre/road Noise: Reference Book. Kisa, Sweden: INFORMEX.

SAS Institute, Inc. No date. Statistical Analysis Package. Cary, NC 27513, www.sas.com.

SPSS, Inc. 2010. SPSS Version 19 – Statistical Package for Social Sciences. http://www-01.ibm.com/software/analytics/spss/.

The MathWorks, Inc. 2012. MATLAB Release 2012b. Natick, MA.

List of Abbreviations, Acronyms, Initialisms, and Symbols

µPa	micropascal
AASHTO	American Association of State Highway and Transportation Officials
ADOT	Arizona Department of Transportation
AICc	Akaike's (corrected) Information Criterion
ANSI	American National Standards Institute
ASTM	ASTM International (formerly American Society for Testing and Materials)
CTIM	continuous-flow traffic time-integrated method
dB	decibel
dBA	A-weighted decibel
ETG	expert task group
FHWA	U.S. Federal Highway Administration
ISO	International Organization for Standardization
MSE	mean square error
NCHRP	National Cooperative Highway Research Program
NIST	National Institute of Standards and Technology
Pa	pascal (SI unit of pressure)
R&R	repeatability and reproducibility
RSS	reference sound source
SIP	statistical isolated pass-by method
SIPI	statistical isolated pass-by index
SLM	sound level meter
SNR	signal-to-noise-ratio
SPL	sound pressure level
TTC	temporary traffic control

APPENDIX A

Tire-Pavement Noise Measurement Background and Literature Review

Introduction

The FHWA conducted two tire-pavement noise strategic planning workshops in 2004 and 2006, which lead to the creation of an Expert Task Group (ETG) on tire-pavement noise. The ETG was tasked with developing tire-pavement noise measurement standards for use in the United States. Uniform tire-pavement noise measurement methods were developed to facilitate a comparison of tire-pavement noise measurements performed throughout the U.S. Measurement methods developed by the ETG include the CTIM measurement method, and the SIP measurement method (Rochat 2009; Rochat et al. 2012b).

AASHTO Standard Test Method TP 98-13 (Standard Method of Test for Determining the Influence of Road Surfaces on Vehicle Noise Using the Statistical Isolated Pass-By (SIP) Method) describes procedures for measuring the influence of the pavement surface on traffic noise and provides a quantitative measure of the sound pressure level at locations adjacent to a roadway. The SIP method allows for comparison of roadways of varying surfaces by comparing measured sound levels to those representing the tire-pavement noise for the average U.S. pavement (Rochat 2009). Due to its ability to compare results across test sections and study areas the SIP method is preferred over CTIM (Rochat 2012). Microphones are positioned 25 and 50 feet from the roadway in SIP measurements.

AASHTO Standard Test Method TP 99-13 (Standard Method of Test for Determining the Influence of Road Surfaces on Traffic Noise Using the Continuous-Flow Traffic Time-Integrated Method (CTIM) describes procedures for measuring the influence of pavement surface on traffic noise at a specific site adjacent to a roadway. The CTIM measurements are intended for use on high traffic roadways to capture sound from all vehicle types traveling on all roadway lanes during the measurement period. Measurements are performed at a distance of 50 to 100 feet from the roadway; therefore, measured sound levels include propagation effects over the road surface and ground between the traffic and the measurement locations. Measurement techniques aimed at measuring traffic noise, versus isolated pass-byes, are generally applied to roadways where measuring isolated pass-by events would not be possible due to the traffic density. The method currently allows for comparison of varying or aging pavement surfaces on a single roadway; extension of the normalization process to include site variation to allow for site-to-site comparisons are being examined (Rochat et al. 2012b).

This appendix presents a review of the available literature in support of the development of precision and bias statements for AASHTO Standard Test Methods TP 98 and TP 99. The literature review collected and evaluated information relevant to the measurement of the influence of road surface on traffic and vehicle noise and associated precision and bias. This chapter presents the results of the literature review. Background research focused on: sources of vehicular traffic noise, the effect of roadway surface on vehicular traffic noise, various methods of tire-pavement noise measurement and the variance in wayside noise measurement due to environmental factors, fleet vehicle variance, site specific conditions and instrumentation.

While extensive research has been performed concerning the individual factors which influence tire-pavement noise and wayside measurement of traffic noise, few studies addressing the precision and bias of the CTIM and SIP test methods have been performed. The limited studies of precision and bias for statistical pass-by measurement of tire-pavement noise utilized alternate international standards which vary somewhat from the current AASHTO test procedures. The differences between TP 98 and other methods are likely to influence the precision and bias somewhat.

The research team collected and evaluated information relevant to the measurement of the influence of road surface on traffic and vehicle noise and associated precision and bias. The background research focused on the following topics:

- Sources of vehicular traffic noise;
- The effect of roadway surface on vehicular traffic noise;
- Methods of tire-pavement noise measurement and
- The influence of human factors, environmental factors, site specific conditions, test vehicles and measurement instrumentation on measurement precision and bias.

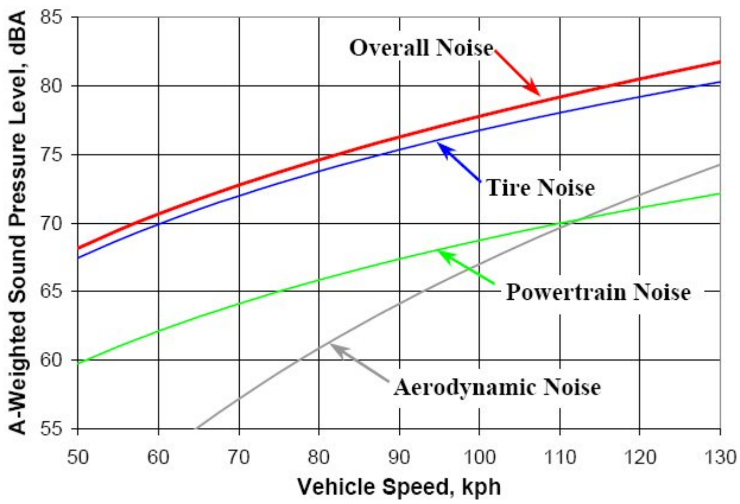
The background research consisted of a literature review of past and current information and phone interviews to collect additional information. Relevant literature consisted of research reports, journal articles, conference presentations and technical papers, measurement standards and reference books.

Background research related to the measurement of the influence of road surface on tire-pavement noise included a number of reports focusing on quiet pavement technologies. The literature review included studies of various roadway materials and textures in which the influence of road surface on traffic and vehicle noise has been documented through either wayside measurement or direct measurement at the source.

The research team also reviewed current measurement methods relating to the influence of road surface on traffic and vehicle noise including: AASHTO tire pavement measurement standards; ISO standards; FHWA measurement methods and others.

Mechanics of Vehicle Traffic Noise

Traffic noise is the total sound resulting from all vehicles traveling on a roadway. Traffic noise is influenced by the number of factors including the traffic volume, the vehicular mix, speed, vehicle operating and maintenance characteristics, and pavement type. Vehicle noise is the sound resulting from a single vehicle on a roadway and includes sound created from various sources of noise on and surrounding a vehicle as shown in Figure A-1.



(Source: Donavan 2004)

Figure A-1: Vehicle Noise Sources

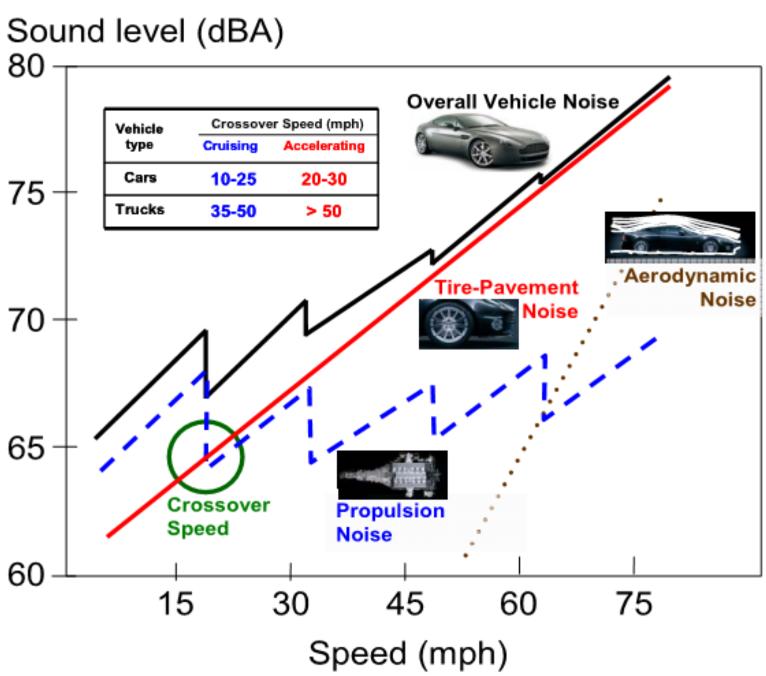
Automobiles and heavy vehicles generate noise from several sources. Sources of vehicular noise in passenger vehicles fall into three general categories: power train noise, aerodynamic noise (air turbulence), and tire-pavement noise (Sandberg and Ejsmont 2002). Alternative classification of vehicle noise sources exist but generally include similar components. For the purpose of this research vehicle noise source categories will be defined consistently with AASHTO TP 98 and TP 99.

Aerodynamic noise or air turbulence is a source of automotive noise. Aerodynamic noise, with reference to traveling vehicles, is caused by turbulent airflow over, around and through a vehicle (Tire-Pavement Noise Research Consortium 2011). Generally aerodynamic noise is not a significant contributor to exterior vehicle noise at legal speeds due to the aerodynamic design of vehicles (Sandberg and Ejsmont 2002).

AASHTO TP 98 and TP 99 define power train noise as the noise generated from the power train including the vehicle engine, exhaust system, air intake, fans, transmission, differential and axles. Power train noise is also sometimes referred to as propulsion noise, power-unit noise or drive-train noise. At travel speeds which are constant and less than 30 mph, the power train can be the dominant noise source in the overall traffic noise level (Tire-Pavement Noise Research Consortium 2011). Power train noise is also the dominant vehicular noise source under stop-and-go travel conditions, near intersections and when climbing grades.

Tire-pavement noise is the sound generated by the interaction of the tire with the pavement surface as it traverses the pavement. While noise produced from tire-pavement interaction is just one source of traffic noise, it becomes the primary noise source for vehicles traveling at high speeds (CDOT 2012; Leasure 1975; Tire-Pavement Noise Research Consortium 2011).

Figure A-2 depicts the overall sound pressure level associated with vehicular pass-by events and the relative contribution of aerodynamic noise, powertrain noise and tire-pavement noise at various operating speeds.



(Source: Rasmussen et al. 2007)

Figure A-2: Vehicle Noise Contribution

As shown in Figure A-2, at low travel speed power train related noise is the dominant source in overall vehicle noise. There is a point at which the increasing tire-pavement noise equals and then exceeds the power train noise, referred to as the crossover speed. Beyond the crossover speed, around 30 mph for many vehicles, tire-pavement noise becomes the largest contributor to overall vehicle noise. At greater speed aerodynamic noise begins to increase but is still a lesser contributor to overall noise than tire-pavement interaction at legal traveling speeds. Naturally, the contributions of various noise sources are highly particular to individual vehicles, in terms of make, model, age, state of repair, and even the habits and behaviors of individual drivers.

In addition to the three primary categories of vehicle sound sources (power train, aerodynamic and tire-pavement) heavy vehicles, such as trucks and busses, also create noise associated with special equipment such as pneumatic and hydraulic auxiliary systems. Despite the presence of additional sound sources in heavy vehicles statistical analysis of the vertical distribution of noise indicates that for the majority of truck pass-byes measured at highway speeds tire-pavement interaction was the dominant sound source (Gurovich et al. 2009). These findings support that tire-pavement noise typically dominates vehicular noise for both automobiles and heavy vehicles at moderate to high speeds provided the vehicle is equipped with a reasonably good exhaust muffler and is in a good state of repair (Leasure 1975). Regarding heavy truck noise sources, roadside acoustic beamforming measurements identified a minority of trucks which exhibit much stronger power-train noise than tire-pavement noise through their engine casing and exhaust systems (Gurovich et al. 2009). In roadside noise measurements performed for this project, this small proportion of trucks is clearly distinguishable by their strong low-frequency rumble which is apparent even without acoustic beamforming to the naked ear as well as through a spectrum analyzer.

Tire Pavement Noise

Tire-pavement noise generation has been studied in detail since the 1970s (Sandberg and Ejsmont 2002). While the research for NCHRP project 10-88 does not address the particular tire-pavement noise generation and amplification mechanisms in detail, the primary mechanisms for tire-pavement noise are listed below for reference (Rasmussen et al. 2007; Leasure 1975; Sandberg and Ejsmont 2002).

- Tread / texture impact;
- Air turbulence,
- Air pumping,
- Stick-slip,
- Stick-snap,
- Acoustical horn
- Helmholtz resonance
- Pipe resonance
- Sidewall vibrations
- Tire resonance

Use of quiet pavement technologies is most effective at highway speeds with free flowing traffic, when tire-pavement noise is the primary contributor to the total traffic noise (Tire Pavement Noise Research Consortium 2011). Similar to other noise mitigation or abatement strategies, the influence of quieter pavement will be most beneficial closer to the roadway, due to the dominance of the road traffic noise, whereas further from the roadway the perceived benefit will be diminished where other ambient noise sources begin to dominate the noise environment. Several studies, including the Arizona Quiet Pavement Pilot Program and FHWA Pavement Effects Implementation Study, have shown that distance from the roadway also plays a role in the effectiveness of quiet pavement technologies in this manner.

Road Surface Properties

The primary road surface properties which affect tire-pavement noise are: roadway texture, roadway material and the presence of joints or non-continuous roadway features. Differences in these roadway properties are the cause of variability in tire-pavement noise between roadway types. Likewise variations in road surface texture and construction can cause measurable differences within a single pavement type or test section.

Roadway surface texture is generally categorized by texture wavelength. Road surface texture categories are: megatextures, macrotextures and microtextures. Megatextures have the largest wavelength with wavelengths ranging from 50 to 500 mm. Macrotexture, with wavelengths ranging from 0.5 to 50 mm, is the roughness or texture that encompasses the tire tread elements and road aggregate up to the size of the tire/road interface area. Microtextures have the smallest wavelength, less than 0.5 mm, and are obtained by the surface properties of the individual chippings or surface particles (Sandberg et al. 2002).

Road surface texture parameters such as texture wavelength, direction, uniformity and peak amplitudes influence tire-pavement noise. Furthermore, texture with predominantly transverse linear texture exhibits greater tire-pavement noise than an otherwise identical texture but with predominantly longitudinal linear features. More precisely, tire-pavement noise emission on an anisotropic surface will be higher if the texture is traversed in a direction perpendicular to its main features (transversely) and lower if traversed along the direction of its main features (longitudinally), provided all other features except texture direction are the same (Sandberg and Ejsmont 2002).

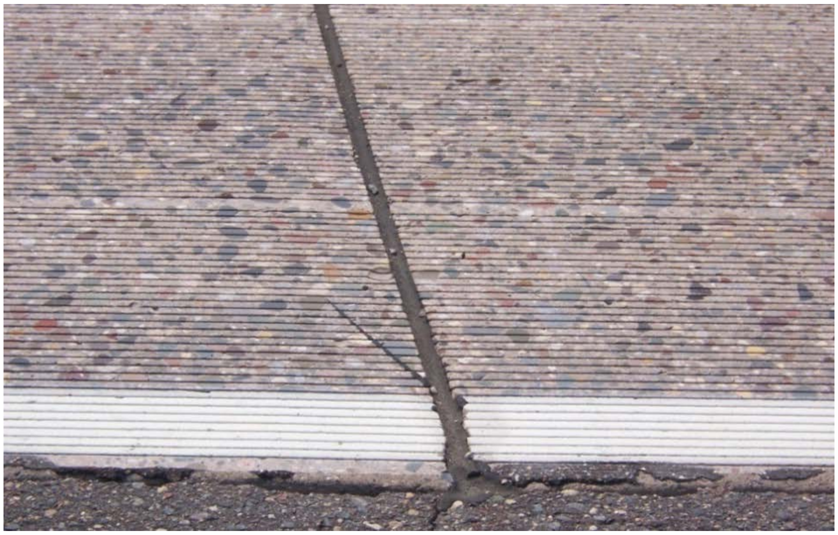
Roadway material properties such as porosity, layer thickness and stiffness of a roadway material also play a role in tire-pavement noise generation and propagation. Porosity is defined as the ratio of the volume of air voids to the total volume of a material. Materials used in most pavement surfaces have porosity less than 5 percent. When increased to 20 percent or more and/or when air can flow through the material, the result can be a benefit in noise reduction (Rasmussen et al. 2007). Tire-pavement noise is reduced with porous roadways through two mechanisms: the modification of tire/road generation at the source, and the partial elimination of sound waves reflected off the road surface in the propagation to the receiver. Porous pavements can provide reductions to A-weighted sound levels of 3 to 5 dB, compared to a dense pavement structure (Peeters and Kuijpers 2008; Berengier et al. 1997).

The thickness of the surface layer also influences the acoustical absorption for a roadway. The increased or decreased thickness of a porous layer affects the frequency of peak sound absorption for a road surface. Increasing the thickness moves peak absorption to lower values. Road surfaces can be engineered optimize noise reduction by using a porous layer thickness which increases sound absorption in the dominant frequency spectra emitted by traveling vehicles.

The mechanical impedance of the road, or stiffness, also influences tire-pavement noise generation. Generally the stiffness of road surfaces is several orders of magnitude higher than that of the rubber in a tire (Morgan 2006). Lowering the road mechanical impedance generally reduces the tread block impact forces transmitted into a tire which reduces tire vibration. Pavements that have stiffness characteristic approaching that of a tire can be quieter than those that are more typical of asphalt and concrete in use today (Rasmussen et al. 2007).

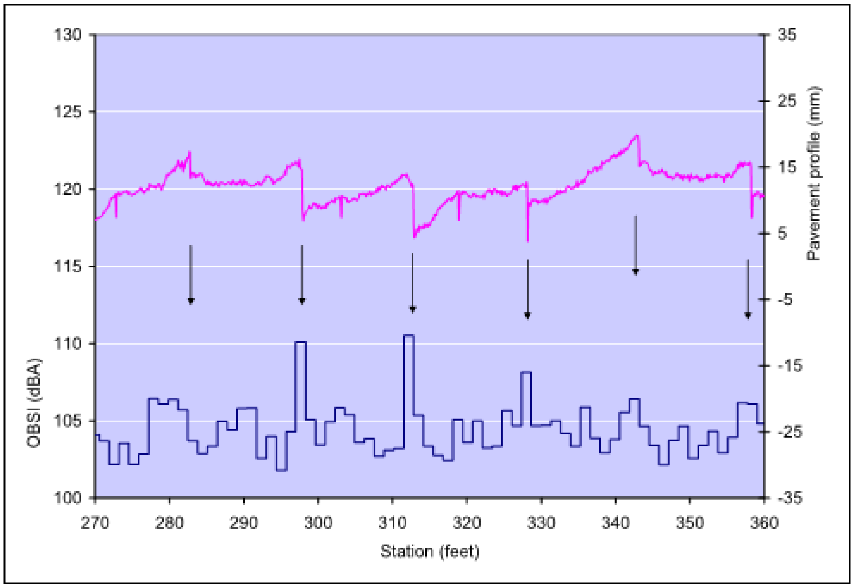
Roadway joints and discontinuous road features such as cracks can cause an audible and measureable increase in tire-pavement noise.

The joints between pavement sections, as shown in Figure A-3, can cause a distinct sound that has often been described as an audible clap or slap (Donavan 2011; Sandberg and Ejsmont 2002). The increase in sound due to joints is dependent on surface type and texture.



(Dick et al 2010)
Figure A-3. Typical Roadway Joint

Wayside and near-field tire-pavement measurements of joint noise have recorded spikes in sound level ranging from 5 dB to 10 dB over the constant tire-pavement noise due to the sound of rolling over a joint (Harvey and Kohler 2011; Sandberg and Ejsmont 2002). Figure A-4 depicts the increase in sound level and corresponding roadway profile for an on-board sound intensity (OBSI) tire-pavement noise measurement of a faulted jointed concrete pavement.



(Harvey and Kohler 2011)
Figure A-4. Sound Level Spike due to Roadway Joint and Roadway Profile

Figure A-4 shows a maximum increase in noise level of 8 A-weighted decibels (dBA) due to joints in the roadway.

A Caltrans study examined overall A-weighted sound levels with and without joints for three PCC (concrete) surface textures: longitudinal tining, burlap drag, and broom surfaces using the OBSI tire-pavement noise measurement technique. The study found that the increase in sound due to joints was dependent on surface texture. Joints for the longitudinal tine surface increased the overall level by 1.3 dB while for the burlap drag and broom surfaces, the joints added 1.9 and 2.0 dB, respectively (Donavan 2011).

The manner in which roadways maintain their acoustical properties over time, or acoustic longevity, has also been studied through measurement. The changes in the acoustical characteristics of a road surface over time can be attributed to changes in roadway texture and material due to vehicular traffic, climate changes and wear. Generally tire-pavement noise increases over time.

Acoustical absorption is linked to the porosity of the roadway material; therefore, air voids or inclusions in the surface material are necessary for absorption to occur. Over time porous road surfaces may not keep their acoustical characteristics. The variation in the acoustical characteristics is due to the clogging of the structure by various wastes causing the flow resistivity to increase and the porosity to decrease (Berengier et al 1997).

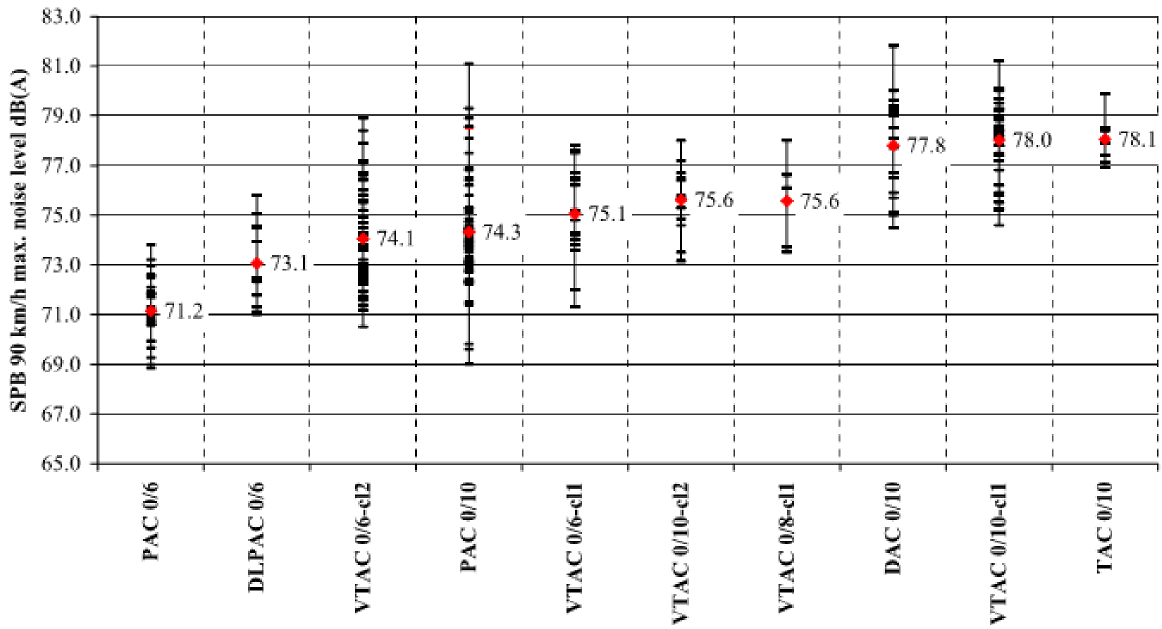
Changes in roadway texture and stiffness can also cause an increase in traffic noise level for roadways over time. With use and changes in temperature road surface texture may change such as the development of cracks or the exposure of course aggregate. Asphalt pavements may increase in stiffness due to traffic loading. All of these factors: porosity, stiffness and texture will influence the acoustic longevity, and acoustical performance of a road surface.

