UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
NATIONAL VEHICLE AND FUEL EMISSIONS LABORATORY
2565 PLYMOUTH ROAD
ANN ARBOR, MICHIGAN 48105-2498

OFFICE OF
AIR AND RADIATION

May 15, 2013

MEMORANDUM

SUBJECT: EPA Response to Comments on the peer review of *Estimates of the Fraction of the Fleet with High Evaporative Emissions based on the Ken Caryl Station (Denver, Colorado) Field Study (High Evaporative Emissions Field Study)*

FROM: Constance Hart, Assessment and Standards Division
Office of Transportation and Air Quality, U.S. Environmental Protection Agency

THRU: Glenn Passavant, Director, Data and Testing Center, Assessment and Standards Division
Office of Transportation and Air Quality, U.S. Environmental Protection Agency

TO: William Charmley, Director, Assessment and Standards Division
Office of Transportation and Air Quality, U.S. Environmental Protection Agency

EPA contracted with Systems Research and Analysis (SRA), Inc to conduct and coordinate the peer review for the subject report. The three peer reviewers selected by SRA were Dr. H. Christopher Frey (North Carolina State University), Dr. Eric Fujita, (Desert Research Institute), and Mr. Keith Knoll (Czero). EPA would like to extend its appreciation to all three reviewers for their efforts in evaluating this report. The three reviewers brought useful and distinctive views in response to the charge questions.

In cooperation with the contractor who performed the study, Eastern Research Group (ERG), the report has been substantially revised in response to the substantive comments of the three peer reviewers. The final report is now titled "Estimated Summer Hot-Soak Distributions for Denver's Ken Caryl IM Station Fleet" and has a new date of February 8, 2012 with an annotation of May 2013 to acknowledge the aforementioned revisions in response to peer review comments.

In response to the peer reviewers' comments as well as our own realizations of the remote sensing measurement (RSM) weaknesses, the analysis is now based on the portable SHED (PSHED) measurements instead of the RSMs of all vehicles entering the Ken Caryl IM Station during the summer study of 2009. A fundamental change was a revised approach to the analysis which bases the prevalence rates on the PSHED measurements of the 175 participating vehicles in the 2009 study, appropriately weighting the results to account for differential sampling and response rates. The revised approach places less reliance on interpretation of the remote-sensing measurements (RSM) of the 5,000 vehicles that passed the remote sensing device (RSD) monitor during the study, and attempting to relate them

to the PSHED results. This change resulted in substantial reductions in the estimated prevalence rates for newer vehicles, yet higher incidence of leaks in the older model year vehicles.

Responses are given in bold below for individual comments or groups of comments to point to where this was addressed in the revised report.

Responses to Section 3.1 "Specific Technical Comments" in the Peer Review Report

Frey: [1] There are some fundamental questions related to this work that should be part of the objectives and that should be addressed in the technical results and conclusions:

1. Is PSHED a good surrogate for SHED?
2. Can an RSD, if appropriately interpreted, be a good surrogate for a PSHED measurement?

The first question presumes that SHED is the reference method to which all other methods should be compared. What, however, is really measured in a SHED measurement? There are many evaporative processes. Some, such as refueling, are not addressed by SHED. Which processes are addressed?

In what ways are PSHED measurements similar to those of SHED measurements, and in what ways do they differ? Is PSHED effectively just as good as SHED?

What kinds of evaporative processes can be measured using RSD? There is an unstated hypothesis in this report that RSD measurements can provide information on evaporative emissions in a manner comparable to that of PSHED, if only the RSD measurement is appropriate interpreted. What is the basis for this hypothesis? What evaporative processes affect the quantity of HC that is detected by remote sensing? If there was no error in the measurement, would strong concordance be expected between RSD and PSHED? If so, why? A clearer statement of hypothesis and the theoretical underpinning for it would be helpful when interpreting results.

RESPONSE:
During conduction of the study, the PSHED was compared to the laboratory SHED using sets of paired measurements. The results show the PSHED to be a reasonable surrogate for SHED, despite a slight negative bias in the PSHED results, apparently attributable to lower rates of retention and recovery rates in the PSHED. Nonetheless, the degree of correlation between paired PSHED and SHED results is similar to that seen in repeat measurements in either the PSHED or SHED. These comparisons are presented in detail in Appendix A to the revised report. The PSHED is not viewed as a replacement for the laboratory SHED or the evaporative emission test specifications of 40 CFR 86 Subpart B. It appears to serve well as an inexpensive field measurement tool.

Given the nature of evaporative emission processes and the manner in which remote sensing operates, we would not expect that any evaporative emissions arising solely from parked vehicles (e.g., refueling or diurnal emissions) would be reflected in RSD results. It is likely that running losses and fuel vapor arising from fuel/evaporative control system leaks would be reflected. Vapors from these same leaks would also be emitted during the

hot soak and diurnal portions of the traditional evaporative emissions SHED test, but would clearly be more pronounced in the hot soak portion. Based on this limited testing, one might hypothesize a correlation between running losses, hot soak and RSM but this hypothesis would not extend to the other sources of evaporative emissions from parked vehicles.