Differences in Tire-Pavement Noise between Various Road Materials and Textures

Studies have found tire-pavement noise to vary greatly between different road materials and textures and variations among similar road surfaces. A study of tire-pavement noise performed by TUG measured a range in sound levels of 17 dB between various road surfaces using the close-proximity measurements method (Sandberg and Ejsmont 2002).

Several countries including France, Germany and the U.S. have begun developing databases of tire-pavement noise for various road types. These databases allow for the comparison of road surfaces based on their acoustical characteristics including absorption and sound intensity for new pavements and through the life cycle of the road. The French road laboratory (LRPC) in Strasbourg has one of the most comprehensive databases on road surface influence on tire-pavement noise. Figure A-5 presents the results of statistical pass-by measurements on various road surfaces including porous asphalt concrete (PAC), thin and very thin layer asphalt concrete (TAC, VTAC) and dense asphalt concrete (DAC).

Figure A-5 presents the maximum pass-by level (L_{Amax}) for each road surface test section. Pass-by levels represent light weight vehicles traveling at 90 km/hr. Each dot in the figure represents a measured road surface section, the bar represents the range of sound levels measured for each road surface type and the red symbols represent the average noise level.



(Brosseaud and Anfosso- Lédée n.d.)

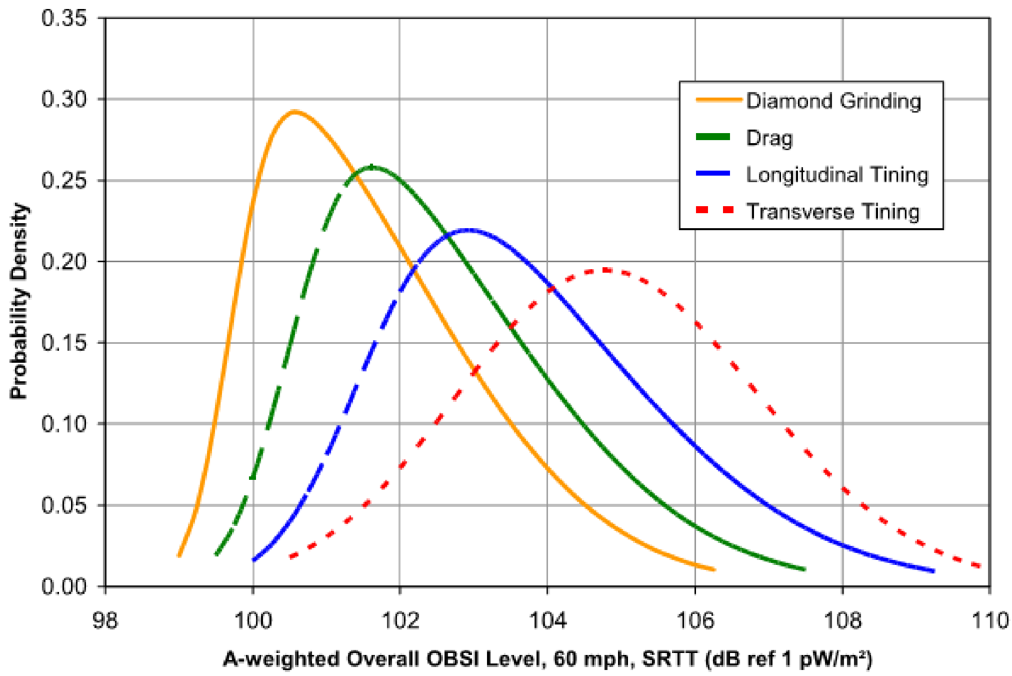
Figure A-5. Light Vehicle SPB Level at 90 km/h at 20 °C, for Different Low Noise Road Surfaces: Output from the French Acoustic Database of the LRPC Strasbourg

Figure A-5 illustrates how statistical pass-by measurements can be used to compare the influence of road surface materials on vehicle noise. Measurement results from the French National Road Laboratory (LCPC) database found new porous asphalt concrete to be quieter than other road surfaces at equivalent grading (Brosseaud and Anfosso- Lédée n.d.).

A study performed for the Ohio Department of Transportation (ODOT) included noise measurements, utilizing the ISO 11819-1 Statistical Pass-By Method at 12 sites with various pavement types, textures, and pavement age. Tested pavements consisted of dense-graded asphaltic concrete, open-graded asphaltic concrete, stone mastic asphaltic concrete and Portland cement concrete (PCC) with various textures. Based on Statistical Pass-by measurements, a difference of 6.7 dB was measured between the lowest and the highest SPBI (Statistical Pass-By Index) for all of the pavements measured. New open-graded asphalt concrete had the lowest sound levels while PCC with random transverse grooves resulted in the highest sound levels. The LCPC close proximity measurements and ODOT statistical pass-by measurements demonstrate that variations in tire-pavement noise between road surface materials can be measured both at the source and at a distance from the road (wayside) (Herman and Ambroziak 1999).

Differences in Tire-Pavement Noise for Various Textures and Grindings

Several studies have been aimed at comparing the influence of road surface texture on traffic and vehicle noise. Research performed by the National Concrete Pavement Technology Center has found large variability in sound intensity between various textures and even among similar textures (Herman and Ambroziak 1999; Harvey and Kohler 2011; Rassmussen et al. 2011). Figure A-6 depicts the normalized distribution of sound intensity levels, measured using the OBSI measurement method, of four concrete textures (diamond grinding, drag, longitudinal tining, and transverse tining).



(Rasmussen et al. 2012)

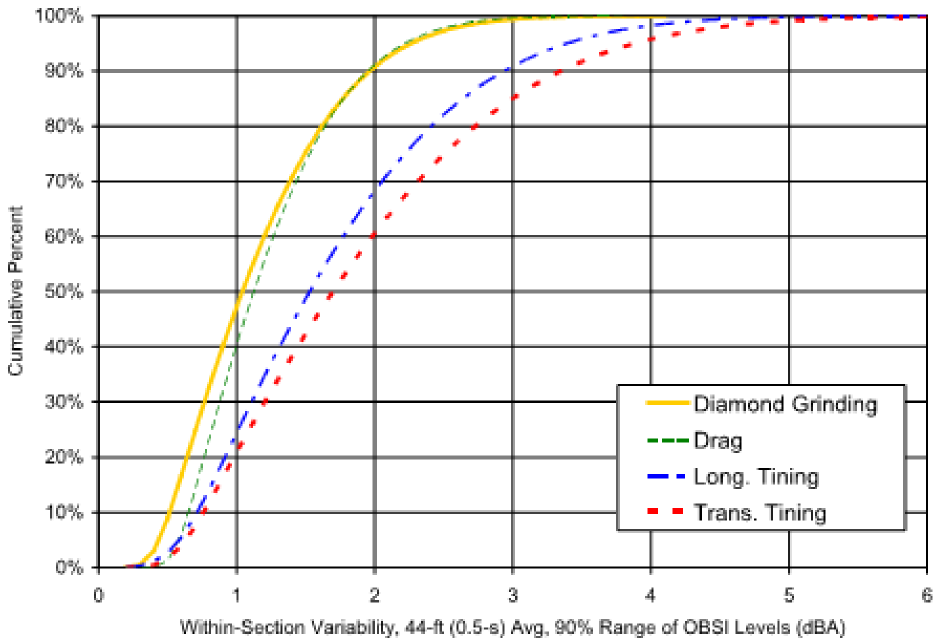
Figure A-6. Normalized Distribution of OBSI Noise Levels for Convention Concrete Pavement Textures

As shown in Figure A-6 all conventional nominal textures of concrete have the potential to be constructed as quieter concrete surfaces, though some are more likely to be quieter than others (Rasmussen et al. 2012). Alternate studies of concrete textures have been performed using the statistical pass-by measurement method found in similar trends. A study performed by Ohio University compared the noise of three types of cement grooving types. Of those tested longitudinal grooves produced the lowest sound levels, followed by transverse, then random-transverse grooves (Herman and Ambroziak 1999).

Tire-Pavement Noise Variability within Similar Roadway Materials and Textures

Measurements performed in California as part of the Quieter Pavement Research program also found that there is a large variation in tire/pavement noise level for the texture types evaluated. The difference between the lowest and highest OBSI levels for the same nominal texture type exceeded 5 dBA without considering texture conditions, pavement age, and traffic and climatic conditions (Harvey and Kohler 2011). Greater variability in sound level among the same nominal texture can be found in studies including road surfaces of various ages. For example the noise distributions in Figure A-6 depict that among the same nominal texture OBSI sound levels can vary more than 8 dB.

Differences in tire-pavement noise have been measured among the same nominal texture type and within a single test section. Figure A-7 depicts the within section variability of OBSI levels collected by the Concrete Pavement Surface Characteristics Program. The figure illustrates the cumulative within-section variability of four concrete textures: diamond grinding, drag, longitudinal tining and transverse tining.



(Rasmussen et al. 2012)

Figure A-7. Distribution of Within-Section Variability of OBSI Noise Levels by Texture

Figure A-7 illustrates that diamond grinding technique typically results in sections that have low within-section variability. Tined sections can also have low within-section variability, but the distributions also indicate they have a greater probability of having higher within-section variability (Rasmussen et al 2011). Variability within test sections result in differences in measured sound levels at the source as shown in the OBSI measurements. This variability is also likely to influence sound levels measured at a distance from the roadway.

Differences in Tire-Pavement Noise with Age (Acoustic Longevity)

The acoustic longevity of road surfaces has been studied through measurement using both measurement at the source and wayside measurement (Rochat et al. 2010; Donavan 2011; Donavan 2012). The measured acoustic longevity of road surfaces varies between studies but trends in the rate of noise level increase over time have been measured among road surfaces of similar texture. For example, the California Mojave Bypass Acoustic Longevity Study found PCC textured pavements to have lower rates of overall noise level increase over time relative to asphalt concrete pavements (Donavan 2011).

The California Thin Lift Study measured statistical pass-by noise levels at five pavement sections to document the acoustic longevity of the test sections between the ages of 4 months and 52 months. Road surfaces included in the study were: dense-graded asphaltic concrete, two open-graded asphaltic concrete of varying thickness; rubberized asphaltic concrete, open rubberized asphaltic concrete and bonded wearing course. Results from the study indicate the relative noise reduction between dense-graded asphaltic concrete and the test road surfaces over 52 months when comparing pavements of the same age. Although the noise reduction was maintained, each of the pavements tested showed some deterioration in noise benefit. The deterioration of the noise benefit over 52 months was up to 1.5 dBA depending on the pavement type (Rochat et al. 2010).

Measurement Descriptions

This research is concerned with the two AASHTO standard test methods: TP 98 and TP 99. They are both road wayside noise measurements; the former is a test of a pavement's influence on vehicle noise, whereas the latter is a test of a pavement's influence on traffic noise. The appropriate measurement method for a particular study will vary depending on the parameters of interest (tire-pavement noise versus single vehicle noise versus total traffic noise) as well as measurement conditions and roadway traffic conditions. In many instances multiple measurement methods may be used in the same study to determine the influence of tire-pavement noise at a receiver of interest along the roadway and sound intensity at the tire.

Other measurement methods to quantify tire-pavement noise and the influence of road surface on traffic and vehicle noise exist in the U.S. and abroad. Tire/pavement Source measurements noise quantify the tire-pavement noise in isolation, either through on-board noise measurements over in-situ pavement or through laboratory noise measurements of pavement specimens. These methods, while related to the vehicle noise and the total traffic noise, are not addressed in this research. Additionally there are other standardized measurements of wayside noise, some of which are only briefly addressed in the following overview of the two AASHTO wayside noise test methods.

Arithmetic Mean Sound Level versus Equivalent-Average Sound Level

Mainly for the benefit of statisticians unfamiliar with decibels, there are two ways to determine an average sound level, each appropriate in context. The simplest computation is the arithmetic mean of the decibel-levels. However, an arithmetic mean of levels is only appropriate in the context of a statistical analysis. The more common computation is an equivalent-average sound level, which is generally a much more appropriate method of averaging sound pressure levels (SPL) or any other type of level on a decibel-scale. A decibel is itself a logarithmic transformation of the sound pressure (specifically, the SPL in decibels is twenty times the base-ten logarithm of the ratio of sound pressure in micropascals to a reference pressure of 20 micropascals). The equivalent-average sound level is the preceding logarithmic transformation after finding the mean of the pressures, or stated another way it is the decibel-level of the mean micropascals.

Naturally in any statistical analysis the arithmetic mean of sound levels is preferred. Otherwise, in any time-integration or spatial-integration of sound levels, the equivalent-average sound level is appropriate.

Statistical Isolated Pass-By Test Method

The method expressed in AASHTO Standard Test Method TP 98, *Determining the Influence of Road Surfaces on Vehicle Noise Using the Statistical Isolated Pass-By (SIP) Method*, reflects the noise of a single isolated vehicle traveling over a road surface. The SIP test method focuses on the effect of pavement on vehicle noise and reduces the effects of interfering factors. Tire/pavement noise is only one of many factors which influence vehicle noise; therefore, the SIP test should be conducted to minimize influences which mask or interfere with tire/pavement noise. Other than the pavement, two of the largest influences on vehicle noise at a wayside measurement position are the conditions of the vehicle itself, and the traveling speed of the vehicle.

The SIP test method minimizes the influence from the conditions of the vehicle itself in two ways: it separates similar vehicle types and utilizes the concept of an average vehicle. The vehicle categories are consistent with FHWA vehicle noise categories: automobiles, heavy trucks, etc. Separate analyses are performed for these categories of similar vehicles.

The average vehicle is determined by a sample set from the population of the fleet of vehicles in any one category. It is average in the vehicle's other noise-producing sources such as the power train and the vehicle's tires, including the effects of age and maintenance conditions on those noise sources. The average vehicle is not any particular vehicle that one can find; it is analogous to stating an average family had 1.3 children in 1955 (U.S. Bureau of the Census 2004). It is obviously impossible for any family to actually have 1.3 children; it is an average from the population of families. Similarly, the SIP test result reflects the noise of an average vehicle from the vehicle fleet, and does not represent any particular vehicle.

The SIP test method controls the influence of the traveling speed on vehicle noise with a liner regression of the vehicle pass-by sound level versus the common logarithm of the vehicle speed. The test method then compares the regression model with a reference model.

In the broadest terms, statistical pass-by tests are noise measurements on the road wayside of a statistically average vehicle as a function of speed. This is useful to compare the test results of two pavements. The results represent the noise of two average vehicles which are in all other respects identical (or at least statistically representative of identical vehicles) traveling over the different pavements. Therefore, any difference in the test results should be mainly attributable to the influence of the road surface on the vehicle noise level.

Overview of AASHTO Standard Test Method TP 98

The AASHTO Standard Test Method TP 98 method includes microphones positioned adjacent to the roadway to capture the maximum sound pressure level of isolated vehicle pass-by events. Additionally, vehicle category and traveling speed of vehicle pass-by events must be measured; several options for speed measurement are available.

Several procedures within the method are aimed at maximizing the tire/pavement noise and minimizing interfering or masking sounds. For instance, free-flowing traffic conditions allow the greatest potential influence of pavement on vehicle noise, whereas vehicles under acceleration or maneuvering may produce more power-train noise which would mask the influence of pavement on vehicle noise. Additionally, vehicle pass-by events which are not isolated are discarded from the sample set because nearby vehicles have the potential to influence the wayside noise measurement during the pass-by event. Vehicles are categorized into groups of similar vehicle types and each vehicle category analyzed separately.

Statistical analysis of the measurement data is a linear regression model. The response variable is the measured vehicle pass-by sound levels, and the predictor variable is the common logarithm of the vehicle speed. The final test result of the SIP method is the Statistical Isolated Pass-By Index (SIPI), which is the difference of the linear regression model from a reference curve, each calculated at the same designated speed. The designated speed is the average vehicle speed from the sample set. While other regression diagnostics are reported none of this information is used in the eventual test result.

The reference curve is fixed for each vehicle category. It is a simplified linear model from FHWA standardized models developed in the early 90s. The reference curve is based on conducting a multitude of similar tests over many different pavement types, aggregated to an average pavement. The reference curve is generally considered the “National average.”

Differences between Standard Methods of Test

There are other standardized test methods for statistical pass-by measurements; two in particular are International Organization for Standardization (ISO) standards: ISO 11819-1, *Acoustics – Measurement of the influence of road surface on traffic noise – Part 1: Statistical Pass-By Method*; and ISO 11819-4, *Acoustics – Method for measuring the influence of road surfaces on traffic noise – Part 4: SPB method using backing board*. All these standards operate under the same general premise: an average vehicle pass-by noise level determined statistically.

The selections of standards to measure statistical isolated pass-by noise have similar approaches. They differ from each other only slightly in their execution. Most of the differences between the

AASHTO TP 98 and ISO 11819-1 test methods are generally minor differences between site selection, accepting or discarding vehicle pass-by events, vehicle classification criteria and equipment setup. The most pronounced difference between the two methods is in the final test determination.

The final test determination for the ISO standard is the Statistical Pass-By Index (SPBI), which is a vehicle-mix weighted average of the vehicle sound levels for each vehicle category. This is perhaps the most distinguishing characteristic between the AASHTO test method and the analogous ISO test method. The vehicle sound level is the predicted value of the regression model at one of three prescribed speeds. The three prescribed speeds represent three categories of road speeds, each with its own distinct reference speed for either passenger cars or heavy vehicles. The reference speed is used for calculating the regression model value, not the mean speed. However, to avoid the issue of increasing uncertainty with increasing departure from the mean speed, the standard specifies the reference speed shall be within one-and-a-half standard deviation of the mean speed for passenger cars (one standard deviation for heavy vehicles).

The ISO 11819-4 test method differs from ISO 11819-1 in that it allows provisions to measure in urban areas or areas with noise barriers or other reflecting surfaces. In addition, there are several test methods described as controlled pass-by measurements (CPB). The CPB test procedure differs in that the vehicle pass-by events are not random vehicles on the road; rather the traffic is controlled and known vehicles are driven over the pavement specimen repeatedly in front of the measurement microphones. This and other tire/pavement source noise measurements are used for noise classification of tires; this is required for some export markets.

Continuous-Flow Traffic Time-Integrated Method

The test method is expressed in AASHTO Standard Test Method TP 99, Standard Method of Test for Determining the Influence of Road Surfaces on Traffic Noise Using the Continuous-Flow Traffic Time-Integrated Method (CTIM) reflects the noise of all traffic which travels over a road surface, as measured at a road wayside location. Traffic noise is the combined noise of all the vehicles on the roadway. Traffic noise is often quantified through community noise measurements, which measures the overall noise at a location within a community. While community noise measurements include the influence of pavement on traffic noise, they neither control for other noises which contribute to the overall community noise level, nor do they control for the variations in traffic over the roadway. The CTIM test method offers more control over these variables in order to focus the measurement results on the influence of pavement on traffic noise.

The most influential factor on traffic noise is the vehicles which comprise the traffic, specifically the density of traffic, the types of vehicles in traffic, and the traveling speeds of the vehicles in traffic. In addition, highly influential factors on wayside noise levels are non-traffic interfering noise events. Both of these factors are controlled by measuring sound levels over a number of relatively shorter time-blocks (an equivalent-average). The shorter time-blocks allow time-blocks to be discarded when interfering non-traffic noise events occur.

In addition, the CTIM test method includes a normalization procedure. Normalization reduces the influence (on individual time-blocks) from differences in the vehicles which comprise the traffic between time-blocks. The CTIM normalization is based upon predictive modeling of traffic noise for each time-block. Each time-block measured sound level is normalized by the modeled traffic sound levels. A time-block with heavy traffic will be both measured and modeled at a louder noise level. Therefore, this normalization should bring extreme values closer to the overall mean. This normalization procedure has the effect of reducing the overall variance of the time-blocks around the mean time-block sound level.

The application of the CTIM test method is very limited. It is not recommended for comparing pavements at two different sites. Its best use is to compare a change in pavement over time at a single site (such as the wear of pavement over years, or before-and-after resurfacing or replacing the pavement). A comparison requires knowledge of the previous normalization process in order to normalize each subsequent set of time-blocks to the same reference traffic.

There is general recognition that CTIM is a less precise method than SIP, and that if possible SIP is preferred from these two methods. But where traffic is too dense to capture isolated vehicle pass-by events, the CTIM method is available. The SIP method measures vehicle noise, whereas the CTIM measures traffic noise.

In broad terms, to compare the influence of pavements on traffic noise levels, the measurement conditions need to be relatively comparable, especially the nature of the traffic over the pavement under test. Since traffic cannot be controlled, two measurements can be normalized with the CTIM test method using common modeled traffic noise levels. Then after normalization, the influence of the two pavements on traffic noise is more comparable.

Overview of AASHTO Standard Test Method TP 99

The AASHTO Standard Test Method TP 99 method includes microphones positioned adjacent to the roadway to capture the equivalent-average sound pressure level over a number of relatively short time-blocks (5 to 15 minute blocks). Additionally, vehicle counts, vehicle categories and traveling speed of vehicles must be observed during the course of each time-block; several options for speed measurement are available.

Time-blocks with interfering noise events are discarded. Traffic noise modeling software is used to predict the difference in noise levels for the remaining valid time-blocks, based upon the traffic observations for each time-block. The reference level for normalization is the equivalent-average modeled sound level from all time-blocks in the sample set. The difference between the reference and each modeled time-block noise level is then applied to the measured time-block. The test result is simply the arithmetic mean of the normalized time-blocks in decibels.

When comparing two measurements, the reference for one measurement is used as the normalization reference for both test results.

Factors Contributing to Measurement Variability

The aim of this research is to generate statements of precision and bias in two measurement methods for the influence of road surface on traffic and vehicle noise. ASTM E177-10 defines precision as the closeness of agreement between independent test results, whereas bias is the difference between expected test results and an accepted reference value. The accepted reference value may be a theoretical value based upon scientific principles, an assigned value from some authority, or a consensus value of the scientific community. There are occasions where the accepted reference value cannot be known and therefore the full extent of bias cannot be known. The accepted reference value is often called the “true value,” though ASTM E177-10 discourages this usage. Precision and bias are products of measurement variability due to different factors. Factors contributing to measurement variability are grouped into the following categories:

- Environmental factors,
- Test site,
- Vehicle fleet variability, and
- Measurement instrumentation.