The revised approach to the analysis does not rely on treating the remote-sensing values as direct surrogates for PSHED measurements. Rather, the remote-sensing was used as an index to guide sampling and improve the efficiency with which vehicles with "elevated" emissions could be identified for measurement. While there is some degree of correlation between the remote sensing and PSHED measurements, this association is stronger for the older vehicles and weaker for newer vehicles having the benefits of improved fuel systems and more stringent emission controls.

Finally, one very important distinction arising from this work is how the PSHED results reflect potential evaporative emission rates. The PSHED test covered measurements for a warmed vehicle parked in an enclosure, with measurement conducted for only 15 minutes. Thus, the emission test results would only reflect hot soak emissions and since most vehicles in the sample were fuel injected it is reasonable to conclude that emissions in the SHED resulted from liquid leaks from the fuel system or vapor leaks from the fuel and/or evaporative emission control systems.

[2] Over the years, EPA has been criticized for making public policy and developing modeling tools to support public policy that are based on proprietary data and methods. The use of proprietary methods precludes a full understanding and review of the underlying science. A case in point are the "Method A" and "Method B" exhaust plume analysis methods associated with the ESP remote sensing instrumentation. Since the distinction between Method A and Method B appears to be an important technical consideration in this study, the lack of disclosure of what these methods are is unacceptable.

RESPONSE:
The differences between Methods A and B are useful in attempting to distinguish exhaust and evaporative components of the emission plume when screening vehicles. However, the differences in the approaches are not inaccessible to readers of the report and are explained in Appendix A. The key difference between the two methods relates to the way in which hydrocarbon (HC) attenuation is related to CO_2 attenuation in calculating the ratio [HC]/[CO_2]. The older method, Method B, used with series 3000 instruments, uses the slope of a simple regression of HC on CO_2 attenuation, which can allow positive intercepts in some cases, presumably attributable to background contamination in the currently measured plume. Method A, by contrast, used with series 4000 instruments, corrects for background levels with each plume measurement. In addition, it assumes that HC attenuation must be zero if CO_2 attenuation is zero. Accordingly, this method relates HC to CO_2 attenuation using a simple ratio-of-means estimator, which can also be thought of as simple regression with the intercept forced to 0.0.

[3] The purpose of the report is to estimate, not develop, fractions of various levels of high evaporative emissions. However, nowhere is any justification or rationale given as to why this

report is focusing on the Denver fleet. Since Denver is at high altitude, and barometric pressure is a factor in evaporation, it is not clear that data from Denver would be representative of other parts of the U.S.

RESPONSE:
Regarding the distinction between the terms "develop" and "estimate", the reviewer's comment is well taken and the title and text has been revised accordingly.

As discussed in the beginning of the report, this work was performed in Denver because it was initiated by the State of Colorado with the purpose of improving the ability to identify vehicles with "elevated" evaporative emissions in the implementation of a repair program implemented in the context of the I/M program. As EPA has shared research interests in this area, EPA collaborated in the effort, and has reported an analysis of the results in this document. EPA is aware that the results of this work, particularly emission rates, cannot necessarily be generalized to other conditions without accounting for differences in conditions such as temperature and pressure.

[4] What is the purpose of "stratification."? Why is achieving stratification a goal in itself? E.g., page 4-3, "to achieve stratification, a higher fraction of vehicles..." The reader can eventually figure this out, but why can't the authors communicate this more clearly? The purpose seems to be to evaluate a screening procedure for identifying vehicles with high evaporative emissions rates, but what about goals for false positives or false negatives?

RESPONSE:
The term "stratification" as used in the draft report was a misnomer and the associated discussion has been fully revised. The correct term is "probability proportional to remote-sensing" or "probability proportional to Index" (ppEI). The reviewer is also correct in surmising that the purpose was to screen vehicles before sampling, to allow vehicles with higher values of the index to be sampled at higher rates, as the revised discussion makes clear. The revised presentation of the analysis also evaluates the index using parameters typically applied to screening measures in epidemiology, sensitivity, specificity, negative predictive value and positive predictive value. Results show that use of the index in sampling was reasonably effective in improving the efficiency with which vehicles with "elevated" evaporative emissions could be identified, and in reducing the cost and effort of measurement. Not unexpectedly, screening is more efficient for older vehicles (at greater ages) than for newer vehicles (at younger ages) having the benefits of improved materials and fuel systems, coupled with advanced emissions controls.

It is important to note that the presence of false positives and negatives in the screening process does not obviate the usefulness of the measurements on these vehicles in estimating leak frequencies. In analyzing the results, the probability with which each vehicle was sampled determines its assigned weight (i.e., inverse-probability weighting), not the value of PSHED measurement obtained after sampling and recruitment. Accordingly, a "false negative" receives the same (higher) weight in the analysis as a "true negative" and a "false positive" receives the same (lower) weight as a "true positive." Thus, if the index classifies vehicles accurately, it can greatly reduce the effort and cost required to conduct inspections and measurements and estimate the prevalence of elevated evaporative emissions based on PSHED results. However, if the index

performs poorly if it fails to guide sampling efficiently, resulting in a situation more similar to sampling fully at random. Nonetheless, because the sampling probabilities for each measured vehicle are known, the resulting set of measurements can be still be used to estimate the prevalence of elevated emissions based on PSHED results.

[5] Is it literally the case that six RSDs were used? i.e., six remote sensing devices at six locations? Or were the two highway "RSDs" based on repeated passes by the same RSD? The authors need to stop using the term "RSD" to refer to a measurement. RSD = Remote Sensing Device and refers to an instrument. A measurement made using an RSD could be described as a remote sensing measurement. What is an RSD beam block? This is shop jargon (I know what it means, but most readers won't).

RESPONSE:
We appreciate the commenter's perspectives on terminology and have revised the report accordingly. There were three remote sensing devices, one where the vehicle passes at approximately 12 mph, one at approximately 34 mph and one at approximately 55 mph. The study collected readings for two passes of each devise after the initial recruitment pass. In the draft analysis provided to the reviewer, multiple remote-sensing measurements were used, one for sampling, and additional measurements to characterize vehicles' emissions levels. In the revised analysis, however, only the initial measurement used to calculate the screening index was used; the six additional remote-sensing measurements were not used. As described in the revised text, the initial measurement obtained for screening purposes prior to sampling was used in assigning sampling fractions (and weights) to vehicles. However, as mentioned, the additional measurements were not used in the analysis and are not discussed further in the body of the revised report.

The term "beam block" has been removed from the body of the revised report but is now defined in the glossary of terms in front of the report since it is used in the appendices.

[6] What is the 'standard I/M inspection' – for those of us not from Denver, please explain what this is. Also, explain the "Modified California Method" – both of these should be documented in the new methods chapter that needs to be written. Who does the olfactory examination? What is an 'electronic HC sniffer"? Is this relevant to the report? If not, then delete mention of these.

RESPONSE:
The parameters of the Denver I/M program are described in the background section of the revised report but not discussed at length because they are not germane to the current project. The role of the inspecting mechanic in the process has been explained and the "Modified California Method" is also described in the methods section. Also, the informal term "sniffer" has been replaced with "hydrocarbon detector" and the instrument identified.

[7] Page 4-4: Method A was used on ESP 4000 and 4600 instruments, and Method B was used on ESP3000 series instruments. Yet, results for both Methods A and B are reported in Table 4-2. Were two RSD instruments used at each RSD site? Or were both Methods A and B applied to the same data measured from just one RSD instrument at each site? At the end of the

paragraph is it mentioned that 'code' was 'added' to the 4000 and 4600 series instruments – it would have helped if this was mentioned up front, and if there was a prior section that more clearly disclosed the study design in terms of what instruments were deployed at what locations and what the vehicle path was through each RSD site. It would help if this text were reorganized so that there was an intro paragraph, one paragraph on Method A, one paragraph on Method B, and then a paragraph that compares Methods A and B. Are the CO, NO, and CO2 results shown in Table 4-2 based on Method B? The distinction between Methods A and B with respect to how they deal with exhaust versus evaporative concentrations of HC is not clear. To merely state that "ESP believes" that one method is responsive to exhaust and another is not is quite tenuous.

RESPONSE:
As mentioned above in the response to a previous comment on this topic, Methods A and B refer to algorithms used to calculate the ratio of hydrocarbon concentration to CO2 concentration in the emissions plume by relating attenuation for HC to that for CO2. Each begins with a regression of HC attenuation to CO2 attenuation. An initial difference is that Method B allows non-zero intercepts for this regression whereas Method A forces it through the origin (other differences follow in the calculation of the concentration ratio). As algorithms applied to the raw spectroscopic absorbences, either calculation can be applied to results obtained from 3000 series or 4000 series ESP instruments (or any other instrument). In this project, software for a 4000 series instrument was modified so as to allow conduction of a "method B" calculation of the HC concentration ratio.

The calculation of the screening index (EI23) for each vehicle entering the station was based on a single pass by the remote-sensing van (and collection of a single set of absorbances). After screening, additional remote-sensing measurements were collected but they were not used in the analysis reported in the revised report and are not further discussed. The calculation of EI23 is also based on a simple regression of HC on CO2 attenuation. This regression is similar to the Method B regression in that it allows non-zero intercepts but not identical in that not all absorbances are included. However, in transforming and classifying raw EI23 values into values of EI23 Bin, the index actually used, a "Method B" calculation of the HC:CO_2 ratio is used to distinguish the exhaust component of the plume.

The material in the draft concerning methods A and B and their responsiveness to exhaust vs. evaporative hydrocarbons is of interest as background material but is not relevant to the manner in which the screening index EI23 Bin was calculated. The procedure used to assign raw EI23 values to corresponding values of EI23 Bin involves use of the HC concentration ratio for each vehicle. The value used was calculated by method B, but a value calculated by another method could also have been used. These steps are described in more detail in Appendix A to the revised report.

[8] Page 4-4 (bottom): regression toward the mean.... This is stated as if it is an underlying principle in a rather didactic manner, but the actual concept is poorly explained here. A measurement is biased if it is systematically high or systematically low. If the error is randomly distributed with a mean of zero, then the measurement is subject to random error, not bias. The random error can lead to false positives or false negatives if used in the context of a binary decision (e.g., vehicle is a high emitter). This context is not clearly articulated.

False positives or false negatives are not necessarily a result of bias, but rather a result of imprecision (random error). The discussion here of bias is thus without sufficient context and therefore is unclear.