Environmental Factors

Atmospheric conditions may affect traffic noise level even very close to the roadway (Wayson and Bowlby 1990). The primary environmental factors which influence tire traffic noise propagation and tire-pavement noise generation are: temperature, humidity, wind speed, wind direction, and precipitation.

Background noise from adjacent noise sources such as other roadways and railways are also subject to variation in meteorological conditions. Therefore background noise may change substantially over time (Sandberg and Ejsmont 2002).

Temperature

Tire-pavement noise generation and propagation can be influenced by temperature including air temperature, road surface temperature, and tire temperature. Several researchers have found a trend of decreasing tire-pavement noise levels with increasing temperatures. The decrease of tire-pavement noise with increasing temperatures occurs for several reasons including: changes in tire hardness, pavement properties such as stiffness and joints expansion (CDOT 2012).

The strength of the temperature effect on tire-pavement noise generation varies by temperature measurement medium, vehicle type, and pavement type. The rate of decrease in tire-pavement noise level has been measured to exceed 0.1 dBA/degrees Celsius (°C) for various pavement types. For cars, the influence of temperature on sound level data is greater for dense-graded asphalt pavements than for porous or open-graded pavements and is greater for porous or open-graded asphalt pavements than for cement concrete pavements. Recent research found variations in the decrease of tire-pavement noise levels ranging from 0.06 dBA/°C for porous pavements to 0.1 dBA/°C for pavements with higher density surfaces. The same study found no temperature effect on cement concrete pavements (Anfosso-Lédée and Pichaud 2007; Rochat 2010).

Changes in air temperature, for measurements performed between 5°C and 30°C can result in sound pressure level variations of up to 3 dBA for a same tire – road configuration (Anfosso-Lédée and Pichaud 2007). Without applying temperature corrections, there may be error due to temperature variations which can affect comparisons among data sets, whether the comparison is among different sites/pavements or whether it is the same site/pavement over time (Rochat 2010). To account for the air temperature conditions many tire-pavement noise-measurement standards and groups recommend the inclusion of a temperature correction which attempts to normalize measurement results to a reference temperature. The application of various temperature corrections to wayside measured data shows that it is possible to reduce error related to temperature variations, but one should do so cautiously (Rochat 2010). It should be noted that TP 98 and TP 99 do not include a temperature correction, only temperature bounds.

Temperature can also affect sound propagation by its effect on atmospheric attenuation and the effect of temperature gradients on diffraction of sound waves. Temperature effects on sound propagation are generally larger at greater distances from the source. Generally, the short distances for SIP and CTIM measurements will not be significantly affected by either atmospheric attenuation or temperature gradients. However, temperature gradients over pavement surfaces can become fairly significant. Furthermore, Sandberg and Ejsmont (2002) cite modeling of extreme temperature gradients which has shown reductions up to 10 dB in the 3150 Hz band at distances as short as 7.5 meters, the standard distance of SIP measurements. The authors continue to say an effect of this magnitude has not been observed in practice, but nonetheless suggest using the difference between pavement temperature and air temperature at the microphone position as an indicator of the magnitude of temperature gradients between the pavement surface and the microphone height (Sandberg and Ejsmont 2002).

Humidity

Atmospheric attenuation, or the attenuation of sound due to atmospheric absorption, depends strongly on the frequency of the sound, ambient temperature and relative humidity in air (ISO 1996). Higher frequency sounds are attenuated at a higher rate than lower frequencies. The maximum absorption of sound by air is observed for air with a relative humidity around 10 to 20 percent (Sandberg and Ejsmont 2002). Wayside traffic noise measurement performed under varying atmospheric conditions, including humidity, will results in varying noise levels due to atmospheric attenuation.

Wind

Wind speed, direction and wind shear influence the propagation of sound. Deviations in noise levels due to atmospheric refraction have been measured at 6.6 dB at distances of 125 feet from roadways (Wayson and Bowlby 1990). Larger differences in traffic noise levels, due to atmospheric refraction, have been measured at greater distances from roadways. These deviations in noise level are more apparent in microphone positions close to ground.

Wind can also be a significant influence on outdoor acoustical measurement including wind-generated vortices or noise produced as air moves past the microphone's protective grid (Harris 1998). Wind-induced noise may also result from surrounding vegetation. To limit the influence of wind induced noise and wind related propagation effects on measurement results TP 98 and TP 99 limit measurements to wind speeds of less that 5 meters per second (m/s) regardless of wind direction. Both standards also require the use of a wind screen to limit wind induced noise.

Precipitation

Pavement moisture due to precipitation can cause a measureable increase in traffic and vehicle noise level. This increase in noise is due to the incidence of splash at the tire – road surface interface. The increase in vehicular noise level due to pavement moisture differs dependent on tires type, road surface, travel speed, and water depth. Wayside measurements of vehicle pass-byes have recorded increases in sound level varying from 1 to 10 dBA between wet road surface and the same road surfaces when dry. Typically increases in noise due to precipitation occurred in the high frequency bands above 2 kHz (Nelson and Ross 1981; Underwood 1981; Leasure 1975). TP 98 and TP 99 limit the influence of precipitation on the measurement by requiring the test section to be dry.

Site Factors

Test site factors can also influence measured vehicle and traffic sound levels. In particular, the presence of reflective surfaces and the ground between the road surface and microphone locations can cause differences in measured sound levels. Differences in road surface within a roadway test section may also influence measured sound levels. Section 7 of TP 98 and TP 99 describe site selection requirements which minimize the influence of site factors on measured sound levels.

Ground Absorption

Sound propagation between a road surface and microphone location can be greatly influenced by the presence of reflecting surfaces and ground absorption. Ground attenuation is the attenuation of sound due to ground effects resulting from sound reflected by the ground surface interfering with sound propagating direct from the source to the receiver (ISO 1996). The term hard ground refers to ground surfaces which are acoustically reflective and have a low porosity such as non-porous pavement, water and ice. Soft ground refers to a ground surface which is acoustically absorptive and porous. Mixed ground refers to a propagation path which includes both hard and soft areas (Harris 1998).

The influence of ground absorption on measured sound levels increases with distance. Even between small distances differences in sound level in excess of 1 dBA have been measured between hard and soft ground surfaces. Differences in A-weighted sound levels between microphone distances of 7.5 and 15 meters, as specified in the SIP measurement standard, can vary 1.5 dB when performed over acoustically hard ground versus soft ground (Sandberg and Ejsmont 2002).

The CTIM measurement procedure uses a microphone mounted 50 feet from the centerline of the nearest travel lane at 12 feet above the surface of the pavement; this height is intended to reduce the effect of ground reflections. The SIP measurement procedure allows for either the same microphone position or a position 25 feet from the centerline of the nearest travel lane at 5 feet above the surface of the pavement; at this position the microphone is usually not far from the shoulder of a typical highway so most of the ground between the vehicle and the microphone is pavement except for the area closest to the microphone position. For comparison, the international version of the statistical pass-by measurement procedure (ISO 11819-1) requires a microphone position at 7.5 meters from the centerline of the nearest travel lane and 1.2 meters above the plane of the pavement, roughly comparable to the 25 feet SIP microphone position. The ISO standard states the measurement should essentially be representative of free field conditions (free of reflective surfaces), specifically the ground surface within 3.75 meters of the microphone position should be significantly absorptive and within 25 meters of the microphone position free from reflecting surfaces other than the ground is usually adequate "to ensure that approximate free field conditions exist."

Reflective Surfaces

Reflective surfaces not directly associated with the sound source of interest, in this case traffic and vehicle noise, can contribute significantly to measured sound levels. The affect of reflective surfaces on measured sound levels is dependent on the location of the reflective surface relative to the source and microphone position and the size of the surface. The CTIM and SIP measurement standards aim to decrease in influence of reflective surfaces by requiring that sites consist of an open space free of large sound-reflecting surfaces.

Alignment (Straight and Level)

Both measurement procedures require test sections to be "essentially level and straight." This promotes constant-speed driving, without acceleration or maneuvering which could increase vehicle power train noise. Furthermore, compared to a time-integrated measurement of a straight alignment, a curved alignment will elevate or reduce time-integrated measurements, depending upon whether the measurement position is inside or outside the curve, respectively.

Pavement Specimen

As described elsewhere in this report, differences in tire-pavement noise have been measured within a single test section and similar test sections constructed on the same project. Variability within test sections results in differences in measured sound levels at the source, measured using OBSI measurements, and at a distance from roadways. Statistical pass-by measurements have recorded differences in sound level of up to 1.6 dB between two pavement sections constructed on the same project. The noise variability observed within sections could be found with all road surface textures to varying degrees (Herman and Ambroziak 1999).

Fleet Variability

The traffic present during the test period, the vehicle sample set or the test fleet can contribute to variability in the CTIM and SIP measurement methods. Both the CTIM and SIP measurement methods rely on the collection of noise measurement data from an uncontrolled sampling of vehicles. The sample size therefore must be large enough to achieve a desired level of significance. Fleet variability such as vehicle speed, tire types, vehicle loads, power train noise, and even driving behaviors may influence the variability in CTIM and SIP measurement methods. Table A-1 summarizes the change in tire-pavement noise level due to various fleet parameters.

Table A-1. Increases in Tire-Pavement Noise Levels with Changes in the Parameters Studies

Parameter	Change from a to b			Increase in Peak Pass-by Noise, dBA
Vehicle speed[a]	30	to	100 km/h	16-20
Load	5.9	to	13.2 tonnes	1-7
Tread pattern	Smooth	to	Traction	~3
Tire construction	Radial	to	Cross-ply	~3
Tread belt material[a]	Steel	to	Rayon	~3
	Steel	to	Nylon	~5
Tread material	Natural rubber	to	High hysteresis rubber	~2
Wear	0	to	25 percent	1-5

Source: Underwood, M. C P. Lorry Tyre Noise. Rep. no. TRRL Laboratory Report 974. Crowthorne, Berkshire: Transport and Road Research Laboratory, 1981.

Pass-by speed normally 100 kilometers per hour (km/h)

[a]not normalized to a pass-by speed of 100 km/h

The results in Table A-1 demonstrate that tire-pavement noise can be influenced greatly by various fleet factors.

Speed

As discussed previously, the travel speed of vehicles greatly influences the noise level, mainly due to tire/pavement noise interaction. Both AASHTO test methods TP 98 and TP 99 account for the effect of speed respectively with a linear regression procedure and with a normalization procedure. Therefore, the observations of speed have their own level of variability and consequently affect the resulting measurement and test results.

TP 98 and TP 99 state that vehicle speed shall be measured using a measuring instrument with an accuracy of ±1 mph. Vehicle speed may be determined using a number of methods including pneumatic tubes, radar gun, stopwatch, light sensors, and video processing. In many instances speed measurement instrumentation may interfere with sound level measurements. For example, pneumatic tubes must be placed at a distance from the microphone position due to noise generation. Handheld radar gun within eyesight of drivers could also influence behavior and interfere with pass-by measurements at a continuous speed. Past tire-pavement sound studies have noted concerns surrounding vehicle speed measurements methods and instrumentation. For example, the Arizona Quiet Pavement Pilot Program determined speed using three different techniques: global radar, handheld radar gun, and video reduction. The global radar approach proved inadequate and was later discontinued. Speed measurement through video reduction was also noted as “suspect” (ADOT 2004).

Other Vehicular Noise Sources

Vehicular noise sources, aside from the influence of road surface, may also cause variations in measured sound levels. The traffic present during the test period is an uncontrolled sampling of vehicles. Test fleets on same road surface may have varying contributions of power-train noise and aerodynamic noise depending on the design and condition of the vehicle. Power-train noise may contribute substantially to measured sound levels depending on travel condition and vehicle type (Sandberg and Ejsmont 2002). Differences in vehicle sample sets between test periods may create variations in measurement results due to differences in the contribution of power-train noise.

Likewise, variations in aerodynamic noise between test fleets may also create variations in measurement results. Generally, influence of aerodynamic noise to total wayside noise is negligible at lower travel speeds. However, at higher travel speeds, above 80 kilometers per hour (km/h) for trucks and 120 km/h for automobiles, aerodynamic noise may become important (Sandberg and Ejsmont 2002). The test fleet in uncontrolled measurement data will include vehicles of varying aerodynamic design. Tested vehicles may also have roof racks and other unique features influencing the aerodynamic noise produced by the vehicle.

Tire Noise

Tire-pavement noise will vary based on tire type, tire inflation, load, and wear. Table A-1 shows that tire-pavement noise levels may differ up to 7 dB due to changes in vehicle load. Alternate tests performed by General Motors found an increase in pass-by noise levels up to 15 dB due to changes in load. Tread design and road texture are also major factors in the production of tire noise (Leasure 1975). Measurement methods aimed at the measurement of tire-pavement noise direction, such as OBSI, use a standardized tire to eliminate the variability due to tire type. Both the CTIM and SIP measurement procedures rely on the measurement of an uncontrolled fleet of vehicles; therefore, tire type is not standardized. Among the test vehicle fleet variations in load, tire type, tire wear, and inflation pressure are likely to occur resulting a variation in measured sound level.

Current Knowledge of Measurement Uncertainty

The literature search yielded some knowledge of measurement uncertainty for test methods similar to SIP, but little knowledge of measurement uncertainty for test methods similar to CTIM.

SIP Measurement Uncertainty

Measurement uncertainty associated with ISO 11819-1 has been quantified within the measurement standard. Systematic errors due to instrumentation are ±1 decibel (dB) for instrumentation systems specified within the standard. Table A-2 summarizes the expected random errors for single-site measurement.

Table A-2. ISO 11819-1 Expected Random Errors in A-weighted Sound Pressure Level

Vehicle Class	Standard Deviation for Individual Vehicles Around, L_{veh}	95% Confidence Interval Around L_{veh}
Cars	1.5 dB	0.3 dB
Heavy Vehicles, dual-axle	2.0 dB	0.7 dB
Heavy Vehicles, multi-axle	2.0 dB	0.7 dB

The confidence intervals for L_{veh} assume that the number of vehicles is 100 cars and a total of 80 heavy vehicles. The error of the calculated SPBI will be a combination of the errors presented in Table A-2 according to the chosen weighting factors.

Repeatability according to ISO 5725-1 is better than 1.0 dB. Variation in vehicle fleet composition can also create variation in measured sound levels. According to ISO 11819-1 the error due to variations in fleet composition may amount to 0.3 to 0.8 dB within a typical European country (ISO 1997).

A French study of controlled pass-by and statistical pass-by measurements performed in 1994 demonstrated the repeatability of the measurement methods to be approximately 1 dBA. The same study determined reproducibility to be between 2.5 and 5 dBA. A more recent study was performed to determine the repeatability and reproducibility associated with the controlled pass-by and statistical pass-by measurement methods, per the French standard, Association Française de Normalisation (AFNOR) S 31119. Statistical pass-by methods described in AFNOR S 31119-2 are similar to ISO 11819-1 procedure. It should be noted that the difference in reference speeds, vehicle categories and reference levels may bring significant discrepancies in the final results. The French study of controlled pass-by and statistical pass-by measurements performed in 1996 demonstrated the repeatability of the measurement methods to be approximately 1 dBA. The same study determined reproducibility to be approximately 1.5 dBA (Anfosso-Lédée n.d.).

CTIM Measurement Uncertainty

A literature search yielded no specific information on the uncertainty associated with the CTIM test method. In lieu of specific information, the systematic error due to instrumentation, calibration imperfections and deviations from ideal acoustical propagation conditions is ±1 dB for the precision instrument specified in AASHTO TP 99.

Summary

The project team conducted a literature review in support of the development of precision and bias statements for AASHTO Standard Test Methods TP 98, and TP 99. The team collected and evaluated information relevant to the measurements of the influence of road surfaces on traffic and vehicle noise and associated precision and bias. The background research included topics related to vehicle traffic noise, the effect of pavement on traffic noise, the measurement of the effect of pavement on traffic noise, and factors contributing to measurement variability with respect to wayside noise.

While extensive research has been performed concerning the individual factors which influence tire-pavement noise and wayside measurement of traffic noise, few studies addressing the precision repeatability and reproducibility of the CTIM and SIP test were found. The limited studies of precision and bias for statistical pass-by measurement of tire-pavement noise have utilized alternate international standards which vary from the current AASHTO test procedures. The differences in instrumentation requirements, vehicle categories, number of pass-by events and other variations are likely to influence the precision and bias of the test method.

References

[ADOT] Arizona Department of Transportation. 2004. *Quiet Pavement Pilot Program*. Rep. Vol. Progress Report No. 1. Arizona Department of Transportation.

Anfosso-Lédée, F. No date. "SILVIA PROJECT REPORT: A Former LCPC Experimental Campaign about Repeatability and Reproducibility of SPB and CPB Measurement Methods." Rep. no. SILVIA-LCPC-002-00-WP2-170403.

Anfosso-Lédée, Fabienne, and Yves Pichaud. 2007. "Temperature Effect on Tyre-Road Noise." Applied Acoustics 68: 1–16.

Berengier, M. C., M. R. Stinson, G. A. Daigle, and J. F. Hamet. 1997. "Porous Road Pavement: Acoustical Characterization and Propagation Effects." Journal of Acoustical Society of America 101, no. 1: 155–62.

Brosseaud, Y. and F. Anfosso-Lédée. No date. "Review of Existing Low Noise Pavement Solutions in France." Rep. no. SILVIA-LCPC-011-01-WP4-310505.

[CDOT] Colorado Department of Transportation. 2012. DTD Applied Research and Innovation Branch. Report No. CDOT-2012-5. *Tire Pavement and Environmental Traffic Noise Research Study*. By R. O. Rasmussen and R. C. Sohaney.

Dick, E. B., B. I. Izevbekhai, and T. Casey. 2010. "Influence of Pavement on Traffic Noise – Statistical Pass-By Measurement of Traffic on Several Interstate Pavements." Rep. no. MN/RC 2010-40.

Donavan, P. R. 2004. "Vehicle Exterior Noise," Handbook of Noise and Vibration Control. Editor Malcolm Crocker, John Wiley and Sons. Accessed: http://www.fhwa.dot.gov/environment/noise/tire_pavement_noise/fhwa_policy/tpnwks2004/wrkshop01.cfm.

Donavan, P. R. 2011. Initial and Long-Term Evaluation of the Tire-Pavement Noise Produced by Various Portland Cement Concrete Surface Textures.

Donavan, P. R. 2012. SPR-577 Quiet Pavement Pilot Program: Progress Report 3. Rep. no. FHWA AZ-12-577, Phoenix, AZ: Arizona Department of Transportation.

Donavan, P. R., and D. M. Lodico. 14 July 2011. "Measuring Tire-Pavement Noise at the Source: Precision and Bias Statement". Final Report, NCHRP Project 1-44(01). Illingworth & Rodkin, Inc., Petaluma, CA. http://onlinepubs.trb.org/onlinepubs/nchrp/docs/nchrp01-44(01)_FR.pdf.

Gurovich, Y. A., K. J. Plotkin, D. H. Robinson, W. K. Blake, and P. R. Donavan. 2009. *NCHRP Report 635 Acoustic Beamforming: Mapping Sources of Truck Noise*. Transportation Research Board of the National Academies, Washington, D.C.

Harris, C. M. 1998. Handbook of Acoustical Measurements and Noise Control, 3rd ed. McGraw Hill.

Harvey, J. T., and E. Kohler. 2011. "Quieter Pavement Research: Concrete Pavement Tire Noise." Rep. no. CA12-1200 A.

Herman, L. A., and M. J. Ambroziak. 1999. "Effects of Pavement Type on Traffic Noise Levels." Rep. no. FHWA/OH-99/012.

Leasure, Jr., W. A. 1975. "Tire-Road Interaction Noise." The Journal Of The Acoustical Society of America 58, No. 1: 39–50.

Morgan, P. (ed.). 2006. Guidance Manual for the Implementation of Low-Noise Road Surfaces. Rep. no. FEHRL Report 2006/02. FEHRL.

Nelson, P. M. and N. F. Ross. 1981. Noise from Vehicles Running on Open Textured Road Surfaces. Rep. no. TRRL Supplemental Report 696, Crowthorne, Berkshire: Transport and Road Research Laboratory.

Peeters, B., and A. Kuijpers. 2008. "The Effect of Porous Road Surfaces on Radiation and Propagation of Tyre Noise." Acoustics '08 Paris: 155th ASA Meeting. Paris, France.

Rasmussen, R. O., R. J. Bernhard, U. Sandberg, and E. P. Mun. 2007. "The Little Book of Quieter Pavements." Tech. no. FHWA-IF-08-004.

Rasmussen, R.. O., R. Sohaney, and P. Wiegand. 2011. *Variability and Visualization of Tire-Pavement Noise Measurements*. Concrete Pavement Surface Characteristics Program.

Rasmussen, R. O., P. D. Wiegand, G. J. Fick, and D. S. Harrington. 2012. "How to Reduce Tire-Pavement Noise: Better Practices for Constructing and Texturing Concrete Pavement Surfaces." Rep. no. DTFH61-06-H-00011 Work Plan 7.

Rochat, J. L. 2009. "Developing U.S. Wayside Methods for Measuring the Influence of Road Surfaces on Traffic Noise." Inter-noise 2009: Innovations in Practical Noise Control. Washington: Institute of Noise Control Engineering.

Rochat, J. L. 2010. "Investigation of Temperature Correction for Tire/Pavement Noise Measurements." Rep. no. FHWA-HEP-11-005.

Rochat, Judith L., D. Lodico, P. Donavan, and R. O. Rasmussen. 2012. "Overview and Application of the Continuous-Flow Traffic Time-Integrated Method (CTIM) for Determining the Influence of Road Surfaces on Traffic Noise." Internoise 2012: Quieting the World's Cities. New York City.

Rochat, J. L., D. R. Read, and G. G. Flemming. 2010. "Caltrans Thin Lift Study: Effects of Asphalt Pavements on Wayside Noise." Rep. no. CA 10-0146.

Sandberg, U. and J. A. Ejsmont. 2002. Tyre/road Noise: Reference Book. Kisa, Sweden: INFORMEX.

Sandberg, U., J. A. Ejsmont, L. Herman, and J. Withers. 2002. Effectiveness of Tire/Road Noise Abatement Through Surface Retexturing by Diamond Grinding for Project SUM-76-15.40 2005.

Tire-Pavement Noise Research Consortium. Jan. 2011. Questions and Answers About Quieter Pavements. Developed under TPF-5 (135). Publication no. January 2011.

U.S. Bureau of the Census. September 15, 2004. "Table HH-6. Average Population Per Household and Family: 1940 to Present." http://www.census.gov/population/socdemo/hh-fam/tabHH-6.pdf (accessed April 5, 2013).

Underwood, M. C P. 1981. Lorry Tyre Noise. Rep. no. TRRL Laboratory Report 974, Crowthorne, Berkshire: Transport and Road Research Laboratory.

Wayson, R. L. and W. Bowlby. 1990. "Atmospheric Effects on Traffic Noise Propagation." Transportation Research Record 1255. TRB, National Research Council, Washington, D.C. pp. 59–71.

Bibliography

Acoustical Society of America. 11 Jan 2013. *Proceedings of Meetings on Acoustics*. 159th Meeting of the Acoustical Society of America / NOISE-CON 2010, Baltimore, Maryland. 040008th ed. Vol. 9. N.p.: Acoustical Society of America, 2010. N. pag. *ASA Digital Library*. Web. http://asa.aip.org.

Ferragut, T., R. O. Rasmussen, P. Wiegand, E. Mun, and E. T. Cackler. 2007. ISU-FHWA-ACPA Concrete Pavement Surface Characteristic Program Part 2: Preliminary Field Data Collection. Rep. no. DTFH61-01-X-00042 (Project 15, Part 2).

[FHWA] U.S. Department of Transportation Federal Highway Administration. 4 Feb 2013. *FHWA-HEP-10-025Highway Traffic Noise: Analysis and Abatement Guidance*. Web. http://www.fhwa.dot.gov/environment/noise/regulations_and_guidance/analysis_and_abatement_guidance/revguidance.pdf.

FHWA. 2011. Highway Traffic Noise: Analysis and Abatement Guidance. Rep. no. FHWA-HEP-10-025.

Herman, L. A. and J. M. Withers. 2005. "Effectiveness of Tire/Road Noise Abatement Through Surface Retexturing by Diamond Grinding for Project SUM-76-15.40." Rep. no. FHWA A/OH-2005/009.

Lee, C. S. Y., and G. G. Fleming. 1996. "Measurement of Highway-Related Noise." Rep. no. FHWA-PD-96-046, 1996.

Procedures for Abatement of Highway Traffic Noise and Construction Noise, 75 23 CFR 772 §772.15. 2010.

Rasmussen, R., S. Garber, and R. Whirledge. No date. "Session 1pNCd: Noise."

Rochat, J. L., A. L. Hastings, D. R. Read, and M. Lau. 2012. "FHWA Traffic Noise Model (TNM) Pavement Effects Implementation Study: Progress Report 1." Rep. no. DOT-VNTSC-FHWA-12-01.

APPENDIX B

Description and Statistical Analysis of Limited Pilot Testing

Introduction to Pilot Test

The research team conducted limited preliminary tests with TP 98 and TP 99 and performed a pilot repeatability and reproducibility (R&R) study. These statistical evaluations serve as the basis for preliminary estimates of precision and bias and for a more extensive test program. The results of the R&R study were used as an *a priori* variance component estimate, or planning values, to perform simulations of experimental design plans. There were two simulations: a Power and Estimation Simulation, and a Signal-Noise-Ratio Simulation. The purpose of these tests is to measure the variance of the test methods. The actual loudness or quietness of any particular pavement is less relevant than the differences between a multitude of test results.

The pilot R&R study, in particular, is designed to provide guidelines for developing the more comprehensive R&R study to be conducted in an expanded R&R study (in the present research, referred to as the full R&R study). The objective of the estimates from the pilot R&R testing is not to assess the adequacy of either of the methods, but to use the results to determine optimal experimental designs which meet satisfactory levels of precision within approved budgets and timelines. Example candidate experimental design plans are 4 operators, 6 recordings and 2 replicates or 5 operators, 5 recordings and 2 replicates. The results of the expanded/full R&R study will be used to assess the adequacy of both methods.

The research team gathered test results from the following experimental design for both test methods:

- 2 operators qualified to perform acoustic measurements
- 2 process samples which are calibrated audio and synchronized video to present the operator with a repeatable measurement opportunity of a particular pavement during a period with its own environmental conditions and vehicle fleet
- 2 replicates – each operator will read each sample twice; the order that samples will be processed will be randomized and will be unknown to the operator

The premise of the test method evaluation is to use recordings of calibrated audio and synchronized video to evaluate repeatability where each operator is presented with the same measurement subjects. Each recording presents the operator with a particular pavement specimen under reproducible measurement conditions. The measurement conditions for each recording are comprised of the environmental conditions present during rerecording and common the vehicle sample sets which drove by.

Test Procedures

The research team gathered field recordings with instrumentation microphones connected to a digital audio recording system. At the same time, a video camera recorded all the traffic at the measurement position. To facilitate accurate reproduction from the digital audio recordings, recordings were calibrated. Calibration signals were recorded at the beginning and end of the measurement and at approximately hourly intervals. These calibration signals were used normally with the measurement instrumentation in the office, just as if the calibrator were connected directly to the instrumentation. The calibrated audio and synchronized video was presented to noise analysts (operators), each with a digital playback system connected to an analyzer. Each operator played the audio and video (as if he/she were in the field) and initiated calibrated measurements using the analyzer.

The recording and playback systems have been qualified using procedures which conform to and expand on the 2005 version of ANSI S1.13, §8.7, to ensure that the microphone signal playback in the office precisely replicates the original acoustic signal in the field. See the main body of this report for additional discussion regarding the use of calibrated audio and synchronized video in the present research.

Process Samples/Test Sites

The research team maximized the volume of available data by partially using traffic recordings for one measurement site from the team's library of recordings gathered for other research. In addition, the research team gathered new calibrated audio and synchronized video recordings for another measurement site. The prior recording used for this research was from a site in Minnesota (MN Site); the new recording was from a site in Iowa (IA Site). Reusing existing recordings has the great advantage of saving the cost of field work. It also has the detriment of imprecise environmental data.

Both the previously-gathered field recordings (MN Site) and newly-gathered field recordings (IA Site) need to conform to the field procedures, microphone requirements and measurement site environment requirements of TP 98 and TP 99. The previously-gathered data collected at the MN Site followed the measurement standard ISO 11819-1; the field procedures therein very closely resemble the field procedures for TP 98 and TP 99. The requirements are nearly identical between TP 98 and TP 99 for the Apparatus (Section 6), Selection of Test Sites (Section 7), and Meteorological Conditions (Section 9). Differences between TP 98 and TP 99 are exhibited in the Microphone Positions (Section 10). Other sections are not necessarily relevant to gathering field recordings, though they certainly have bearing on the measurement.

These field recordings include some minor and some other very important deviations from the standards. While deviations from the standard could introduce a bias to the result, the newly-gathered field recordings were performed to match the same deviations as the previously-gathered field recordings. Any bias presented by these deviations should be nearly identical and therefore should have negligible effect on the variance in the pilot R&R study. The results will only reflect differences between road/traffic, differences between operators, and differences between iterations. This is the data necessary to satisfactorily design the full R&R experiment. The research team believes that deviation in field collection methods for the pilot study will have a negligible effect on the variance in repeatability and reproducibility; the actual level resulting from each test is not as important as the differences in levels between tests. The deviations are a small sacrifice in engineering rigor, but that is to keep a satisfactory level of statistical rigor. Specific deviations are discussed in the following sections. (All data and field procedures for the full R&R study will strictly adhere to requirements of TP 98 and TP 99. Minor deviations from prescribed field collection methods will be less tolerable in the full R&R study. Intentionally deviating from the test method is only necessary in the pilot R&R study to prevent inconsistencies between test procedures which would mask other variance components.)

A minor deviation from TP 98 and TP 99 was the meteorological conditions were not measured hourly on-site during the field data collection. Wind speed was measured on-site using a handheld anemometer, and temperature was measured at the nearest airport. The full R&R study collected hourly metrological data at the measurement site.

Measurement Test Site in Minnesota

The location for the MN test site was previously recorded in the late summer of 2009 with the express consent of the state department of transportation. A site was selected on a straight, flat section of I-94, between mile markers 194 and 195. This was a Portland cement concrete (PCC) pavement with a burlap-drag surface finish. The measurement location was on the eastbound traffic, with microphone positions on the south side of the travel lanes. To the south of the microphone positions was an open field, and across the interstate to the north there was a frontage road then a large parking lot. The nearest reflecting surface was over 300 feet away (a road sign oriented parallel with traffic). The MN Site conformed to Section 7 of TP 98 and TP 99. Traffic data at the MN Site were gathered using microwave sensors. This primarily provided speed data, since traffic volumes and vehicle categories were determined with the video playback.

Measurement Test Site in Iowa

The location for the new recording was determined by weather forecasts as much as other site requirements. A location in Story County, IA, offered a location that met the test site requirements but also avoided the precipitation and a cold-front that was beginning to move south; this offered a one-day window of acceptable meteorological conditions at the selected IA Site in the late fall of 2012. Due to the short notice, only an informal arrangement could be made with the County Engineer to work in the right-of-way; work conformed to all verbal instructions of the engineer. The IA Site was specifically on County Road E41, a two-lane road also known as Lincoln Highway, between Ames and Nevada. The microphone positions were on the south side of the travel lanes and matched the distances and heights used for the MN site. The test site was surrounded by flat, open cropland. The nearest reflecting surface was over 350 feet away (a storage shed to the northeast of the microphone positions). The IA Site conformed to Section 7 of TP 98 and TP 99. Traffic data at the IA Site were gathered only via video playback, including speed data using video image analysis.

Microphone Positions

The requirements for Microphone Positions (section 10) differ between TP 98 and TP 99, and also differ from ISO 11819-1, the procedure followed for MN Site field data collection. As stated previously, the IA Site field data collection followed the same procedures as the MN Site to ensure consistency between recordings before they're presented to the test operators. Table B-1 summarizes the actual microphone placements and compares to the specified placements from TP 98 and TP 99.

Table B-1. Microphone Placements

Microphone positioning		Distance from nearest lane's centerline (ft)	Height above pavement (ft)
Actual Placement	Used for TP 98	25	4
	Used for TP 99	50	4
TP 98 (SIP)	Primary option 1[(a)]	25	5
	Primary option 2[(a)]	50	12
	Additional location[(b)]	50	5
TP 99 (CTIM)	Preferred location[(c)]	50	12
	Alternate location[(d)]	Between 50 and 100	$\geq$ 12

[(a)] TP 98 offers two possible primary microphone positions; use of only one is also acceptable

[(b)] Recommended due to compatibility with FHWA-PD-96-046 and historical data; not mandatory. Refer to TP 98 for details.

[(c)] Preferred by TP 99; not mandatory.

[(d)] These are alternate measurement locations allowed by TP 99. Microphone must be at least 5 feet above the ground, and at least 12 feet above the center of the near travel lane.

The research team used the position 25 feet away and 4 feet above the pavement for the TP 98 (SIP) procedure in the pilot study. For the TP 99 (CTIM) procedure in the pilot study, the research team utilized the microphone position 50 feet away and 4 feet above the pavement. As the table shows, these positions deviate from the specifications in TP 98 and TP 99. To reiterate the previous discussion on deviation from the test method, this intentional deviation is a small sacrifice in engineering rigor for the limited preliminary testing, which allows more time in the pilot study to be spent on a statistically rigorous approach. Variability between sites, vehicles, operators, and replicates was be largely unaffected by the deviation in microphone height, as long as the microphone height was consistent throughout the pilot study.

Microphones conformed to the Type 1 precision requirement in Section 6 of TP 98 and TP 99, but utilized free-field microphones instead of the prescribed random-incidence or pressure-response microphones. All three microphone types only exhibit difference at frequencies 4000 Hz and above. , however there are commonly-used correction factors between the microphone types (the instrumentation used by the research team has a selectable option to automatically apply the correction). Because each measurement uses the same microphones with the same frequency responses between recordings, the R&R experiment will not be affected.

Random-order Presentation

Each operator viewed each of the two recordings twice (two replicates), which yielded 8 test results to analyze for one test method. The operators reused the same recording for each of the two test methods (TP 98 and TP 99). The presentation of recordings was randomized for the experiment. Order of presentation was randomized using a sequence of random numbers generated with the R software package for statistical computing using the “Mersenne-Twister” random-number generator and a seed value 7465915. Table B-2 and Table B-3 show the order of presentation for each operator.

Table B-2. Operator 1 Order of Presentation

Test ID	Test Method	Recording	Test ID	Test Method	Recording
1-1	SIP	MN	1-2	CTIM	MN
1-3	SIP	IA	1-4	CTIM	IA
1-5	SIP	IA	1-7	CTIM	MN
1-6	SIP	MN	1-8	CTIM	IA

Table B-3. Operator 2 Order of Presentation

Test ID	Test Method	Recording	Test ID	Test Method	Recording
2-2	SIP	IA	2-1	CTIM	IA
2-5	SIP	MN	2-3	CTIM	MN
2-6	SIP	IA	2-4	CTIM	MN
2-7	SIP	MN	2-8	CTIM	IA

To further mix the testing, the operators could optionally perform the tests in the order of the test ID number. However, the operators were instructed to at least strictly adhere to the order of presentation within each measurement method.

Data Structure for Test Method TP 98

Essentially, each sample (there are only two) is processed by two different operators and each operator processes the each sample twice. There are a total of 8 tests in our design. The measurement procedure is the same as described in the main body of this report. The final measurement used as input into the pilot R&R study is the SIPI for automobiles (the pilot R&R study only evaluated one vehicle category). Table B-4 presents the final measurements per test sample. The average of the test measurements across the 8 tests is 3.7 with the lowest value being 2.9 and the highest value being 4.4.

Table B-4. R&R Dataset for Test Method TP 98

Sample Class	Sample	Operator	Measurement
MN	1	1	3.0
IA	2	1	4.4
IA	2	1	4.2
MN	1	1	3.4
IA	2	2	4.1
MN	1	2	3.0
IA	2	2	4.2
MN	1	2	2.9

Data Structure for Test Method TP 99

As with the pilot test for Test Method TP 98, each of the two samples is processed by two different operators and each operator processes each sample twice. There are a total of 8 tests in the design. The measurement procedure is the same as described in the main body of this report. Table B-6 lists the final measurements in decibels. The average of the 8 test measurements is 71.2 dBA. The lowest recorded test value is 50.7 dB while the highest is 87.5 db.

Table B-5. R&R Dataset for Test Method TP 99

Sample Class	Sample	Operator	Measurement
MN	1	1	85.1
IA	2	1	62.6
MN	1	1	85.2
IA	2	1	67.0
IA	2	2	46.9
MN	1	2	84.4
MN	1	2	87.5
IA	2	2	50.7

Pilot Repeatability and Reproducibility Study

Overall, the purpose of an R&R study is to determine whether or not the measurement procedure variation is small relative to the variation that can be attributed to the test objects or samples. The variation arising from the differences in the samples is called process variation. If measurement variation is relatively large, then the procedure must be improved before it can be useful for monitoring purposes. See the main body of this report and also Appendix C for additional discussion of R&R studies.

The research team used NCSS (Hintze 2009) to generate estimates of the variance components and their levels of precision. The following section describes the data used to generate the variance component estimates and the estimates themselves. The estimated variance components will then be used to assess candidate experimental designs.

Pilot R&R Study Results for Test Method TP 98

Using the NCSS software, the variability across the 8 results in Table B-5 can be broken down into their variance components. Table B-5 lists the results. These variance components are preliminary and are used as planning values to design robust experimental plans for the full R&R study.

Table B-6. Variance Component Estimates for Test Method TP 98

Term	Symbol	Estimate
Process	$\hat{\sigma}_P^2$	0.678079
Operators	$\hat{\sigma}_O^2$	0.017889
Interaction	$\hat{\sigma}_{PO}^2$	0
Reproducibility	$\hat{\sigma}_O^2 + \hat{\sigma}_{PO}^2$	0.017889
Repeatability	$\hat{\sigma}_\varepsilon^2$	0.020306
R and R	$\hat{\sigma}_O^2 + \hat{\sigma}_{PO}^2 + \hat{\sigma}_\varepsilon^2$	0.038195
SNR	$\sqrt{\hat{\delta}}$	5.958705

Note: the hat above the sigma symbol indicates the value is an estimate and not the underlying population value of the statistic.

The *a priori* variance component estimate for the interaction variance component is zero; however, the true value lies in the interval of [0, 0.2689) at a 95% confidence level.

Pilot R&R Study Results for Test Method TP 99

Using the NCSS software, these numbers generated the variance component estimates in Table B-7. These variance components are preliminary and are used as planning values to design robust experimental plans for the full R&R study.

Table B-7. Variance Component Estimates for Test Method TP 99

Term	Symbol	Estimate
Process	$\hat{\sigma}_P^2$	378.6646
Operators	$\hat{\sigma}_O^2$	0
Interaction	$\hat{\sigma}_{PO}^2$	67.96368
Reproducibility	$\hat{\sigma}_O^2 + \hat{\sigma}_{PO}^2$	67.96368
Repeatability	$\hat{\sigma}_\varepsilon^2$	5.495743
R and R	$\hat{\sigma}_O^2 + \hat{\sigma}_{PO}^2 + \hat{\sigma}_\varepsilon^2$	73.45942
SNR	$\sqrt{\hat{\delta}}$	3.210839

The table shows the *a priori* variance component estimate for the operator effect in the TP 99 R&R study is zero. The standard deviation of this estimate has a 95percent confidence interval of (0 - 85.3751). In order to simulate power test levels per candidate design, the *a priori estimate* for the variance component due to the effect of the operator must be non-zero. The team imputed the value by applying a scaled chi-square distribution with one degree of freedom to the confidence interval for the standard deviation of this variance component. By selecting the 50th percentile from this scaled chi-square distribution with one degree of freedom, the team obtained a probable estimate for the required *a priori* variance component.

Power and Estimation Simulation

As detailed in *Applied Linear Statistical Models* by Kutner et al. (2005), there are two approaches to determining adequate sample size for a designed experiment---the power approach and the estimation approach. Adequate sample size refers to the number of samples from the process (which in our case are the selected road segments selected during a particular time with particular vehicle fleets), the number of operators and number of replicates that define the experimental design plan.

In the power approach, the probability of correctly rejecting the null hypothesis of a zero variance component for a given experimental design plan can be computed given a planning value for the variance component. The planning value is an *a priori* estimate of the variance component, which for us is available from the pilot R&R study. (A conservative approach is sometimes employed in which the planning values for the process variance components are increased or decreased by 30 percent to allow for the uncertainty in the estimated values.) Typically, we seek an experimental design plan that has at least 70 percent or 80 percent power.

In the estimation approach, for a given experimental design plan, we use the planning values of the various variance components to compute the expected confidence limits that will be obtained for key measures of interest. We look for a design, as defined by the number of operators, number of recordings, and number of replicates, that provides confidence limits for the variance components for recordings and for operators that are sufficiently precise.

Power and Estimation Simulation Approaches

The power values and expected confidence limits do not exist in statistical tables and must be computed using Monte Carlo simulation methods. The research team developed a MATLAB code that carries out these computations for any numbers of recordings, operators, and replicates.

The first step in the design selection and sample size determination process for both methods of test TP 98 and TP 99 is to use their *a priori* variance component estimates, which in our case, are the variance component estimates from this pilot R&R study. For a candidate experimental design plan, we use these values to generate 10,000 random data sets for a measurement system which is characterized by these values. Example candidate experimental design plans are 4 operators, 6 recordings and 2 replicates or 5 operators, 5 recordings and 2 replicates. The research team ran these simulations using experimental plan combinations ranging from 2 to 6 operators with 2 to 6 recordings while keeping the number of replicates fixed at 2. (Adding another replicate increases testing costs and may not improve the precision of the final results. However, the research team ran the simulations under a three replicate scenario to assess if there is a benefit or not.)

The research team then analyzed each of the 10,000 data sets, obtaining estimates of all of the variance components, confidence intervals for the operator and the process (i.e., the selected sample) variance components, and signal-noise ratios (SNRs) and their respective confidence intervals. The research team also determined whether or not the hypotheses of zero variance components for operators and process was rejected or accepted for each data set. The percentage of the 10,000 tests that are rejected at the $\alpha = 0.05$

level of significance provides an accurate Monte Carlo estimate of the power of the tests. The research team used the MLS confidence intervals as detailed in Chapter 25 of Kutner et al. (2005)

In summary, each candidate experimental design plan had the following statistics tracked:

- Power values for the $\hat{\sigma}_P^2$ tests for recordings (Power P)
- Power values for the $\hat{\sigma}_O^2$ tests for recordings (Power O)
- Expected confidence limits for the $\hat{\sigma}_P$ of recordings (LL Sig P and UL Sig P)
- Expected confidence limits for the $\sqrt{\hat{\sigma}_O^2 + \hat{\sigma}_{PO}^2}$ of reproducibility (LL Reprod and UL Reprod).

The variance component planning values from the pilot R&R study were input into the Monte Carlo simulations to produce the desired evaluative measures per candidate plan. Then the research team evaluated the statistical merits of the candidate experimental plans which meet the recommended criteria.

Power and Estimation Simulation for Test Method TP 98

The variance component planning values from the TP 98 pilot R&R study were input into the Monte Carlo simulations to produce the desired evaluative measures per candidate plan. As shown in Table B-8, there are multiple designs which achieve at least 75 percent power for both tests. The possible candidate designs are as follows:

- 3 samples, 6 operators;
- 4 samples, with 4 to 6 operators;
- 5 samples, with 4 to 6 operators; and,
- 6 samples, with 3 to 6 operators.

The precision of the confidence intervals for both $\hat{\sigma}_P$ and reproducibility improves as the number of operators and samples increases. For example, fixing replicates at 2, the width of the expected confidence interval for $\hat{\sigma}_P$ for the design with 6 samples and 6 operators is only 6 percent of that which could be expected under a 2 sample, 2 operator design. There is a similar improvement in the precision for the reproducibility estimate as the width of the expected confidence interval under the 6 sample, 6 operator design is only 4 percent of that found with the expected confidence interval under the 2 sample, 2 operator design.

The precision of the confidence intervals for the recommended designs based on the power approach are all reasonable based on statistical judgment. The confidence interval ranges for both $\hat{\sigma}_P$ and reproducibility in the plans with 5 to 6 samples and 5 to 6 operators are of similar magnitude (differences are within 50 percent of their values) and provide the best precision given the observed *a priori* variance component estimates. Therefore, they are the optimal designs to implement for the full R&R study. Table B-8 presents the power results and expected confidence intervals under the two replicates scenario for Test TP 98. The yellow shading indicates power values greater than 0.7070

If we make an adjustment to the planning value of the variance component for the process and reduce it by one third and re-run the simulations, some of the same designs are recommended as found with the simulations using the original variance component estimates. These plans are highlighted in yellow in Table B-9. Based on the relative precisions of the expected confidence intervals, designs of 5 to 6 samples with the 5 to 6 operators are still acceptable designs.

Conversely, if we inflate the planning value for the process variance component by a third and re-run the simulations, all the same plans that are recommended using the original variance component estimates and the reduced versions are observed in this scenario with the addition of the design having 3 samples with 5 operators. Table B-10 shows the power results and expected confidence intervals under the two replicates scenario for Test TP 98 where the planning value for the process variance component has been inflated by one third of the recommended designs. The plans with greater than 70 percent power are highlighted in yellow below in Table B-10.

Table B-10 shows that once more, designs of 5 to 6 samples with the 5 to 6 operators offer the best potential for reliability and precision given that these designs meet the power test criterion and have the most precise confidence intervals

The team investigated candidate plans for the same range of operators and recordings under the three-replicate scenario. Under this scenario, the same candidate plans were identified. Given the added expense of the additional replicate, the team felt no value would be gained by requiring the operators to conduct three repeat tests in order to derive the necessary power or range of confidence intervals.