RESPONSE:
The reviewer is correct that the discussion in this paragraph is unclear and does not convey the intended meaning. What was meant was that there was potential to misclassify some vehicles based on the remote sensing as having "elevated" evaporative emissions if the remote-sensing value used for selection was unusually high. The misclassification could occur because successive remote-sensing values would be unlikely to be similarly high and would suggest that the initial value had overestimated the vehicle's actual level, i.e., if the initial value was unusually high, successive values would "regress to the mean."

However, the discussion of "regression to the mean" in the draft report is not relevant to the analysis presented in the revised report and has thus been removed.

[9] What role does ambient temperature have in contributing to variability in estimated evaporative emissions based on RSD measurements? Since the "Temperature" in Table 4-2 (ambient temperature at the time of each RSD measurement?) differs from the PSHED "Seal Temperature", what role might this have in confounding the results?

RESPONSE:
Remote-sensing measurements are highly variable in all circumstances and it is possible that ambient temperature contributes to this variability as do other uncontrolled factors such as wind or vehicle speed. Thus, any factor, such as temperature, that reduces the degree of association between the remote-sensing and PSHED measurements would also tend to reduce the efficiency of the screening index. However, as mentioned above, as the screening index is used only to assign sampling weights, the degree of "confounding" between the remote-sensing and PSHED measurements does not reduce the utility of the PSHED measurements.

[10] Table 4-2: what is the meaning of negative values for HC Method A (ppmC 3) and how are these interpreted? Table 4-2 values of CO_2 percent appear to be what one would expect in the tailpipe, but this cannot be what was actually measured in the exhaust plume. How is the air-to-fuel ratio inferred, or is it assumed to be stoichiometric? Some discussion is needed. The text barely alludes to this. More detail is needed in a methods chapter. Is RSD temperature the ambient temperature at the date and time of the measurement?

RESPONSE:
The information on the remote-sensing measurements in Table 4-2 in the draft is now in Appendix F as Table F-2. The hyphen (-) is not a negative sign, rather a flag representing a temporary interruption of data collection. This outcome is explained in step (e) of the "Calculation of RSD Evap Index 23 (EI23)" in Appendix B. In the estimation of plume concentrations, stoichiometry is assumed. The "RSD temperature" is the ambient temperature at the time of measurement.

[11] The quantity in Figure 4-1 labeled as "RSD EI23" needs to be clearly defined. Is this based on any numbers given in Table 4-2? Which specific column of Table 4-2 is "RSD EI23"? Which specific column of Table 4-2 is "PSHED Mass (g/Qhr)"? Presumably, "Measured PSHED HC at 15 Minute Soak (grams)" in Table 4-2 is the same as "PSHED Mass (g/Qhr)". However, use consistent terminology in both places to avoid ambiguity. The EI23 values need to be added to Table 4-2.

RESPONSE:
The report has been revised in response to this comment. In Table F-2, (formerly 4-2), "RSD EI23" is in the column labeled "EI23." However, the PSHED measurements (q/Qhr) are no longer in Table F-2. They have been moved to Table F-3.

[12] Page 4-22: what role does ambient temperature have in the estimation of EI23? The RSD measurements are made at ambient temperature. Evaporative emissions are proportional to ambient temperature (something that needs to be introduced and discussed in a background or methodology section of this report). Is the EI23 metric less responsive to evaporative emissions at lower ambient temperature? Speed is not the only factor that affects inference of evaporative emissions.

RESPONSE:
At this time, the effect of ambient temperature on EI23 is not known. In addition to other factors, it is plausible that temperature would contribute to the variability of the remote-sensing measurements, at least with respect to any unburned hydrocarbon vapors detected. If this were true, increased temperature could reduce the efficiency of screening. However, it is also plausible that leaks, if they exist, become more obvious and easy to detect with increasing temperature, increasing the efficiency of screening. None the less, one must be careful not to infer a strong relationship between evaporative emission rates from a moving vehicle and ambient temperature. While there may be a modest effect, the primary relationship would be between fuel temperature and emissions. Fuel temperatures would likely be higher than ambient due to fuel recirculation, proximity of other warm components, road radiant heat load, etc. These effects would be tempered by the cooling effect of ambient air moving past the warmer surfaces.

[13] Page 4-23: Why is model year important? Earlier, a note was made that model year was not part of the EI23 binning method.

RESPONSE:
We analyzed the results by Model-Year group because model-year is a surrogate for important changes in fuel-system and emissions control technologies. In this study it also serves as a surrogate for vehicle age. Results show large differences in hot-soak distributions by model-year group.

Ideally, we would have stratified vehicles by model year while simultaneously screening for emission levels. However, in this project the determination of whether vehicles were drawn into the sample was made immediately after they passed the remote-sensing unit to allow recruitment to occur before vehicles entered the I/M station. It was thus not practical to acquire model-year information for incoming vehicles to incorporate into

sampling. However, as the sampling assignments for all vehicles were made independently, we have analyzed the subsamples of vehicles in each model-year group as independent sub-samples of the fleet.

[14] If there are multiple EI23 bin values available for some vehicles, these data should be analyzed separately to determine the robustness with which a vehicle is assigned to an EI23 bin. Ambiguity in assignment to an EI23 bin would be a significant factor to consider in evaluating the usefulness of this method.

RESPONSE:
It would have been preferable to have employed multiple remote-sensing measurements in the calculation of EI23 and assignment of EI23 Bin. However, in this project, the assignment of the screening indices was made based on single measurements acquired as vehicles entered the I/M station. We agree that improving the reliability of bin assignment would probably improve the performance of the screening index.