Table B-8. Power Results and Expected Confidence Intervals under the Two Replicates Scenario for Test TP 98

Process Samples	Statistic	Operators				
		2.00	3.00	4.00	5.00	6.00
2	Power P	0.48	0.78	0.86	0.87	0.89
	Power O	0.12	0.20	0.29	0.36	0.45
	LL Sig P	0.38	0.39	0.38	0.37	0.36
	UL Sig P	27.10	26.67	26.91	26.37	25.68
	LL Reprod	0.12	0.14	0.14	0.15	0.15
	UL Reprod	5.18	0.99	0.60	0.46	0.40
3	Power P	0.86	0.97	0.98	0.99	0.99
	Power O	0.23	0.41	0.56	0.68	0.74
	LL Sig P	0.42	0.43	0.42	0.42	0.43
	UL Sig P	5.08	5.22	5.09	5.12	5.16
	LL Reprod	0.13	0.14	0.15	0.15	0.15
	UL Reprod	4.76	0.94	0.57	0.45	0.38
4	Power P	0.97	1.00	1.00	1.00	1.00
	Power O	0.35	0.58	0.71	0.80	0.87
	LL Sig P	0.47	0.46	0.47	0.47	0.45
	UL Sig P	3.12	3.07	3.10	3.11	3.00
	LL Reprod	0.13	0.14	0.15	0.15	0.16
	UL Reprod	4.59	0.93	0.55	0.43	0.38
5	Power P	0.99	1.00	1.00	1.00	1.00
	Power O	0.44	0.67	0.76	0.89	0.90
	LL Sig P	0.50	0.50	0.49	0.49	0.49
	UL Sig P	2.43	2.41	2.35	2.38	2.37
	LL Reprod	0.14	0.15	0.15	0.16	0.16
	UL Reprod	4.63	0.91	0.54	0.44	0.37
6	Power P	1.00	1.00	1.00	1.00	1.00
	Power O	0.48	0.72	0.84	0.92	0.95
	LL Sig P	0.52	0.51	0.51	0.51	0.51
	UL Sig P	2.05	2.01	2.02	2.00	2.01
	LL Reprod	0.14	0.15	0.15	0.16	0.16
	UL Reprod	4.47	0.89	0.55	0.43	0.37

Table B-9. Power Results and Expected Confidence Intervals under the Two Replicates Scenario for Test TP 98 where the Planning Value for the Process Variance Component Has Been Reduced by One Third

Process Samples	Statistic	Operators				
		2.00	3.00	4.00	5.00	6.00
2	Power P	0.43	0.75	0.84	0.87	0.88
	Power O	0.10	0.16	0.21	0.27	0.32
	LL Sig P	0.26	0.32	0.30	0.30	0.29
	UL Sig P	21.57	22.29	21.45	21.50	21.20
	LL Reprod	0.12	0.13	0.13	0.14	0.14
	UL Reprod	4.44	0.87	0.52	0.41	0.35
3	Power P	0.82	0.95	0.97	0.98	0.99
	Power O	0.16	0.33	0.45	0.53	0.65
	LL Sig P	0.35	0.34	0.35	0.35	0.34
	UL Sig P	4.29	4.16	4.31	4.23	4.15
	LL Reprod	0.12	0.13	0.14	0.14	0.15
	UL Reprod	3.97	0.79	0.49	0.38	0.34
4	Power P	0.96	0.99	0.99	0.99	1.00
	Power O	0.27	0.45	0.59	0.70	0.77
	LL Sig P	0.37	0.38	0.38	0.37	0.37
	UL Sig P	2.49	2.52	2.50	2.46	2.44
	LL Reprod	0.13	0.14	0.14	0.14	0.15
	UL Reprod	3.75	0.77	0.47	0.37	0.33
5	Power P	0.99	1.00	1.00	1.00	1.00
	Power O	0.35	0.54	0.66	0.78	0.85
	LL Sig P	0.40	0.41	0.40	0.40	0.40
	UL Sig P	1.92	1.97	1.96	1.92	1.91
	LL Reprod	0.13	0.14	0.14	0.14	0.15
	UL Reprod	3.70	0.75	0.46	0.36	0.32
6	Power P	1.00	1.00	1.00	1.00	1.00
	Power O	0.40	0.61	0.72	0.83	0.91
	LL Sig P	0.42	0.41	0.42	0.41	0.42
	UL Sig P	1.66	1.64	1.64	1.64	1.64
	LL Reprod	0.13	0.14	0.14	0.15	0.15
	UL Reprod	3.78	0.73	0.44	0.37	0.32

Table B-10. Power Results and Expected Confidence Intervals under the Two Replicates Scenario for Test TP 98 where the Planning Value for the Process Variance Component Has Been Inflated by One Third

Process Samples	Statistic	Operators				
		2.00	3.00	4.00	5.00	6.00
2	Power P	0.53	0.83	0.89	0.91	0.91
	Power O	0.13	0.24	0.35	0.45	0.56
	LL Sig P	0.47	0.44	0.44	0.43	0.43
	UL Sig P	31.54	30.50	30.95	30.83	30.93
	LL Reprod	0.13	0.14	0.15	0.15	0.16
	UL Reprod	5.61	1.12	0.65	0.51	0.44
3	Power P	0.92	0.98	0.99	0.98	0.99
	Power O	0.25	0.48	0.63	0.76	0.83
	LL Sig P	0.51	0.49	0.49	0.49	0.49
	UL Sig P	6.06	5.98	5.95	5.93	5.96
	LL Reprod	0.13	0.15	0.15	0.16	0.16
	UL Reprod	5.00	1.03	0.62	0.49	0.42
4	Power P	0.98	1.00	1.00	1.00	1.00
	Power O	0.41	0.64	0.76	0.85	0.93
	LL Sig P	0.55	0.54	0.53	0.54	0.54
	UL Sig P	3.59	3.57	3.50	3.54	3.57
	LL Reprod	0.14	0.15	0.16	0.16	0.16
	UL Reprod	5.15	1.04	0.62	0.48	0.42
5	Power P	1.00	1.00	1.00	1.00	1.00
	Power O	0.47	0.69	0.84	0.92	0.96
	LL Sig P	0.57	0.57	0.57	0.57	0.57
	UL Sig P	2.75	2.72	2.73	2.74	2.72
	LL Reprod	0.14	0.15	0.16	0.16	0.16
	UL Reprod	5.06	1.00	0.62	0.48	0.42
6	Power P	1.00	1.00	1.00	1.00	1.00
	Power O	0.57	0.79	0.89	0.95	0.97
	LL Sig P	0.59	0.59	0.60	0.59	0.60
	UL Sig P	2.33	2.33	2.37	2.33	2.37
	LL Reprod	0.14	0.15	0.16	0.16	0.17
	UL Reprod	5.28	1.02	0.61	0.47	0.42

Power and Estimation Simulation for Test Method TP 99

The variance component planning values from the TP 99 pilot R&R study were input into the Monte Carlo simulations to produce the desired evaluative measures per candidate plan. Table B-11 shows the power test results and expected confidence intervals under the two replicates scenario for Test TP 99. Applying the same criteria with respect to power tests described above, the research team identified the following 18 experimental design plans:

- 3 samples, with 3 to 6 operators;
- 4 samples, with 3 to 6 operators;
- 5 samples, with 2 to 6 operators; and,
- 6 samples, with 2 to 6 operators.

As with the simulation runs for Test TP 98, the precision of the confidence intervals for both $\hat{\sigma}_P$ and reproducibility improves as the number of operators and samples increases. For example, when the number of replicates is 2, the width of the expected confidence interval for $\hat{\sigma}_P$ for the design with 6 samples and 6 operators is only 5 percent of that which could be expected under a 2 sample, 2 operator design. There is a similar improvement in the precision for the reproducibility estimate because the width of the expected confidence interval under the 6 sample, 6 operator design is only 2 percent of that found with the expected confidence interval under the 2 sample, 2 operator design. The precision of the confidence intervals found in the designs selected using the power test approach are all reasonable based on statistical judgment.

Since the confidence interval ranges for both $\hat{\sigma}_P$ and reproducibility in the plans with 5 to 6 samples and 5 to 6 operators are of similar magnitude (differences are within 50 percent of their values) and provide the best precision given the observed *a priori* variance component estimates, they would be the ideal designs to implement for the full R&R study.

The results of decreasing or increasing the planning values used to run the simulations suggest that when the planning value for the process variance component is increased by a third, another additional design that can be considered are 2 samples with 5 to 6 operators. For each process variance component adjustment, the possible candidate plans are highlighted in Table B-12 and Table B-13. Both simulation runs still suggest that plans with 5 to 6 samples with 5 to 6 operators provide precise and comparable confidence intervals.

Table B-11. Power Test Results and Expected Confidence Intervals under the Two Replicates Scenario for Test TP 99

Process Samples	Statistic	Operators				
		2.00	3.00	4.00	5.00	6.00
2	Power P	0.18	0.43	0.54	0.61	0.67
	Power O	0.24	0.58	0.75	0.90	0.96
	LL Sig P	4.78	7.78	7.81	8.14	8.10
	UL Sig P	670.79	658.46	625.26	631.31	621.58
	LL Reprod	4.97	5.57	6.03	6.22	6.40
	UL Reprod	249.61	47.11	28.29	21.87	18.62
3	Power P	0.41	0.71	0.79	0.84	0.89
	Power O	0.51	0.86	0.95	0.98	0.99
	LL Sig P	6.69	9.10	9.32	9.49	9.78
	UL Sig P	129.31	127.45	124.21	123.55	125.33
	LL Reprod	5.26	5.88	6.30	6.43	6.65
	UL Reprod	157.85	32.99	21.62	17.16	15.47
4	Power P	0.57	0.84	0.91	0.94	0.96
	Power O	0.68	0.91	0.97	0.98	1.00
	LL Sig P	8.03	10.12	10.16	10.50	10.80
	UL Sig P	75.75	75.25	72.84	73.64	74.75
	LL Reprod	5.51	6.07	6.41	6.64	6.83
	UL Reprod	129.01	28.16	18.62	15.46	14.03
5	Power P	0.70	0.93	0.96	0.98	0.98
	Power O	0.75	0.93	0.99	0.99	1.00
	LL Sig P	9.18	10.65	10.90	11.18	11.35
	UL Sig P	58.00	57.04	56.28	56.74	56.86
	LL Reprod	5.71	6.23	6.63	6.84	6.91
	UL Reprod	120.30	25.31	17.28	14.55	13.16
6	Power P	0.84	0.96	0.98	0.99	1.00
	Power O	0.75	0.93	0.99	1.00	1.00
	LL Sig P	10.16	11.54	11.30	11.57	11.82
	UL Sig P	49.90	49.60	47.46	47.77	48.34
	LL Reprod	5.80	6.41	6.68	6.87	7.01
	UL Reprod	109.16	23.68	15.90	13.69	12.46

Table B-12. Power Results and Expected Confidence Intervals under the Two Replicates Scenario for Test TP 99 where the Planning Value for the Process Variance Component Has Been Reduced by One Third

Process Samples	Statistic	Operators				
		2.00	3.00	4.00	5.00	6.00
2	Power P	0.14	0.33	0.47	0.56	0.62
	Power O	0.19	0.46	0.70	0.80	0.90
	LL Sig P	4.14	5.71	6.11	6.37	6.58
	UL Sig P	535.06	541.98	522.04	514.70	522.64
	LL Reprod	4.91	5.69	6.01	6.20	6.31
	UL Reprod	244.94	48.37	28.39	21.75	18.35
3	Power P	0.30	0.58	0.74	0.78	0.84
	Power O	0.50	0.77	0.90	0.96	0.98
	LL Sig P	4.67	6.72	7.29	7.52	7.73
	UL Sig P	102.49	103.30	102.60	101.81	102.07
	LL Reprod	5.33	5.94	6.24	6.47	6.59
	UL Reprod	162.68	32.99	21.40	17.36	15.33
4	Power P	0.44	0.76	0.88	0.92	0.93
	Power O	0.60	0.87	0.94	0.98	1.00
	LL Sig P	5.75	7.75	8.20	8.49	8.61
	UL Sig P	61.75	62.50	61.32	61.50	61.16
	LL Reprod	5.50	6.13	6.42	6.68	6.77
	UL Reprod	132.69	28.04	18.63	15.66	13.92
5	Power P	0.62	0.83	0.94	0.97	0.96
	Power O	0.70	0.90	0.96	0.99	1.00
	LL Sig P	6.72	8.24	8.88	8.84	8.78
	UL Sig P	48.27	47.09	47.53	46.23	45.33
	LL Reprod	5.68	6.31	6.55	6.76	6.93
	UL Reprod	122.35	25.82	16.93	14.44	13.12
6	Power P	0.71	0.90	0.97	0.98	0.99
	Power O	0.73	0.93	0.98	1.00	1.00
	LL Sig P	7.40	8.73	9.15	9.45	9.50
	UL Sig P	41.27	39.99	39.73	40.01	39.67
	LL Reprod	5.82	6.42	6.73	6.87	7.00
	UL Reprod	109.92	23.74	16.49	13.60	12.49

Table B-13. Power Results and Expected Confidence Intervals under the Two Replicates Scenario for Test TP 99 where the Planning Value for the Process Variance Component Has Been Increased by One Third

Process Samples	Statistic	Operators				
		2.00	3.00	4.00	5.00	6.00
2	Power P	0.18	0.44	0.61	0.70	0.70
	Power O	0.26	0.67	0.84	0.91	0.96
	LL Sig P	5.63	9.48	9.73	9.83	9.46
	UL Sig P	740.14	755.47	747.35	742.86	713.85
	LL Reprod	5.12	5.61	5.98	6.18	6.45
	UL Reprod	257.10	47.57	28.10	21.58	18.74
3	Power P	0.46	0.76	0.84	0.88	0.92
	Power O	0.61	0.87	0.96	0.99	1.00
	LL Sig P	8.39	10.78	11.34	11.35	11.24
	UL Sig P	150.57	144.79	146.22	144.65	142.27
	LL Reprod	5.31	5.86	6.24	6.43	6.69
	UL Reprod	153.46	33.13	21.43	17.18	15.46
4	Power P	0.65	0.88	0.95	0.96	0.98
	Power O	0.71	0.92	0.97	1.00	1.00
	LL Sig P	10.38	11.90	12.13	12.38	12.20
	UL Sig P	87.51	85.75	84.90	85.35	83.44
	LL Reprod	5.51	6.18	6.48	6.67	6.73
	UL Reprod	131.62	29.24	18.85	15.63	13.85
5	Power P	0.80	0.96	0.98	0.99	0.99
	Power O	0.78	0.95	0.99	1.00	1.00
	LL Sig P	11.33	12.73	12.88	12.98	13.14
	UL Sig P	66.34	65.92	65.14	64.87	65.11
	LL Reprod	5.70	6.30	6.59	6.77	6.90
	UL Reprod	122.32	25.86	17.06	14.43	13.07
6	Power P	0.88	0.98	0.99	1.00	1.00
	Power O	0.81	0.96	0.99	1.00	1.00
	LL Sig P	12.35	13.33	13.59	13.68	13.89
	UL Sig P	56.76	55.90	55.87	55.75	56.21
	LL Reprod	5.84	6.33	6.66	6.92	7.03
	UL Reprod	111.01	23.46	15.99	13.75	12.54

Signal-to-Noise Ratio Simulation

The research team also utilized the Monte Carlo simulations to find the confidence intervals of the statistical SNR. The statistical SNR provides a way of looking at the *dependability* of a method. See the main body of this report as well as Appendix C for additional discussion of the SNR. Where the SNR is less than 3 and definitely less than 2 indicate that the test method is not adequate. SNR values that are 5 or greater are considered very good. For each candidate design plan, the research team evaluated the SNR and compared if the power, estimation and SNR approaches suggest the same candidate experimental design plans.

Signal-to-Noise Ratio Simulation for Test Method TP 98

Table B-14 shows the planning values of the variance components from the pilot R&R study produced a SNR of 5.9459. While the indicator is above 5, one needs to evaluate if lower limit of its expected 95 percent confidence interval is also not below 3. If it is below 3, there is a risk that future R&R gauge studies with the same candidate plans may not produce acceptable SNR values. This may lead researchers to question the actual adequacy of the test methodology. Table B-14 shows the results of using this estimate under differing candidate design plans; the highlighted areas identify plans where the lower limit of the 95 percent confidence interval around the SNR estimate is above 3.

Table B-14 shows that by applying the SNR criterion, the analysis suggests 5 samples with 6 operators and 6 samples with 5 operators and 6 samples with 6 operators can produce SNR values within R&R Gauge Study norms. The candidate plans overlap with those from the power test simulation approach. However, the number of candidate plans is significantly lower (compare these 3 to the 11 plans based on the power test approach).

Table B-15 shows additional SNR simulation runs using planning values for the process variance component is one third less and Table B-16 shows the process variance component is one third more. In the run where the planning values were increased, several more candidate plans qualified under the SNR criteria:

- 4 samples, with 5 to 6 operators;
- 5 samples, with 4 to 5 operators; and,
- 6 samples, 4 operators.

The research team investigated candidate plans for the same set of SNRs under the three-replicate scenario. Under this scenario, the same candidate plans were identified. Given the added expense of the additional replicate, the research team felt no value would be gained by requiring the operators to conduct three tests in order to derive the necessary power or range of confidence intervals.

Table B-14. Expected 95 Percent Confidence Intervals for the SNR Estimate of 5.9459 for Test TP 98

Process Samples	Statistics	Operators				
		2	3	4	5	6
2	LL	0.6	1.5	1.9	2.1	2.1
	UL	86.6	78.1	80.8	79.1	78.0
3	LL	1.1	2.0	2.4	2.5	2.7
	UL	28.6	26.9	26.1	24.9	25.2
4	LL	1.2	2.1	2.5	2.7	2.9
	UL	19.8	18.4	17.8	17.3	17.7
5	LL	1.2	2.3	2.6	2.9	3.0
	UL	17.6	16.1	15.1	14.9	14.8
6	LL	1.3	2.3	2.8	3.0	3.2
	UL	15.6	14.6	14.1	13.7	13.4

Table B-15. Expected 95 Percent Confidence Intervals for the SNR Estimate of 4.8653 for Test TP 98 Where the Planning Value for the Process Variance Component Estimate Has Been Reduced by One Third

Parts	Statistics	Operators				
		2	3	4	5	6
2	LL	0.5	1.3	1.5	1.6	1.8
	UL	71.5	70.0	64.6	62.3	65.8
3	LL	0.8	1.7	1.9	2.1	2.2
	UL	23.0	22.2	21.1	20.7	20.5
4	LL	1.0	1.7	2.1	2.3	2.4
	UL	16.8	15.4	14.7	14.6	14.7
5	LL	1.0	1.8	2.2	2.4	2.4
	UL	14.0	12.8	12.4	12.5	11.9
6	LL	1.1	1.9	2.3	2.4	2.6
	UL	12.9	11.9	11.6	10.9	11.1

Table B-16. Expected 95 Percent Confidence Intervals for the SNR Estimate of 6.8805 for Test TP 98 Where the Planning Value for the Process Variance Component Estimate Has Been Increased by One Third

Parts	Statistics	Operators				
		2	3	4	5	6
2	LL	0.8	1.8	2.2	2.4	2.5
	UL	101.3	89.7	92.7	91.9	91.7
3	LL	1.3	2.3	2.8	2.9	3.1
	UL	32.5	29.6	30.2	28.6	29.0
4	LL	1.4	2.4	2.9	3.2	3.3
	UL	23.0	21.4	20.6	20.4	19.8
5	LL	1.4	2.5	3.1	3.4	3.5
	UL	19.5	18.1	17.6	17.5	16.8
6	LL	1.4	2.6	3.3	3.5	3.7
	UL	18.1	16.7	16.4	15.8	15.6

Signal-to-Noise Ratio Simulation for Test Method TP 99

The planning values of the variance components from the pilot R&R study produce a SNR of 3.2108 (see Table B-17). While the estimate is above 3, one needs to evaluate if lower limit of its 95 percent confidence interval is also not below 3. Using this estimate under differing candidate plans, no plan has a lower limit greater than 3. In fact, all the lower limits are less than 2. By applying the SNR criterion, the data and analysis suggests that there is no ideal plan within the range of 2 to 6 samples with 2 to 6 operators that can produce good SNR ranges.

Table B-18 shows additional SNR simulation runs using planning values for the process variance component is one third less and Table B-19shows the process variance component is one third more. In both of these runs, none of the candidate plans produce satisfactory confidence intervals for the estimated SNR.

Table B-17. Expected 95 Percent Confidence Intervals for the SNR Estimate of 3.2108 for Test TP 99

Process Samples	Statistics	Operators				
		2	3	4	5	6
2	LL	0.2	0.6	0.8	0.9	1.0
	UL	57.9	49.2	45.0	44.3	43.0
3	LL	0.3	0.9	1.2	1.3	1.4
	UL	17.0	14.6	14.4	13.7	13.6
4	LL	0.5	1.1	1.4	1.6	1.6
	UL	11.8	10.2	9.7	9.9	9.3
5	LL	0.6	1.3	1.5	1.6	1.7
	UL	9.6	8.6	8.2	8.0	8.0
6	LL	0.7	1.4	1.7	1.8	1.8
	UL	8.5	7.8	7.5	7.4	7.2

Table B-18. Expected 95 Percent Confidence Intervals for the SNR Estimate of 2.6216 for Test TP 99 Where the Planning Value for the Process Variance Component Estimate Has Been Reduced by One Third

Parts	Statistics	Operators				
		2	3	4	5	6
2	LL	0.2	0.4	0.6	0.7	0.8
	UL	52.1	39.8	37.1	37.3	35.3
3	LL	0.2	0.6	0.9	1.0	1.1
	UL	14.6	12.2	11.6	11.6	11.0
4	LL	0.3	0.8	1.0	1.2	1.2
	UL	9.2	8.4	8.0	7.9	7.8
5	LL	0.4	0.9	1.2	1.3	1.4
	UL	7.8	6.9	6.7	6.6	6.6
6	LL	0.5	1.1	1.3	1.4	1.5
	UL	7.2	6.4	6.2	6.1	6.0

Table B-19. Expected 95 Percent Confidence Intervals for the SNR Estimate of 3.7075 for Test TP 99 Where the Planning Value for the Process Variance Component Estimate Has Been Increased by One Third

Parts	Statistics	Operators				
		2	3	4	5	6
2	LL	0.3	0.7	1.0	1.2	1.2
	UL	64.5	55.4	52.3	52.8	50.8
3	LL	0.5	1.2	1.4	1.5	1.6
	UL	20.2	17.8	16.3	15.8	15.3
4	LL	0.6	1.4	1.7	1.8	1.8
	UL	13.1	11.9	11.4	10.9	10.7
5	LL	0.8	1.5	1.8	1.9	2.0
	UL	11.1	9.8	9.5	9.2	9.1
6	LL	0.9	1.7	2.0	2.1	2.1
	UL	9.7	9.1	8.7	8.5	8.3

Recommendations

Recommendations for Test Method TP 98

Based on the results of the power analysis and the SNR analysis, the research team recommends designing the full R&R study for Test TP 98 using 5 to 6 samples with 5 to 6 operators. Other designs that can be considered are the following:

- 3 samples, 6 operators;
- 4 samples, with 4 to 6 operators;
- 5 samples, 4 operators; and,
- 6 samples, with 3 to 4 operators.