[15] Table 4-5. The table is actually of EI23 bins and model year groups, not screening remote sensing measurements. Thus, the caption is not consistent with the content of the table.

RESPONSE:
In the revised report, these tables have been replaced with Tables 4-1 and 4-2 and the captions revised accordingly.

[16] Page 4-25: The terms sample and population in the Appendix B need some careful re-thinking or at least more clear definition. Here, the term 'population' is implied to describe the total sample of 5,830 vehicles (which is actually a sample from a larger fleet). That is okay, but at least be clear as to the meaning of the term 'population' as used in Appendix B. W_h is the fraction of the population of vehicles that fall into each EI23 bin. It is not clear as to the definition of "n" in Appendix B – is this the total number of vehicles in the 'population'? (i.e. n=5830?). L=7 (could be stated clearly). The term σ_h is not clearly defined in appendix B in terms of other variables. Is this the standard error of the fraction of elevated measurements in each strata? Appendix B does not actually show how one estimates the estimated fraction of the population that is above the threshold. How was the value 0.127 estimated? This appears to the product $p_h W_h$ summed over all h. Based on the numbers given in Table 4-6, over 75% of the estimated 'elevated PSHEDs' (a sloppy term) are from Bins 1-4, which account for over 96% of the 'population.'

RESPONSE:
We concur with the commenter's basic input regarding term definitions and explanation of the calculations and outcomes. Appendix B in the draft report is no longer relevant and has been removed from the revised report, as the revised report clarifies that the sampling method actually used was "probability proportional to Index," not "stratified sampling."

[17] Table 4-8. It is not very clear as to what variable is implied by "High-PSHED Fraction..." is this based on p_h and W_h defined in some different way compared to Table 4-6?

RESPONSE:
In the draft report, the term "High-PSHED Fraction" indicated the fraction of vehicles in a fleet expected to have PSHED measurements exceeding a specified threshold. This terminology has been revised in the final report.

[18] The assumption of the EI23 bins is that they are bins of EI23 values. Since no assumption is made regarding model year, it is not really correct to imply that if there is a dependency on a model year that somehow the use of EI23 is inherently inappropriate. It could be that the fraction of vehicles with high PSHEDs measurements is correlated with EI23 and with model year, but that does not imply that EI23 would not be a useful indicator. Whether EI23 is a useful indicator can be determined with or without consideration of model year. In fact, if EI23 has a trend with respect to model year that is consistent with the trend with respect to PSHED measurements, then there might be increased confidence in the utility of EI23 as an indicator.

RESPONSE:
We agree with the reviewer that the existence of a dependency of emissions on model-year does not obviate the usefulness of EI23 as an index. Nonetheless, the results do suggest that that hot-soak emissions are dependent on model-year group (as expected), and that this relationship implies that the efficiency of the index would be lower for more recent model years and higher for older model years. One reason for this expectation is that the "positive predictive value" of a screening measure is related in part to the prevalence of the trait being screened, i.e., the rarer a "trait," the more difficult it is to screen for. Thus, vehicles manufactured since 1996 and having the benefits of enhanced emissions controls and OBD systems, have generally lower emissions and are thus screened less efficiently than older vehicles. The index is nonetheless useful in reducing cost and effort.

[19] Section 4.5: the discussion here suffers from a conceptual problem related to not clearly defining what is meant by "uncertainty." The term uncertainty is used inappropriately as if it refers only to imprecision, and the notion of bias is discussed as if it distinct from "uncertainty." Uncertainty refers to lack of knowledge regarding the true value of a quantity, and includes both random and systematic sources of error. Random error is imprecision. Systematic error is bias and also known as lack of accuracy. Thus, bias is a component of uncertainty, not distinct from it.

RESPONSE:
The reviewer is correct that the text in the draft report used the terms "uncertainty" and "bias" and "error" in vague and more or less synonymous senses. Much of the discussion centers on the "uncertainty" or "bias" in attempting to predict hot-soak (PSHED) results from small sets of remote-sensing measurements and then to draw inferences about fleet behavior from the "predicted PSHEDs" rather than from actual hot-soak measurements.

However, the revised analysis obviates this issue by placing much less emphasis on the ability of remote-sensing to accurately predict expected hot-soak results. The remote-sensing measurements obtained prior to the hot-soak measurements have been

relegated to their proper role as a screening index used to assign sampling probabilities for individual vehicles.

The uncertainty in the screening process has been recast in terms appropriate for a screening measure using the four parameters commonly used in epidemiology: sensitivity, specificity, negative predictive value and positive predictive value.

[20] Uncertainties associated with small sample size are typically quantified based on random sampling error. The discussion of the role of 'chance alone' is inappropriate as written. Perhaps the intended statement is that if a different random sample of vehicles had been selected, the number of vehicles with PSHED measurements greater than 2 g/Qhr might have been different from the 2 that were observed in the available sample. Because the fraction of vehicles with PSHED measurements greater than 2 g/Qhr is based on a sample, there is 'sampling error' in the estimate. If the sample is assumed to be random, then the error of the estimate can be estimated based on sampling distributions of the statistics (a statistic is a quantity estimated from a sample). The errors shown in Table 4-11 are of unclear basis. For example, the 'size of error for 'high PSHED Definition' of 2 is given as 0.025. There should be more detail on how this number was estimated, based on the data given in Table 4.6.

RESPONSE:
In the revised report, estimates of sampling error have been recalculated and presented with the summary results. For each fraction, "exact" Clopper-Pearson confidence intervals have been calculated, rather than relying on the more commonly used normal approximation to the binomial.

[21] PSHED measurement error should be more clearly discussed. The text refers to 'two parts' but really only one 'measurement error' is actually addressed. Measurement error typically refers to the imprecision and bias of the measurement method itself. Propane retention and recovery tests are an incomplete indicator of the imprecision and bias of the PSHED method, because actual evaporative emissions are not pure propane. Variability in hot soak emissions is a measurement error only in the context of attempting to assess the repeatability of measurements of the same vehicle under the same conditions. However, it is not clear that such an experiment has actually been done. If there are underlying differences in the state or condition of the vehicle, then the variability in the measurements is not because of the measurement method itself but because of the state of the vehicle being measured. The concept of repeatability of the measurement should be discussed in a separate paragraph or subsection. If the repeatability is only -50% to +200%, then there is significant question as to the usefulness of any kind of PSHED test when compared to a 'brightline' threshold that is a point value.

RESPONSE:
The retention and recovery tests give a indication of the ability of the enclosure to retain emitted gases during the duration of a measurement period. The use of a specific gas, such as propane, facilitates this step, and is standard laboratory practice, even in procedures as rigorous as emissions certification.

However, we agree with the reviewer that assessment of overall repeatability of the PSHED approach involves consideration of additional factors determining vehicle

condition, although not all such factors can be determined or controlled for. In the revised report, we addressed the effects of measurement repeatability by reestimating the fractions of "elevated" emissions for two scenarios. The first, a "lower-bound" scenario, assumes that the measured values systematically overestimated the "true" emissions levels for all vehicles. The second, an "upper bound" scenario, assumes that the measured values systematically underestimated the "true" values for all vehicles. The two scenarios were developed based on a prediction interval obtained from a simple log-log regression applied to the set of paired PSHED measurements presented in Appendix D to the revised report.

The results of this analysis suggest, most importantly, that the calculated frequencies of "elevated" emissions are not highly sensitive to an assumption that all PSHED measurements over-estimated the "true" emissions levels for all vehicles by a margin of 50%.

[22] The discussion of detection limit and how it was inferred is difficult to follow. First, it would help to define what is meant by detection limit. It is not clear how a detection limit can be inferred by making a measurement on a vehicle or any sample for which it is not known as to whether the HC concentration is actually zero. Why not use a 'zero' calibration gas that contains 0 ppm of HC? A baseline before a vehicle enters the PSHED does not guarantee that actual concentration was 0 ppm of HC. However, it does provide a background level. However, the text does not discuss what is background or the role of background in making measurements.

RESPONSE:

We agree that the text in the draft was ambiguous and used the term "detection" in a non-standard sense. The more appropriate term used in the revised report is "quantitation." The text in the revised report clarifies that what this discussion describes is an estimate of a lower limit of "quantitation" for the PSHED, or an estimate of the lowest hydrocarbon emission rate that the method can quantitatively distinguish from expected background levels.

[23] Page 4-32: the analysis of duplicate EI23 measurements is quite important, and the text refers to Appendix A. Appendix A is very poorly written and very unclear. It is not apparent that there are any data regarding the duplicate EI23 values in the main body of this report or in the appendix. The data and findings from these data should be disclosed.

[24] The rational for the bias in the EI23 values and the implication that it would 'tend to elevate the high-PSHED fraction' needs to be more clearly articulated.

RESPONSE to [23] and [24]:
This discussion is irrelevant in the revised report and has been removed. The accuracy with which vehicles were assigned to EI23 Bins was of critical importance in the analysis presented in the draft report, as remote-sensing values were used directly to estimate the fractions of vehicles with "elevated" emissions. As mentioned, in the revised analysis, the values of the EI23 screening index are used solely to assign inverse-probability sampling weights to measured vehicles. The efficiency of the index in guiding

sampling is assessed using parameters typically used for this purpose. Accordingly, in cases where the measured PSHED suggested that vehicles had been flagged as false positives or negatives, these vehicles retained their assigned sampling weights in analysis. Once a vehicle was drawn into the sample, its sampling weight is fixed, and is not modified based on the outcome of its PSHED measurement.

[25] Page 4-33: the apparent confusion regarding detection limit and background level is evident in the second paragraph on this page. One does not subtract a detection limit from a measured value to impute an unbiased estimate. This would be done only for a background level. However, if the background is negligible compared to the measurement, this will have little effect on the results.

RESPONSE:
The reviewer's observation is correct. See response to item [22] above.

[26] The discussion of a possible Monte Carlo simulation is so vague that it hardly merits being in this report. Unless the authors can clearly define terms and propose a meaningful algorithm, the recommendation for future Monte Carlo simulation could be stated briefly, with further development left to those competent to conduct such an analysis.

RESPONSE:
We agree that this discussion did not contribute to the draft report. Accordingly, it has been removed from the revised report.