However, trade-offs may need to be made in terms of lower precision and possibly observing values of SNRs between 2 and 3. SNR values between 2 and 3 are not ideal, but they are marginally acceptable. The research team will use the evidence from the full R&R study coupled with its statistical judgment to ascertain the adequacy of Test TP 98.

Recommendations for Test Method TP 99

Based on the results of the power test analysis and the SNR analysis, the research team recommends designing the full R&R study for Test TP 99 using 6 recordings and 6 operators as a minimum.

While other designs can be considered since they met the power and confidence interval criteria, the fact is that none of the candidate designs even with varying the planning values produce satisfactory SNR estimates and confidence interval widths. By selecting the design with the largest number of recordings and operators we reduce the chance of producing SNRs that are outside accepted norms for a given method when their true values are actually within the recommended guidelines.

Experimental Design for Full R&R Study

Based upon the results of the R&R study and of the statistical simulations, the following experimental design for both the TP 98 and TP 99 test methods should meet satisfactory levels of precision within the constraints of the present research project:

- 5 operators;
- 4 process samples (calibrated audio and synchronized video)
- 2 replicates (samples and replicates will be presented to each operator in random order).

As noted for the TP 99 test method, the simulated SNR indicates that anything less than 6 operators and 6 process samples risks findings of greater variance estimates than the actual variance of the test method. However the constraints of this project (time and budget) do not allow for this level of effort. In lieu of the ideal experimental design, this proposed combination offers the highest power and signal-to-noise ratio for either the TP 98 or TP 99 test methods based upon the results of the R&R study and statistical simulations.

APPENDIX C

Complete Statistical Output for Repeatability and Reproducibility Study

Introduction to Repeatability and Reproducibility Study

A Repeatability and Reproducibility (R&R) study is a special type of analysis of variance (ANOVA) method. R&R studies are designed to partition the total process variation into two main components: (1) variation from the process itself (for example, why are two samples different?) and (2) variation from the measurement procedure (for example, why would two different operators produce different measurements using the same sample?). The latter is further subdivided into (1) variation from the operator or laboratory (reproducibility) and (2) variation due to the measuring device (repeatability). In other words, reproducibility is the variation among operators or labs while repeatability is the variation within operators or laboratories. (For these measurements there is no distinction between operators and laboratories. Each operator will execute a unique realization of the test method. The difference between one operator's realizations with another's is irrespective of whether they are based in the same laboratory.)

The basic design model used for most R&R studies is straightforward:

$$Y_{ijk} = \mu + P_i + O_j + (PO)_{ij} + \varepsilon_{ijk} \tag{C-1}$$

Where:

Y_{ijk} = measured value of the ith sample taken by the jth operator on the kth replicate
μ = grand mean
P_i = the effect of the ith sample (process)
O_j = the effect of the jth operator or laboratory
$(PO)_{ij}$ = the interaction effect between the ith sample (process) and the jth operator or lab
ε_{ijk} = random error

The terms P_i, O_j, $(PO)_{ij}$, and ε_{ijk} are jointly independent normal random variables with means of zero and variances of σ_P^2 , σ_O^2 , σ_{PO}^2 , and σ_ε^2, respectively. These variances are related to the total variance by the following equation:

$$\sigma_T^2 = \sigma_P^2 + \sigma_O^2 + \sigma_{PO}^2 + \sigma_\varepsilon^2 \tag{C-2}$$

Where:

σ_T^2 = total variance of a measurement process
σ_P^2 = variance of actual process (physical differences in samples)
σ_O^2 = variance of different operators
σ_{PO}^2 = variance of interaction between operator and sample
σ_ε^2 = variance of single operator

Repeatability is captured by σ_ε^2 , reproducibility is $\gamma_1 = \sigma_O^2 + \sigma_{PO}^2$, and total variability associated with the measurement process is $\gamma_2 = \sigma_O^2 + \sigma_{PO}^2 + \sigma_\varepsilon^2$. The latter expression is referred to as the "R&R" value. The process (sample-to-sample) variability which arises when the samples have different attributes such as different physical properties or measured under different treatments or conditions is designated as σ_P^2 (The subscript P identifies that the variance is from the process and is attributed to the sample). The sample for the purpose of this study is a *particular road segment, on a particular day with a particular vehicle fleet* for both test methods. This sample source of variability needs to be separately tracked from the sources of measurement variability in order to compare if a method is reliable or not.

For practical purposes, researchers compare the standard deviation of each of the components to the total sample's standard deviation. The standard deviation of a sample of measurements is the square root of the sample's variance. The advantage of using the standard deviation instead of the variance is that the units of the standard deviation are identical to the units of the actual measurements. Precision in terms of deviations in units of decibels versus deviations in units of square-decibels is more easily understood.

If this index is less than 30%, the measurement process is deemed to be acceptable. The lower the value is, the better the measurement process. If the index is higher than 30 percent, then the process is deemed to be unacceptable and not precise.

Signal-to-Noise Ratio

To compare measurement variability to the total process variability, the following ratio is used:

$$\delta = \frac{2\sigma_P^2}{\sigma_O^2 + \sigma_{PO}^2 + \sigma_\varepsilon^2} \qquad \text{(C-2)}$$

The square root of δ is commonly referred to as the Signal-to-Noise Ratio (SNR), a statistical concept used to examine the relative magnitudes of appropriate variance components. The statistical signal-to-noise ratio is a different measurement than the acoustic signal-to-noise ratio. The statistical SNR provides a way of looking at the *dependability* of a method. The AIAG Measurement Systems Analysis Manual (3, p. 117) recommends an SNR value of five or greater. A value less than 2 indicates that the measurement system is of no value in monitoring the process. A value between 2 and including 3 is also questionable (Minitab 2010).

The manufacturing sector tracks the SNR indicator and its confidence interval as it has found that designs where the SNR is less then 3 and definitely less than 2 indicate that the test method is not adequate. SNR values that are 5 or greater are considered very good.

Computation

The research team used the software NCSS (Hintze, J. (2009). NCSS. NCSS, LLC. Kaysville, Utah WWW.NCSS.com) to run a full R&R report for test method TP 98 (automobiles and heavy trucks) and for test method TP 99. Other software packages can be used to produce the same variance components and summary statistics; however, due to minor differences in the statistical formulae, estimates may differ by 3 to 4 places after the decimal.

Datasets for Repeatability and Reproducibility Studies

Table C-1. Test Method TP 98 Dataset for R&R Study.

Sample Class	Sample	Operator	SIPI (Autos)	SIPI (Heavy Trucks)
M	1	1	5.40	2.60
H	2	1	4.40	-0.70
G	5	1	4.92	0.23
E	6	1	1.84	-2.69
G	9	1	4.65	0.00
H	12	1	3.47	-0.27
M	13	1	4.06	2.20
E	16	1	4.05	-1.61
M	1	2	4.18	1.79
G	2	2	2.92	-0.37
H	5	2	2.87	0.04
M	7	2	3.86	1.74
E	9	2	4.60	0.67
G	10	2	2.99	-0.32
E	11	2	5.09	0.66
H	13	2	3.22	0.30
G	4	3	1.98	-0.73
M	5	3	4.15	1.35
E	6	3	4.62	0.87
H	7	3	3.51	0.11
H	10	3	4.08	1.50
E	11	3	2.90	-0.13
G	13	3	4.03	0.13
M	15	3	4.59	-0.69
G	2	4	3.57	-0.07
E	3	4	4.30	0.71
M	4	4	3.75	2.29
M	7	4	3.79	2.15
E	8	4	4.46	0.49
H	11	4	3.21	0.50
G	13	4	1.39	-0.72
H	16	4	3.31	0.33
G	3	5	2.04	-0.17
M	7	5	3.78	1.41
E	8	5	3.97	0.12
H	10	5	2.68	0.35
H	12	5	2.68	0.75
E	13	5	4.06	0.44
G	14	5	2.43	-0.48
M	16	5	3.94	2.28

Table C-2. Test Method TP 99 Dataset for R&R Study

Sample Class	Sample	Operator	CTIM Measurement
E	3	1	70.80
M	4	1	70.18
H	7	1	74.00
E	8	1	69.90
H	10	1	74.30
K	11	1	77.40
M	14	1	70.53
K	15	1	77.20
H	3	2	74.15
M	4	2	70.22
K	6	2	77.29
K	8	2	77.29
E	12	2	70.39
E	14	2	70.51
M	15	2	70.22
H	16	2	74.49
M	1	3	70.11
E	2	3	70.02
M	3	3	70.20
H	8	3	74.01
E	9	3	69.88
K	12	3	77.16
K	14	3	77.16
H	16	3	73.95
M	1	4	70.19
K	5	4	77.10
M	6	4	70.21
H	9	4	74.43
K	10	4	76.91
E	12	4	70.18
H	14	4	74.09
E	15	4	70.39
E	1	5	69.93
K	2	5	77.14
E	4	5	70.01
M	5	5	70.24
H	6	5	74.15
K	9	5	76.95
M	11	5	70.28
H	15	5	74.08

Complete Statistical Results for Repeatability and Reproducibility Study

The results of the R&R analysis below breakdown the variance components by sample site, operator and the interaction between operator and sample site. The values for reproducibility and repeatability and their totals are also provided. The percent total standard deviation attributed to the R&R component is also shown in order to assess the ability of the overall measurement process to be precise. And the performance indicator signal-to-noise ratio is given.

Results for TP 98 (Automobiles) Repeatability and Reproducibility

With respect to the results of the R&R study for the test method TP 98 automobiles, the results indicate that the process is not precise since the percent R&R of the total standard deviation is 91 percent, which is much larger than the accepted maximum percent R&R value of 30 percent.

The SNR based on the study's results is 0.6513. For a process to be precise enough to distinguish samples, the SNR value should not fall below 3. Since even the upper limit of the 90 percent confidence interval for the SNR is less than 3, this process is not adequate and hence not precise.

Table C-3. Test Method TP 98 (Automobiles) R&R Report: Data Summary Section

Item	Actual Count	Expected Count
Total Values	40	40
Site	4	
Operator	5	
Replicates	2	

Table C-4. Test Method TP 98 (Automobiles) R&R Report: Expected Mean Square and Variance Component Section

Source Term	DF	Expected Mean Square	Variance Component	Lower 90% Conf. Limit	Upper 90% Conf. Limit
Site (P)	3	R+2(PO)+10(P)	0.1577501		
Operator (O)	4	R+2(PO)+8(O)	0		
Interaction (PO)	12	R+2(PO)	0.237777		
Replicates (R)	20	R	0.5174177	0.3294559	0.953694

Table C-5. Test Method TP 98 (Automobiles) R&R Report: Analysis of Variance Section

Source Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Site	3	7.711416	2.570472	2.59	0.101402
Operator	4	3.607794	0.9019486	0.91	0.489762
Interaction	12	11.91566	0.9929717	1.92*	0.095258
Replicates	20	10.34835	0.5174177		
Total (Adjusted)	39	33.58323			
Total	40				

Table C-6. Test Method TP 98 (Automobiles) R&R Report: Variance Section

Term	Variance	% Total Variance	Standard Deviation	Lower 90%Conf. Limit of Std Dev	Upper 90%Conf. Limit of Std Dev	% Total Std Dev
Site	0.157750	17.4973	0.3972	0.0000	1.4439	41.8298
Operator	0.000000	0.0000	0.0000	0.0000	0.7104	0.0000
Interaction	0.237777	26.3738	0.4876	0.0000	0.9394	51.3554
Reproducibility	0.226399	25.1117	0.4758	0.0000	0.9641	50.1116
Repeatability	0.517418	57.3909	0.7193	0.5740	0.9766	75.7568
R and R	0.743817	82.5027	0.8624	0.7397	1.2196	90.8310
Total Variation	0.901567	100.0000	0.9495	0.8207	1.7170	100.0000

Note that the variance component for the operator effect is set to zero since its actual value is negative. Theoretically, the variance of sample of observations is positive. However, in conditions of small sample sizes where observations can vary greatly from one another, negative values are possible.

Table C-7. Test Method TP 98 (Automobiles) R&R Report: Percent of Process Variation R & R Section

Term	Lower 90% Conf. Limit	5.15 Std Dev	Upper 90% Conf. Limit	% Total Std Dev	Percent Contribution
Site	0.0000	2.0455	7.4360	41.8298	17.4973
Operator	0.0000	0.0000	3.6585	0.0000	0.0000
Interaction	0.0000	2.5113	4.8377	51.3554	26.3738
Reproducibility	0.0000	2.4504	4.9651	50.1116	25.1117
Repeatability	2.9560	3.7045	5.0293	75.7568	57.3909
R and R	3.8093	4.4416	6.2812	90.8310	82.5027
Total Variation	4.2268	4.8900	8.8426	100.0000	100.0000

Since the % R&R value is greater than 30%, the measurement system is not acceptable. Identify the measurement problems and correct them.

Table C-8. Test Method TP 98 (Automobiles) R&R Report: R & R Indices Section

Index	Lower 90%Conf. Limit	Value	Upper 90%Conf. Limit
Signal-to-Noise Ratio		0.6513	2.3185

Since the lower confidence limit of Signal-to-Noise Ratio is less than 3, the measurement process may be inadequate.

Table C-9. Test Method TP 98 (Automobiles) R&R Report: Means and Bias Section

Term	Count	Mean	Deviation From Target
Overall	40	3.644	
Site			
E	10	3.990	
G	10	3.091	
H	10	3.343	
M	10	4.150	
Operator			
Operator 1	8	4.101	
Operator 2	8	3.716	
Operator 3	8	3.732	
Operator 4	8	3.472	
Operator 5	8	3.197	
Site,Operator			
E,Operator 1	2	2.948	
E,Operator 2	2	4.846	
E,Operator 3	2	3.758	
E,Operator 4	2	4.382	
E,Operator 5	2	4.014	
G,Operator 1	2	4.786	
G,Operator 2	2	2.953	
G,Operator 3	2	3.006	
G,Operator 4	2	2.480	
G,Operator 5	2	2.232	
H,Operator 1	2	3.937	
H,Operator 2	2	3.044	
H,Operator 3	2	3.795	
H,Operator 4	2	3.260	
H,Operator 5	2	2.681	
M,Operator 1	2	4.732	
M,Operator 2	2	4.020	
M,Operator 3	2	4.370	
M,Operator 4	2	3.767	
M,Operator 5	2	3.861	

Since there is no target value for TP 98, the deviation from target column is blank.

Data Plots for TP 98 (Automobiles) Repeatability and Reproducibility

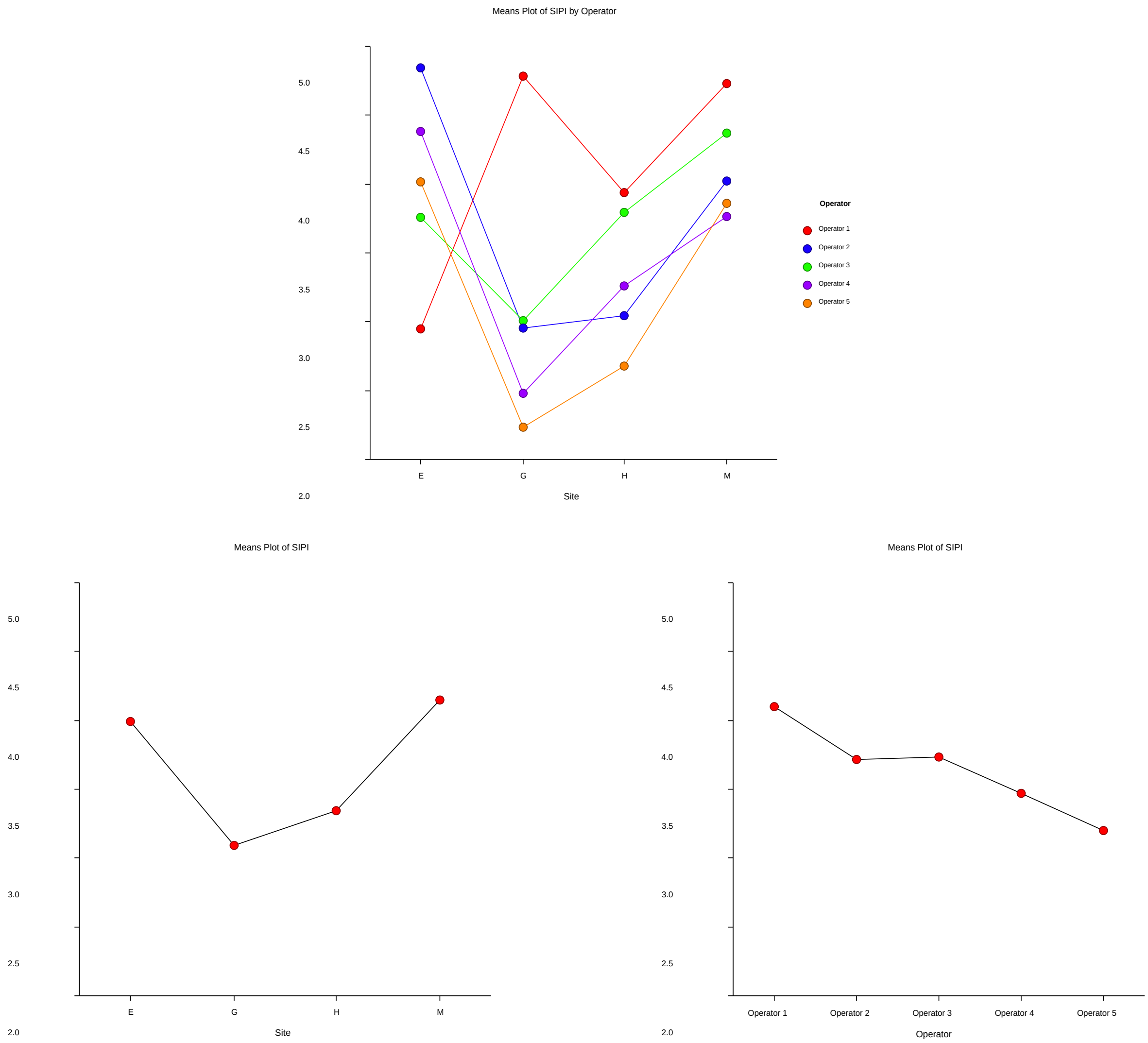


Figure C-1. Test Method TP 98 (Automobile) R&R Study Means Plots

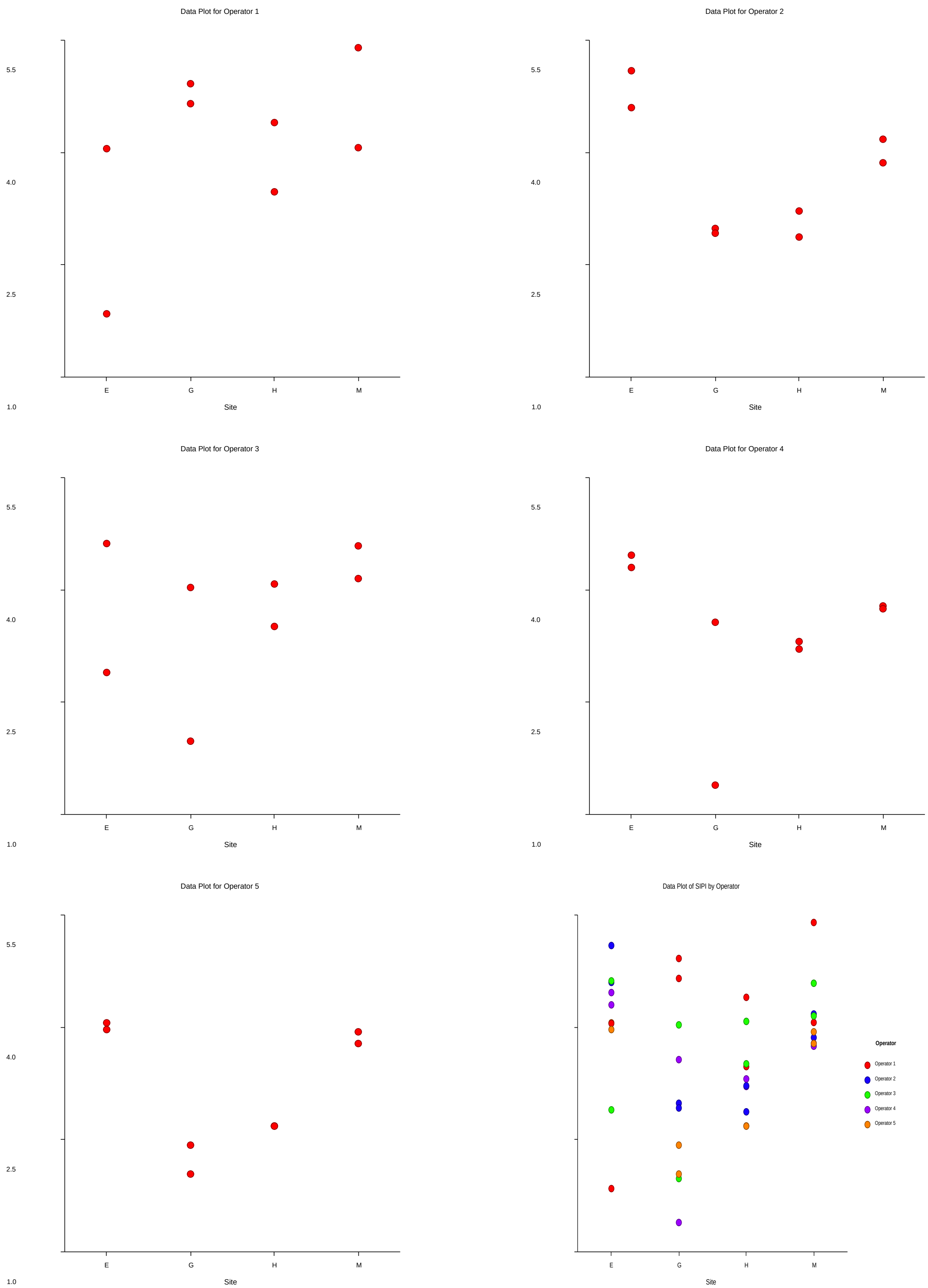


Figure C-2. Test Method TP 98 (Automobile) R&R Study Residual Plots by Operator

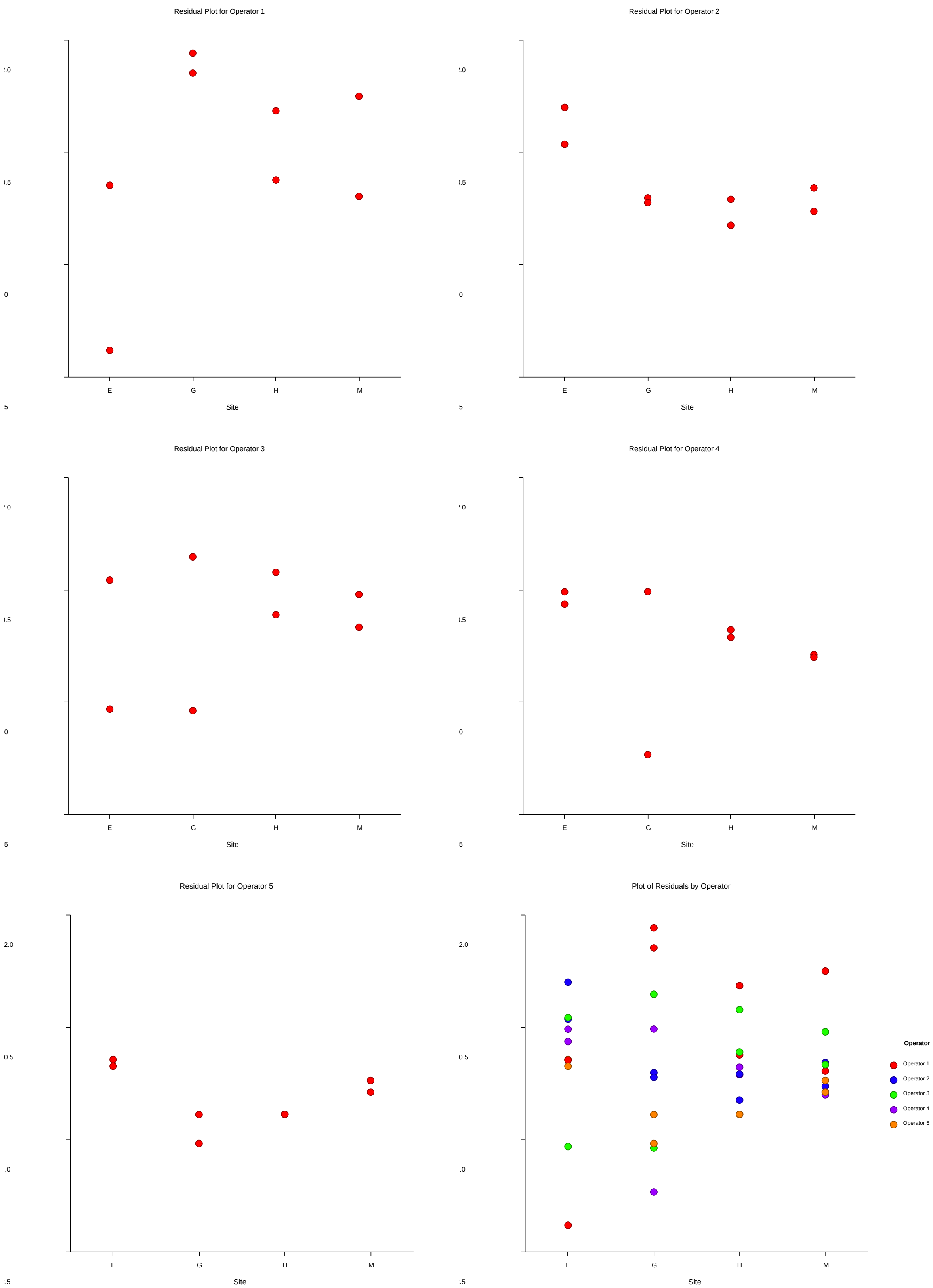


Figure C-3. Test Method TP 98 (Automobile) R&R Study Residual Plots by Operator

Results for TP 98 (Heavy Trucks) Repeatability and Reproducibility

With respect to the results of the R&R study for test method TP 98 heavy truck, the results indicate that the process is not precise since the percent R&R value is 73 percent, which is much larger than the accepted maximum percent R&R value of 30 percent. Of note is that TP 98 Heavy Trucks' percent R&R value is better (meaning closer to the accepted maximum of 30 percent) than that observed for TP 98 Automobiles (91 percent). Table 5-4 summarizes these results for heavy trucks.

The SNR value based on the study's results is 1.3383. For a process to be precise, the lower limit of the SNR's confidence interval should not fall below 3. Since this is the case, this process cannot distinguish across the samples and hence is not precise.

Table C-10. Test Method TP 98 (Heavy Trucks) R&R Report: Data Summary Section

Item	Actual Count	Expected Count
Total Values	40	40
Site	4	
Operator	5	
Replicates	2	

Table C-11. Test Method TP 98 (Heavy Trucks) R&R Report: Expected Mean Square and Variance Component Section

Source Term	DF	Expected Mean Square	Variance Component	Lower 90% Conf. Limit	Upper 90% Conf. Limit
Site (P)	3	R+2(PO)+10(P)	0.6512048	0.1405309	6.548032
Operator (O)	4	R+2(PO)+8(O)	0		
Interaction (PO)	12	R+2(PO)	0.529771	0.2222135	1.394641
Replicates (R)	20	R	0.2771196	0.1764507	0.5107813

Table C-12. Test Method TP 98 (Heavy Trucks) R&R Report: Analysis of Variance Section

Source Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Site	3	23.54613	7.848709	5.87*	0.010478
Operator	4	2.796268	0.6990671	0.52	0.720909
Interaction	12	16.03994	1.336662	4.82*	0.000999
Replicates	20	5.542392	0.2771196		
Total (Adjusted)	39	47.92472			
Total	40				

Table C-13. Test Method TP 98 (Heavy Trucks) R&R Report: Variance Section

Term	Variance	% Total Variance	Standard Deviation	Lower 90%Conf. Limit of Std Dev	Upper 90%Conf. Limit of Std Dev	% Total Std Dev
Site	0.651205	47.2437	0.8070	0.3749	2.5589	68.7340
Operator	0.000000	0.0000	0.0000	0.0000	0.5654	0.0000
Interaction	0.529771	38.4339	0.7279	0.4714	1.1809	61.9951
Reproducibility	0.450072	32.6518	0.6709	0.4344	1.1003	57.1418
Repeatability	0.277120	20.1045	0.5264	0.4201	0.7147	44.8380
R and R	0.727191	52.7563	0.8528	0.7073	1.2254	72.6336
Total Variation	1.378396	100.0000	1.1741	0.9298	2.7056	100.0000

Table C-14. Test Method TP 98 (Heavy Trucks) R&R Report: Percent of Process Variation R & R Section

Term	Lower 90% Conf. Limit	5.15 Std Dev	Upper 90% Conf. Limit	% Total Std Dev	Percent Contribution
Site	1.9306	4.1559	13.1784	68.7340	47.2437
Operator	0.0000	0.0000	2.9117	0.0000	0.0000
Interaction	2.4277	3.7484	6.0819	61.9951	38.4339
Reproducibility	2.2374	3.4550	5.6666	57.1418	32.6518
Repeatability	2.1633	2.7111	3.6807	44.8380	20.1045
R and R	3.6428	4.3917	6.3108	72.6336	52.7563
Total Variation	4.7887	6.0464	13.9336	100.0000	100.0000

Since the % R&R value is greater than 30%, the measurement system is not acceptable. Identify the measurement problems and correct them.

Table C-15. Test Method TP 98 (Heavy Trucks) R&R Report: R & R Indices Section

Index	Lower 90%Conf. Limit	Value	Upper 90%Conf. Limit
Signal-to-Noise Ratio	0.5440	1.3383	4.1824

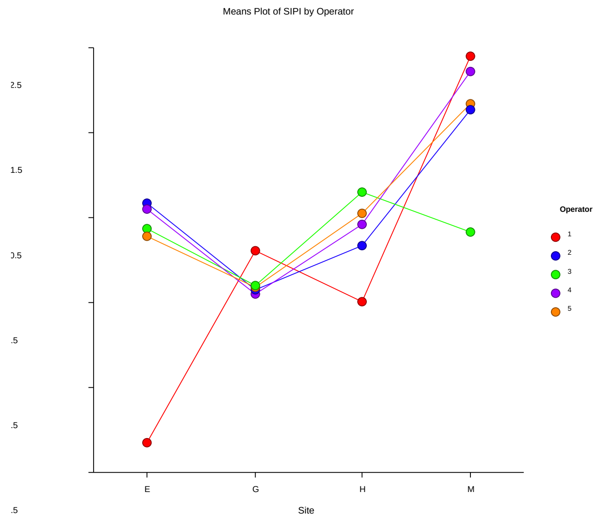
Since the lower confidence limit of Signal-to-Noise Ratio is less than 3, the measurement process may be inadequate.

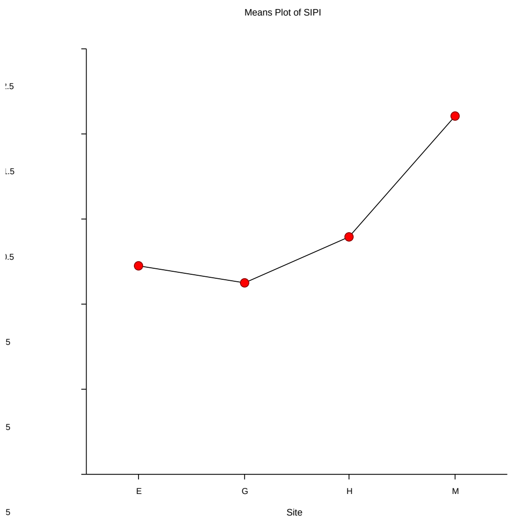
Table C-16. Test Method TP 98 (Heavy Trucks) R&R Report: Means and Bias Section

Term	Count	Mean	Deviation From Target
Overall	40	0.427	
Site			
E	10	-0.047	
G	10	-0.250	
H	10	0.291	
M	10	1.713	
Operator			
1	8	-0.030	
2	8	0.565	
3	8	0.302	
4	8	0.710	
5	8	0.588	
Site,Operator			
E,1	2	-2.148	
E,2	2	0.670	
E,3	2	0.367	
E,4	2	0.598	
E,5	2	0.278	
G,1	2	0.113	
G,2	2	-0.346	
G,3	2	-0.297	
G,4	2	-0.396	
G,5	2	-0.323	
H,1	2	-0.487	
H,2	2	0.168	
H,3	2	0.804	
H,4	2	0.417	
H,5	2	0.553	
M,1	2	2.402	
M,2	2	1.767	
M,3	2	0.332	
M,4	2	2.219	
M,5	2	1.844	

Since there is no target value for TP 98, the deviation from target column is blank.

Data Plots for TP 98 (Heavy Trucks) Repeatability and Reproducibility





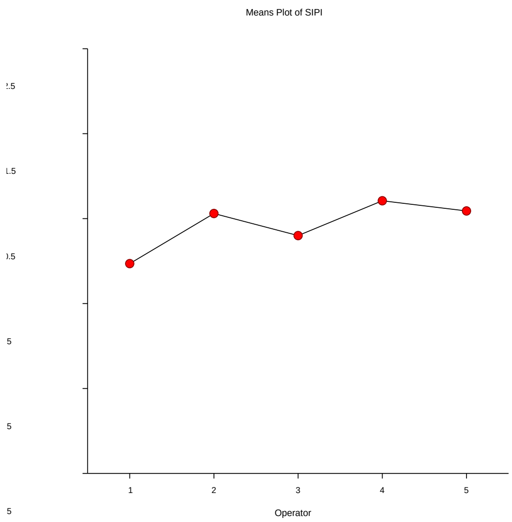


Figure C-4. Test Method TP 98 (Heavy Trucks) R&R Study Means Plots

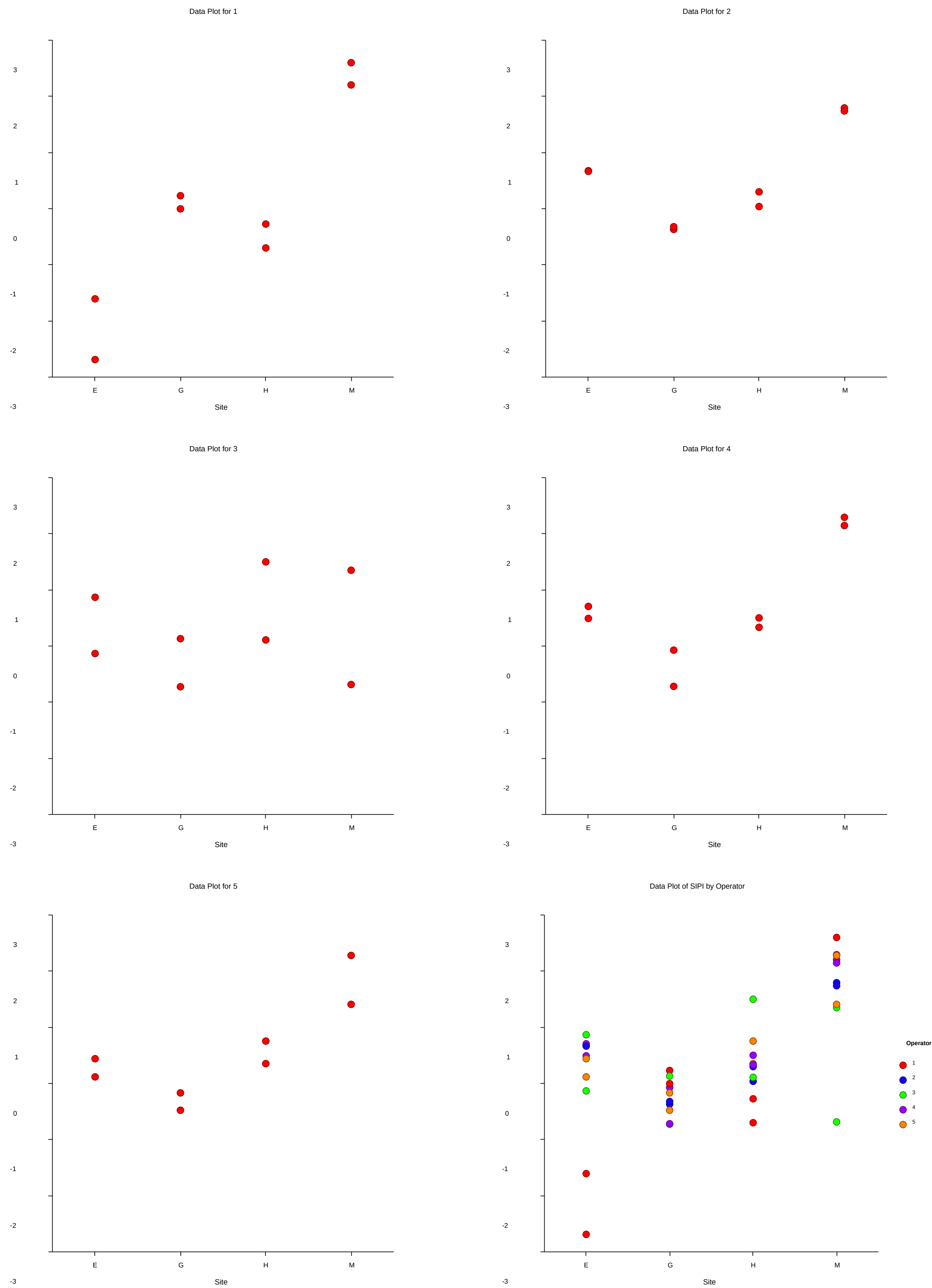


Figure C-5. Test Method TP 98 (Heavy Trucks) R&R Study Data Plots by Operator

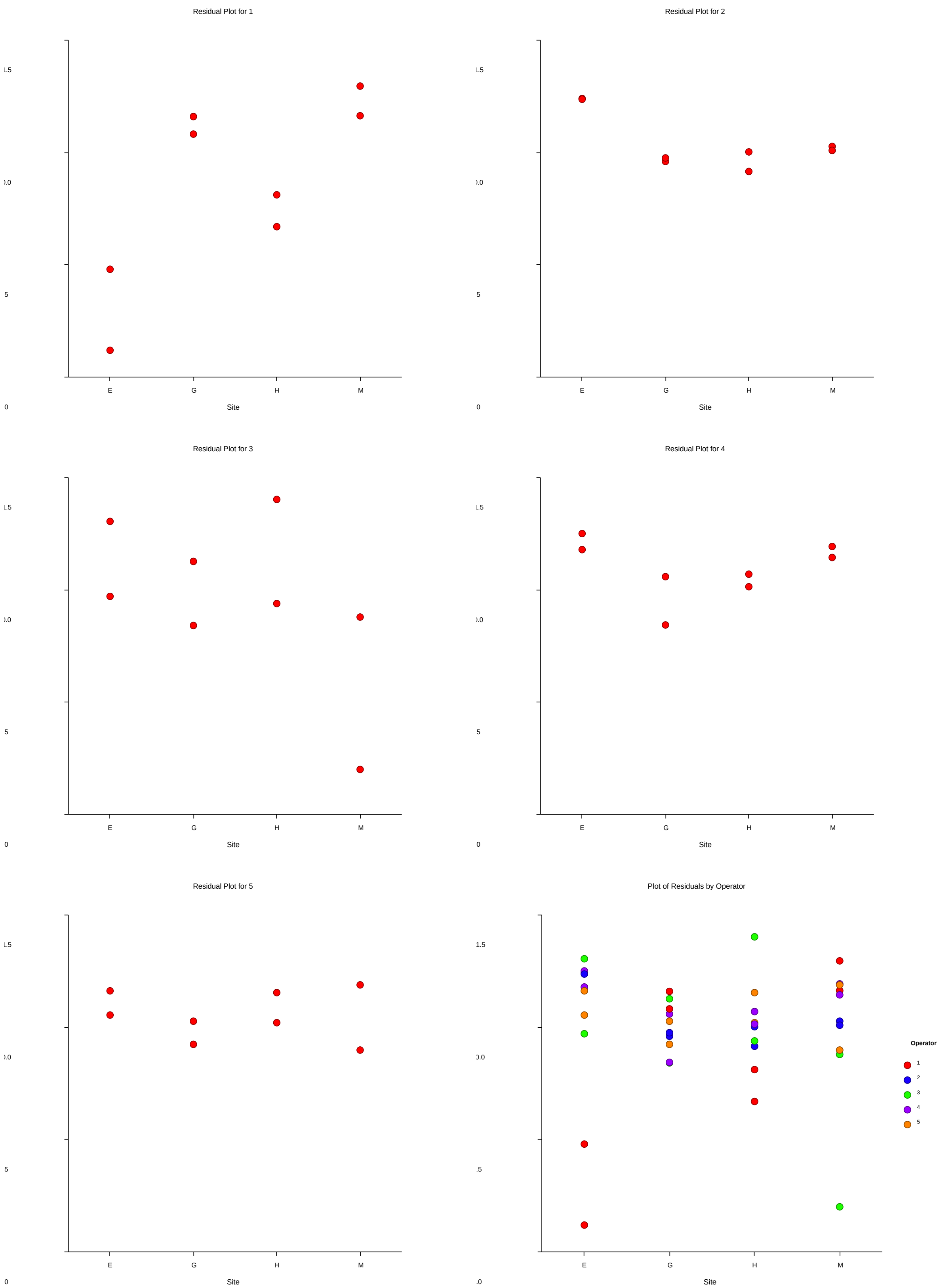


Figure C-6. Test Method TP 98 (Heavy Trucks) R&R Study Residual Plots by Operator

Results for TP 99 Repeatability and Reproducibility

The R&R component's percent standard deviation is a remarkable 6 percent. The percent standard deviation for each of the reproducibility and repeatability effects are only 2 and 6 percent, respectively. The SNR from the test results is a healthy index of 23.2865 which is significantly larger than the minimum cut-off of 3.

Table C-17. Test Method TP 99 R&R Report: Data Summary Section

Item	Actual Count	Expected Count
Total Values	40	40
Site_Code	4	
Operator	5	
Replicates	2	

Table C-18. Test Method TP 99 R&R Report for: Expected Mean Square and Variance Component Section

Source Term	DF	Expected Mean Square	Variance Component	Lower 90% Conf. Limit	Upper 90% Conf. Limit
Site_Code (P)	3	R+2(PO)+10(P)	11.36034	4.357975	96.88753
Operator (O)	4	R+2(PO)+8(O)	0.009400102	0.0007452315	0.06899036
Interaction (PO)	12	R+2(PO)	0		
Replicates (R)	20	R	0.03681739	0.02344278	0.06786108

Table C-19. Test Method TP 99 R&R Report: Analysis of Variance Section

Source Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
Site_Code	3	340.8949	113.6316	4032.02*	0.000000
Operator	4	0.4135324	0.1033831	3.67*	0.035657
Interaction	12	0.3381875	0.02818229	0.77	0.677127
Replicates	20	0.7363478	0.03681739		
Total (Adjusted)	39	342.3829			
Total	40				

Table C-20. Test Method TP 99 R&R Report: Variance Section

Term	Variance	% Total Variance	Standard Deviation	Lower 90%Conf. Limit of Std Dev	Upper 90%Conf. Limit of Std Dev	% Total Std Dev
Site_Code	11.360344	99.6325	3.3705	2.0876	9.8431	99.8161
Operator	0.009400	0.0824	0.0970	0.0273	0.2627	2.8713
Interaction	0.000000	0.0000	0.0000	0.0000	0.1214	0.0000
Reproducibility	0.005083	0.0446	0.0713	0.0000	0.2565	2.1113
Repeatability	0.036817	0.3229	0.1919	0.1531	0.2605	5.6824
R and R	0.041900	0.3675	0.2047	0.1757	0.3243	6.0619
Total Variation	11.402244	100.0000	3.3767	2.0976	9.8453	100.0000

Table C-21. Test Method TP 99 R&R Report: Percent of Process Variation R & R Section

Term	Lower 90% Conf. Limit	5.15 Std Dev	Upper 90% Conf. Limit	% Total Std Dev	Percent Contribution
Site_Code	10.7510	17.3581	50.6922	99.8161	99.6325
Operator	0.1406	0.4993	1.3527	2.8713	0.0824
Interaction	0.0000	0.0000	0.6251	0.0000	0.0000
Reproducibility	0.0000	0.3672	1.3208	2.1113	0.0446
Repeatability	0.7885	0.9882	1.3416	5.6824	0.3229
R and R	0.9051	1.0542	1.6704	6.0619	0.3675
Total Variation	10.8025	17.3901	50.7032	100.0000	100.0000

Since the % R&R value is less than 10%, the measurement system is excellent.

Table C-22. Test Method TP 99 R&R Report: R & R Indices Section

Index	Lower 90%Conf. Limit	Value	Upper 90%Conf. Limit
Signal-to-Noise Ratio	12.7411	23.2865	66.2589

Since the lower confidence limit of Signal-to-Noise Ratio is greater than 3, the measurement process is adequate.

Table C-23. Test Method TP 99 R&R Report: Means and Bias Section

Term	Count	Mean	Deviation From Target
Overall	40	72.942	
Site_Code			
E	10	70.201	
H	10	74.165	
K	10	77.159	
M	10	70.241	
Operator			
1	8	73.040	
2	8	73.070	
3	8	72.811	
4	8	72.938	
5	8	72.849	
Site_Code,Operator			
E,1	2	70.350	
E,2	2	70.451	
E,3	2	69.948	
E,4	2	70.288	
E,5	2	69.971	
H,1	2	74.150	
H,2	2	74.321	
H,3	2	73.979	
H,4	2	74.260	
H,5	2	74.113	
K,1	2	77.300	
K,2	2	77.287	
K,3	2	77.160	
K,4	2	77.004	
K,5	2	77.047	
M,1	2	70.359	
M,2	2	70.220	
M,3	2	70.159	
M,4	2	70.202	
M,5	2	70.263	

Since there is no target value for TP 98, the deviation from target column is blank.