Fujita: [1] While the EI23 evaporative index would be useful for identifying gross evaporative HC emitters, its ability to estimate fractions of high evaporative emissions within various levels of evaporative emission other than the top end of the distribution seems limited.

RESPONSE:
We agree. Reassessment of the capabilities of the remote-sensing measurements and EI23 was a catalyst for us to revisit our analysis, resulting in a complete revision. The analysis in Section 4.0 is now focused on PSHED values of the 175 participants in the Ken Caryl study.

[2]Conversion of EI23 measurements to Bins provides what appears to be clearer summary of the distribution of EI23 values by PSHED-equivalent running loss levels. As I understand this procedure, this classification assigns the estimated evaporative indices into bins with width that each corresponds to one standard deviation of the variability of a single EI23 measurement (after accounting for the effects of the exhaust HC emissions on EI23). The EI23 Bins are then associated with probabilities of exceeding various threshold PSHED hot-soak emission levels. This approach allows the association to be made without regard to the quality of the correlation between EI23 and PSHED hot-soak levels, which we know is poor. EI23 values in at least the first three EI23 Bins (with PSHED thresholds of greater than 1, 2 and 5 g/Qhr) are probably below the method limit of detection and are really random noise. If so, there is about equal chance that any of the EI23 values in the first three Bins has a corresponding PSHED above the threshold. Therefore, it is not unexpected that fractions of elevated PSHED in Table 4-6 are about the same for Bins 1 (6.7%), 2 (7.6%) and 3 (9.6%). These fractions are likely not valid given the measurement sensitivity. If 20g/Qhr is a reasonable

level where the corresponding EI23 values become reliable, then the distribution shown in Table 4-4 for this High PSHED definition is valid for all EI23 Bins. The fractions are progressive less reliable for the lower EI23 Bins at lower thresholds values.

I believe the net result is an overestimation of the fractions of elevated PSHEDS in the lower Bins. Products of these fractions with the proportionally larger numbers of vehicles in these bins for the Random fleet will result in larger fractions of elevated PSHEDs in the larger fleet of vehicles. For example, results of the de-stratification calculations in Table 4-6 shows that 12.7% of the 5830 vehicles in the random sample are estimated to have corresponding high-PSHEDs defined as greater than 2 g/Qhr. If the first three Bins are counted as zero, then this fraction drops to 5.5%. Also dropping Bins 4 and both 4 and 5 reduces the fraction to 2.9% and 1.6%, respectively. The more appropriate fraction is likely between 1.6 to 5.5% rather than 12.7%.

RESPONSE:
We agree with reviewer's comment. Having reassessed the approach presented in the draft report, we too have come to the conclusion that fractions of elevated PSHEDS cannot be reliably estimated using remote-sensing measurements alone. This conclusion applies particularly to lower emission levels in the PSHED (<10 g/Qhr), and for vehicles manufactured since 1996. Accordingly, the analysis in the revised report was fundamentally revised consistent with the study design, with the values of EI23 Bin used solely to assign inverse-probability sampling weights to measured vehicles. However, the emissions level of each vehicle is assigned solely based on its PSHED measurement.

[3] It should also be noted that the distributions are presented without quantitative estimate of uncertainty and bias that are inherent in the study approach. In addition to the poor limits of detection of RSD evaporative index, the following sources of uncertainty and bias were not assessed in the report.

- The distributions are based on static SHED 15-minute hot-soaks and do not include diurnal evaporative emissions and may not fully account for all running emissions.

RESPONSE:
We concur that the 15-minute static PSHED measurement does not account for diurnal or running-loss emissions . The PSHED procedure was not intended to account for all emissions sources simultaneously, but rather to provide a quick and inexpensive assessment of evaporative emissions, for purposes of identifying vehicles with "elevated" evaporative emissions. Additionally, laboratory measurements of vehicles with artificially "implanted" leaks show that emissions through all modes of vapor venting, including diurnal, hot-soak and running loss, show increased vapor emissions when leaks are present. An emissions measurement acquired in the field need not account for all emissions modes to be effective at finding a problem resulting in higher emission levels.

- The residual hydrocarbon signal in the RSD measurements in excess of the regression line of HC with CO_2 results is a crude measure of the diluted mixture of evaporative emissions from fuel permeation, vaporize fuel leaks, and fuel system venting during vehicle operation. Unlike exhaust pollutant, there are no tracers for evaporative HC emissions to account for dispersion rate of emissions.

RESPONSE:
We agree that the measure of evaporative emissions obtained through remote sensing is simple and crude. However, when properly applied, it is not necessary that the index provide a highly accurate or precise measurement of evaporative emissions. To serve its purpose in improving the efficiency of sampling, it is sufficient that it show a reasonable level of association with a more rigorous measure of emissions, at least for hot-soak rates above a certain level. Results presented in the revised report show that the index shows noticeable gains in efficiency for recognizing PSHED levels above about 1.0 g/Qhr.

- Replicate LSHED and PSHED tests have large variability. Section 4.5 does not address the significance of the large variability of replicate SHED tests to distribution of fractions of "high evaps" at various definitions.