Data Plots for TP 99 Repeatability and Reproducibility

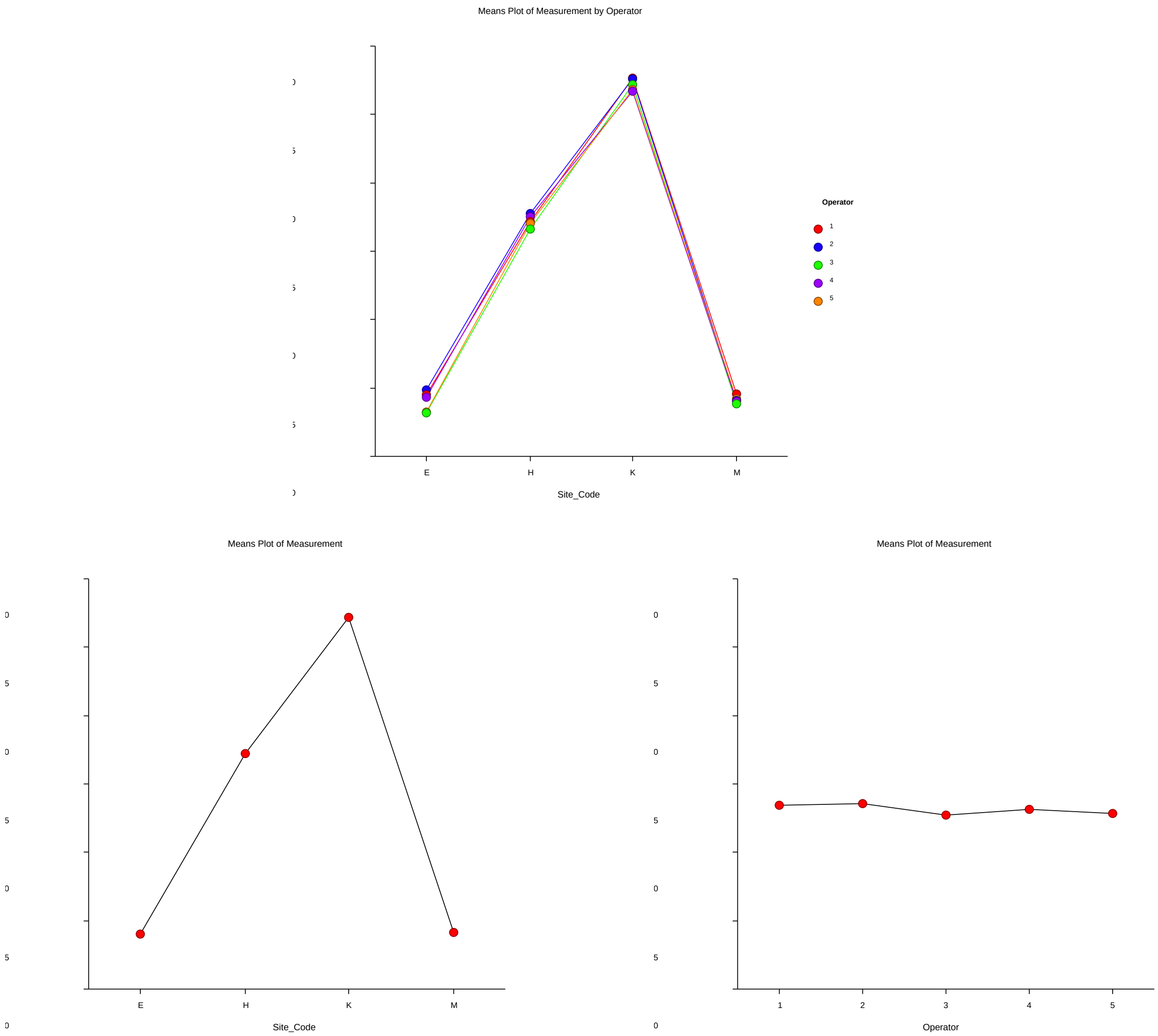


Figure C-7. Test Method TP 99 R&R Study Means Plots

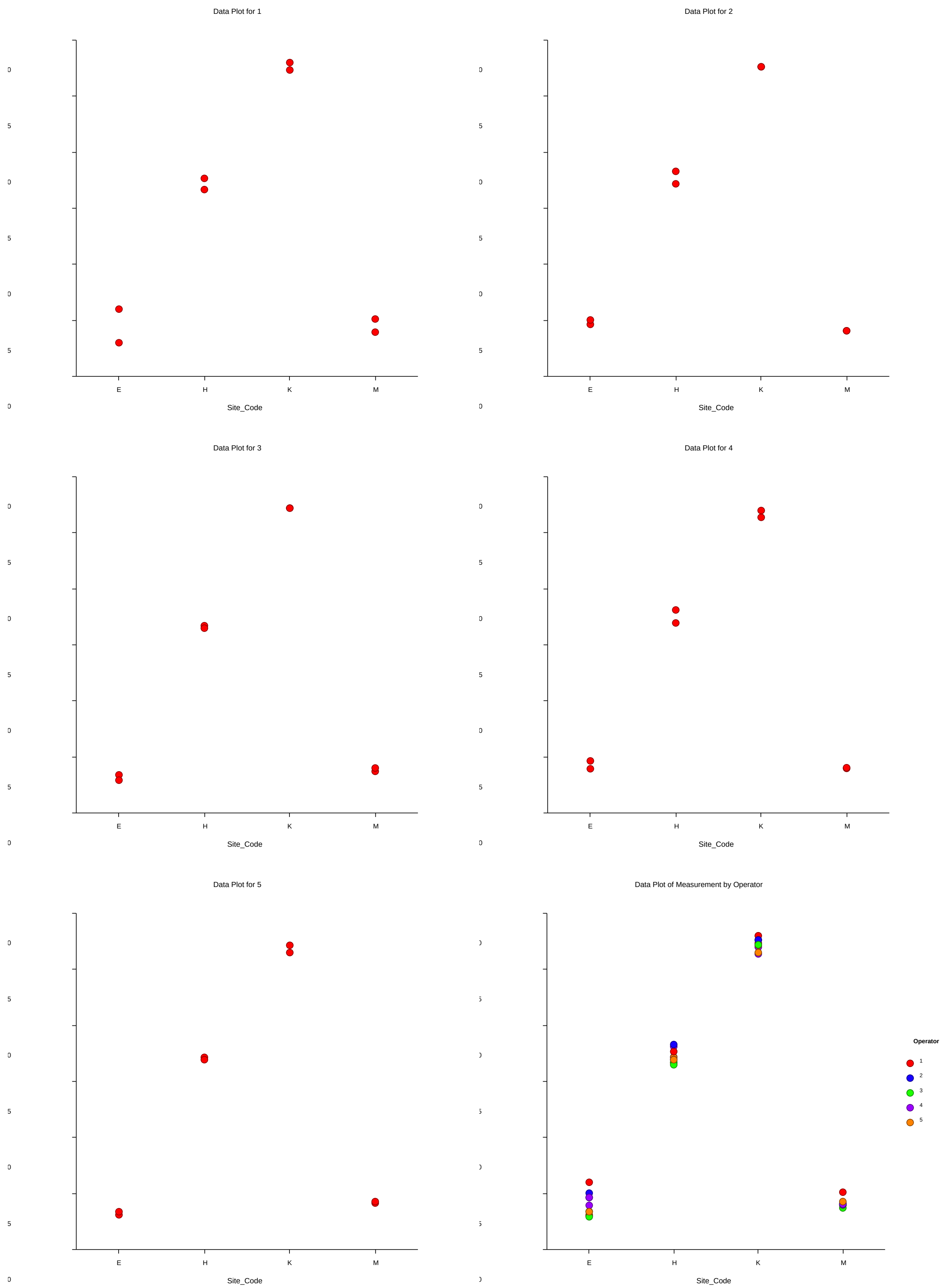


Figure C-8. Test Method TP 99 R&R Study Residual Plots by Operator

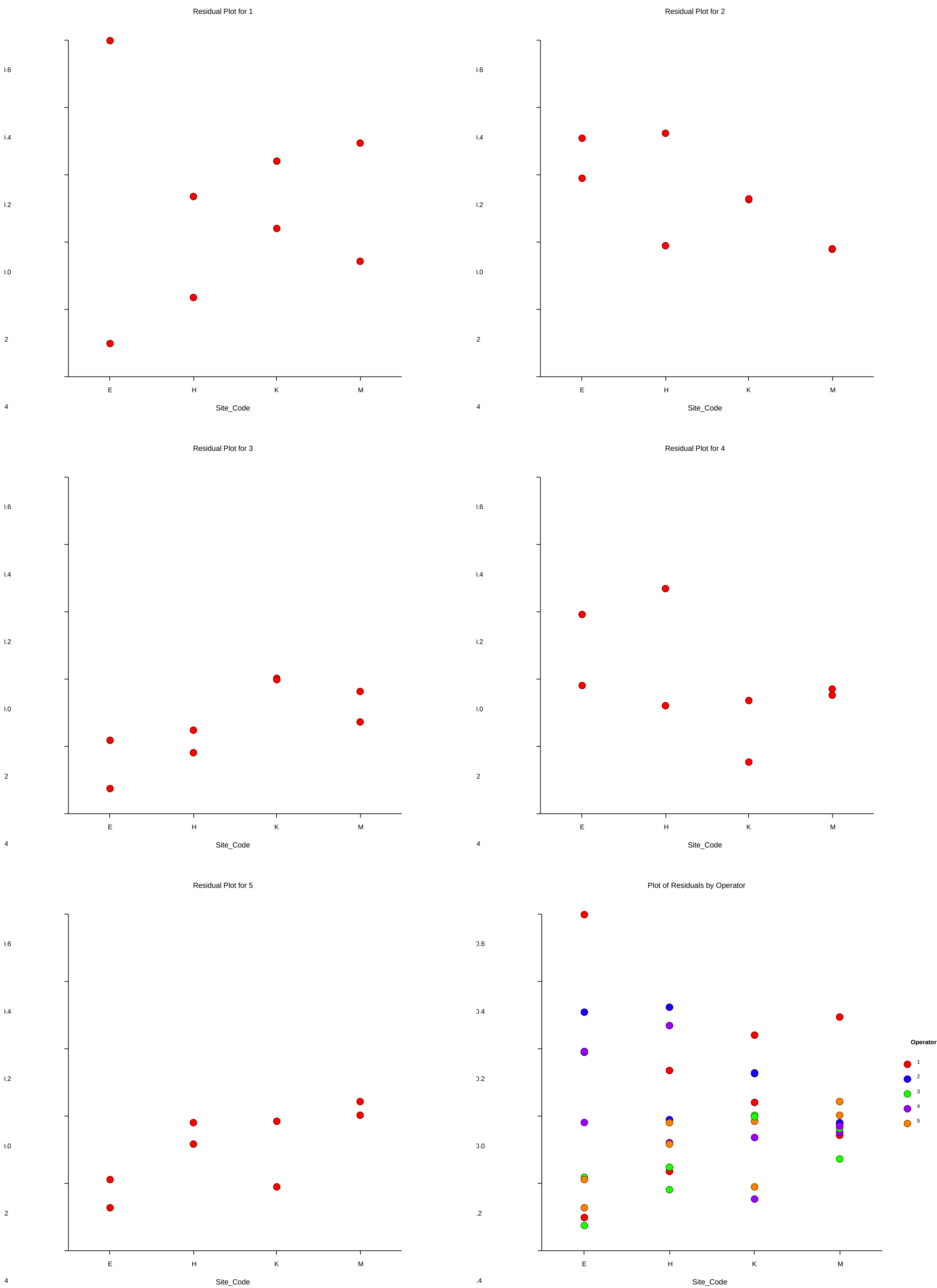


Figure C-9. Test Method TP 99 R&R Study Residual Plots by Operator

APPENDIX D

Complete Statistical Output for Influential Factors/Ruggedness Study

Complete Statistical Results for Regression Analysis Method

Approach to Regression Analysis Method of Ruggedness

The research team used SPSS (IBM® SPSS® Statistics Version 19.0.0) to run a regression analysis. The output from this analysis is used to determine if any of the tested factors influence the measurements.

Results for TP 98 (Automobiles) Regression Analysis Method of Ruggedness

Table C-1. Test Method TP 98 (Automobiles) Ruggedness Regression: Model Summary[b]

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate	Change Statistics				
					R^2 Change	F Change	df1	df2	Sig. F Change
1	.980[a]	.960	.859	1.469	.960	9.543	5	2	.098

a. Predictors: (Constant), Factor_E_air_tempurature, Factor_D_measurement_option, Factor_C_speed_regime, Factor_B_round_impedance, Factor_A_pavement_age
b. Dependent Variable: SIPI

Table C-2. Test Method TP 98 (Automobiles) Ruggedness Regression: ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	102.983	5	20.597	9.543	.098[a]
	Residual	4.317	2	2.158		
	Total	107.299	7			

a. Predictors: (Constant), Factor_E_air_tempurature, Factor_D_measurement_option, Factor_C_speed_regime, Factor_B_round_impedance, Factor_A_pavement_age
b. Dependent Variable: SIPI

Table C-3. Test Method TP 98 (Automobiles) Ruggedness Regression: Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	Constant	.111	.519		.213	.851	-2.124	2.346
	Factor_A	2.148	.519	.586	4.135	.054	-.087	4.383
	Factor_B	-.692	.519	-.189	-1.332	.314	-2.927	1.543
	Factor_C	.765	.519	.209	1.473	.279	-1.470	3.000
	Factor_D	.982	.519	.268	1.891	.199	-1.253	3.217
	Factor_E	-2.496	.519	-.682	-4.806	.041	-4.731	-.262

a. Dependent Variable: SIPI

Table C-4. Test Method TP 98 (Automobiles) Ruggedness Regression: Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-6.972	5.229	.111	3.836	8
Residual	-.764	.764	.000	.785	8
Std. Predicted Value	-1.847	1.335	.000	1.000	8
Std. Residual	-.520	.520	.000	.535	8

a. Dependent Variable: SIPI

Approach for TP 98 (Heavy Trucks) Regression Analysis Method of Ruggedness

Table C-5. Test Method TP 98 (Heavy Trucks) Ruggedness Regression: Model Summary[b]

					Change Statistics				
Model	R	R^2	Adjusted R^2	Std. Error of the Estimate	R^2 Change	F Change	df1	df2	Sig. F Change
1	.986[a]	.972	.900	.806	.972	13.660	5	2	.070

a. Predictors: (Constant), Factor_E_air_tempurature, Factor_D_measurement_option, Factor_C_speed_regime, Factor_B_round_impedance, Factor_A_pavement_age
b. Dependent Variable: SIPI

Table C-6. Test Method TP 98 (Heavy Trucks) Ruggedness Regression: ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	44.316	5	8.863	13.660	.070[a]
	Residual	1.298	2	.649		
	Total	45.613	7			

a. Predictors: (Constant), Factor_E_air_tempurature, Factor_D_measurement_option, Factor_C_speed_regime, Factor_B_round_impedance, Factor_A_pavement_age
b. Dependent Variable: SIPI

Table C-7. Test Method TP 98 (Heavy Trucks) Ruggedness Regression: Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	Constant	-2.138	.285		-7.507	.017	-3.363	-.913
	Factor_A	.345	.285	.145	1.212	.349	-.880	1.571
	Factor_B	-.989	.285	-.414	-3.473	.074	-2.214	.236
	Factor_C	.835	.285	.350	2.932	.099	-.390	2.061
	Factor_D	.254	.285	.106	.892	.467	-.971	1.479
	Factor_E	-1.918	.285	-.803	-6.736	.021	-3.144	-.693

a. Dependent Variable: SIPI

Table C-8. Test Method TP 98 (Heavy Trucks) Ruggedness Regression: Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-6.480	1.696	-2.138	2.516	8
Residual	-.561	.561	.000	.431	8
Std. Predicted Value	-1.726	1.524	.000	1.000	8
Std. Residual	-.696	.696	.000	.535	8

a. Dependent Variable: SIPI

Results for TP 99 Regression Analysis Method of Ruggedness

Table C-9. Test Method TP 98 (Heavy Trucks) Ruggedness Regression: Model Summary[b]

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate	Change Statistics: R^2 Change	F Change	df1	df2	Sig. F Change
1	.993[a]	.987	.953	.765	.987	29.502	5	2	.033

a. Predictors: (Constant), Factor_E_air_tempurature, Factor_D_measurement_option, Factor_C_speed_regime, Factor_B_round_impedance, Factor_A_pavement_age
b. Dependent Variable: SIPI

Table C-10. Test Method TP 98 (Heavy Trucks) Ruggedness Regression: ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	86.233	5	17.247	29.502	.033[a]
	Residual	1.169	2	.585		
	Total	87.402	7			

a. Predictors: (Constant), Factor_E_air_tempurature, Factor_D_measurement_option, Factor_C_speed_regime, Factor_B_round_impedance, Factor_A_pavement_age
b. Dependent Variable: SIPI

Table C-11. Test Method TP 98 (Heavy Trucks) Ruggedness Regression: Coefficients[a]

Model		Unstandardized Coefficients: B	Unstandardized Coefficients: Std. Error	Standardized Coefficients: Beta	t	Sig.	95.0% Confidence Interval for B: Lower Bound	95.0% Confidence Interval for B: Upper Bound
1	Constant	70.909	.270		262.317	.000	69.746	72.072
	Factor_A	1.503	.270	.455	5.559	.031	.340	2.666
	Factor_B	-1.331	.270	-.403	-4.925	.039	-2.495	-.168
	Factor_C	1.628	.270	.492	6.021	.026	.465	2.791
	Factor_D	.366	.270	.111	1.356	.308	-.797	1.530
	Factor_E	-1.991	.270	-.602	-7.366	.018	-3.154	-.828

a. Dependent Variable: SIPI

Table C-12. Test Method TP 98 (Heavy Trucks) Ruggedness Regression: Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	64.090	76.996	70.909	3.510	8
Residual	-.447	.447	.000	.409	8
Std. Predicted Value	-1.943	1.734	.000	1.000	8
Std. Residual	-.585	.585	.000	.535	8

a. Dependent Variable: SIPI

Approach to Stepwise/AICc Method of Ruggedness

The research team used JMP (JMP® Pro 12.0.0) to run the Stepwise/AICc method. The output from this analysis is used to determine if any of the tested factors influence the measurements.

Results for TP 98 (Automobiles) Stepwise/AICc Method of Ruggedness

Table C-13. Test Method TP 98 (Automobiles) Stepwise Regression Control

SSE	DFE	RMSE	R^2	R^2 (adj)	Cp	p	AICc	BIC
107.30833	7	3.915324	0.0000	0.0000	43.71282	1	49.87313	47.63202

Table C-14. Test Method TP 98 (Automobiles) Stepwise Regression Control: Current Estimates

Parameter	Estimate	nDF	SS	F Ratio	Prob.>F
Intercept	0.1105	1	0	0.000	1
A	0	1	36.90264	3.145	0.12653
B	0	1	3.828145	0.222	0.65419
C	0	1	4.6818	0.274	0.6196
D	0	1	7.718521	0.465	0.52073
E	0	1	49.8601	5.207	0.06263

Table C-15. Test Method TP 98 (Automobiles) Stepwise Regression Control: Step History

Step	Parameter	Action	Sig Prob	Seq SS	R^2	Cp	p	AICc	BIC
1	E	Entered	0.0626	49.8601	0.4646	22.614	2	50.4746	44.7129
2	A	Entered	0.0302	36.90264	0.8085	7.5182	3	51.582	38.5664
3	D	Entered	0.1957	7.718521	0.8805	5.9424	4	66.4799	36.8772
4	C	Entered	0.2805	4.6818	0.9241	5.7735	5	118.847	35.3236
5	B	Entered	0.3144	3.828145	0.9598	6	6		32.3243
6	Best	Specific			0.0000	43.713	1	49.8731	47.632

Results for TP 98 (Heavy Trucks) Stepwise/AICc Method of Ruggedness

Table C-16. Test Method TP 98 (Heavy Trucks) Stepwise Regression Control

SSE	DFE	RMSE	R^2	R^2 (adj)	Cp	p	AICc	BIC
16.17561	6	1.6419303	0.6454	0.5863	20.918403	2	40.33552	34.57385

Table C-17. Test Method TP 98 (Heavy Trucks) Stepwise Regression Control: Current Estimates

Parameter	Estimate	nDF	SS	F Ratio	Prob.>F
Intercept	-2.137875	1	0	0.000	1
A	0	1	0.954271	0.313	0.59972
B	0	1	7.826946	4.688	0.08266
C	0	1	5.57947	2.633	0.16561
D	0	1	0.516636	0.165	0.70143
E	-1.918375	1	29.4413	10.921	0.01631

Table C-18. Test Method TP 98 (Heavy Trucks) Stepwise Regression Control: Step History

Step	Parameter	Action	Sig Prob	Seq SS	R^2	Cp	p	AICc	BIC
1	E	Entered	0.0163	29.4413	0.6454	20.918	2	40.3355	34.5738
2	B	Entered	0.0827	7.826946	0.8170	10.861	3	44.3776	31.3621
3	C	Entered	0.0469	5.57947	0.9393	4.2659	4	54.2159	24.6131
4	A	Entered	0.2980	0.954271	0.9602	4.7959	5	106.836	23.3125
5	D	Entered	0.4665	0.516636	0.9715	6	6		22.7119
6	Best	Specific			0.6454	20.918	2	40.3355	34.5738

Results for TP 99 Stepwise/AICc Method of Ruggedness

Table C-19. Test Method TP 99 Stepwise Regression Control

SSE	DFE	RMSE	R^2	R^2 (adj)	Cp	p	AICc	BIC
87.401828	7	3.53355	0.0000	0.0000	143.65986	1	48.23161	45.9905

Table C-20. Test Method TP 99 Stepwise Regression Control: Current Estimates

Parameter	Estimate	nDF	SS	F Ratio	Prob.>F
Intercept	70.909375	1	0	0.000	1
A	0	1	18.06306	1.563	0.25777
B	0	1	14.18048	1.162	0.32247
C	0	1	21.19982	1.921	0.21502
D	0	1	1.073845	0.075	0.79386
E	0	1	31.71663	3.417	0.11401

Table C-21. Test Method TP 99 Stepwise Regression Control: Step History

Step	Parameter	Action	Sig Prob	Seq SS	R^2	Cp	p	AICc	BIC
1	E	Entered	0.1140	31.71663	0.3629	91.351	2	50.2252	44.4635
2	C	Entered	0.1399	21.19982	0.6054	57.05	3	55.7251	42.7095
3	A	Entered	0.1039	18.06306	0.8121	28.12	4	68.4566	38.8538
4	B	Entered	0.0224	14.18048	0.9744	5.8388	5	108.526	25.0025
5	D	Entered	0.3079	1.073845	0.9866	6	6		21.866
6	Best	Specific			0.0000	143.66	1	48.2316	45.9905