RESPONSE:
The reviewer is correct that the discussion in the draft report did not account for the effect of variability in the PSHED results on the estimated fractions of "high evaps." In the revised report, we addressed the effects of measurement repeatability by reestimating the fractions of "elevated" emissions for two scenarios. The first, a " lower-bound" scenario, assumes that the measured values systematically overestimated the "true" emissions levels for all vehicles. The second, an "upper bound" scenario, assumes that the measured values systematically underestimated the "true" values for all vehicles. The two scenarios were developed based on a prediction interval obtained from a simple log-log regression applied to the set of paired PSHED measurements presented in Appendix D to the revised report. The results of this analysis suggest, most importantly, that the calculated frequencies of "elevated" emissions are not highly sensitive to an assumption that all PHED measurements over-estimated the "true" emissions levels for all vehicles by margins of 50%.

[4] Ambient temperature was not included as a variable in the study design and PSHED and replicate RSD measurements were all made within a short time at about the same temperature. The test sets within each EI23 Bin were conducted at ambient temperature spanning a range of up to about 30°C. Evaporative emissions are known to increase with ambient temperature with doubling of permeation for 10°C rise in temperature. This likely would not be issue if ambient temperature was a random variable in the study and test sets within each bin had similar random distribution of temperature. Was this checked? The potential bias due to differences in temperature would be minimal for the high emitter bins, but may be more important for the other bins.

RESPONSE:
There is a significant relationship between fuel temperature and evaporative emissions, but a weaker relationship between ambient temperature and evaporative emissions except for the diurnal mode. Thus, we conclude that especially for this ambient temperature range, the effect of ambient temperature is not a first-order consideration for the hot soak measurements, relative to the importance of the temperatures of the engine, exhaust system, fuel system and the fuel itself.

[5] Most vehicles in Bins 6 and 7 had high exhaust HC emissions, which can contribute to the estimated evaporative emissions. The report asserts that this positive interference is mitigated by the binning procedure. From the relevant discussion in Appendix A, it is difficult to determine the significant of the positive interference or the effective of the binning procedure. [*We interpret this comment to indicate a concern of the reviewer that the presence of high exhaust emissions in the plume would interfere with the ability of the instrument to detect high evaporative emissions.*]

RESPONSE:
The method used to calculate the EI23 Bin index attempts to discount for the presence of exhaust emissions in the emissions plume. It is difficult to assess with certainty the effectiveness with which the calculation can achieve this goal. However, the practical effectiveness of the screening index is assessed in the revised report using appropriate measures, including sensitivity, specificity, negative predictive value and positive predictive value.

[6] P. 1-2, line 5. Are there plans for follow-on uncertainty analysis that can be described here?

RESPONSE:
While it is difficult to assess the effects of specific sources of error, an assessment of the effect of overall measurement repeatability (repeatability) on the estimated fractions of vehicles with "elevated" emissions is presented in the revised report.

[7] P. 2-2, second full paragraph: Describe briefly the evidence, with appropriate references, that previous estimate of "high evaps" were lower that what is occurring in the real world.

RESPONSE:
The analysis presented in this report itself makes the case that the prevalence of vehicles with "elevated" evaporative emissions as estimated in this project are higher than would have been assumed prior to conducting this work.

[8] P. 3-14, last sentence: Meaning is unclear. Why would large variability of PSHED hot-soaks itself result in overestimation of fraction of vehicles with high hot-soak emissions?

RESPONSE:
The discussion in this paragraph in the draft report simply presents this conclusion without justification. If we assume that "measurement error" (from all sources) tends to function as a component of random error, i.e., with a mean of 0.0 and variance proportional to vehicles' emissions levels (assuming no biases), it is not possible to ascertain from single measurements whether vehicles' "true" values are higher or lower than the measured values on hand. However, the uncertainty analysis presented in the revised report suggests that the estimated frequencies of "elevated" emissions are more sensitive to underestimation than to overestimation of "true" values. This result can be attributed to the fact that the prediction interval is not symmetric about the mean level, but is asymmetric, with the upper half several times broader than the lower half of the interval.

[9] P. 4-25, Table 4-6: What is the basis for S_h in the calculation of standard error of the fraction of elevated PSHEDs? What are the sources of the values used in calculating the standard deviation?

RESPONSE:
This table in the draft report is based on the assumption that the sample was collected using "stratified sampling." As the revised report makes clear, however, the method applied is more appropriately called "probability proportional to Index." The calculations in the revised report were modified appropriately, removing the references to "stratification" as presented in the draft report.

[10] P. 4-30, Table 4-10. Unless there is good reason for using natural log, give estimated error for column 2 in units of g/Qhr.

RESPONSE:
This table and its contents have been removed from the revised report. However, in general, there is good reason for using natural logarithms to express variances of emissions in that the logarithmic variance is reasonably stable across a wide range of emissions. This result implies that in "linear" space (e.g., g/Qhr), the variance of emissions is proportional to the emissions level, and therefore cannot be simply stated without reference to a specific emissions level.

11] P. A-1, item i): Residual rather than N?

RESPONSE:
No. In this item N refers to the number of observations, not the residuals.

Knoll: [1] [Section 3] It would be useful to provide some further explanation regarding HE-3555 evaporative emissions behavior. Why did these emissions continue to increase with time? Was the evaporative purge system on the vehicle evaluated for proper functionality? Was any testing done to identify root cause?

RESPONSE:
The behavior of this vehicle is discussed in Appendix D to the revised report (Appendix A to the draft Report). The trend in the measurements for this vehicle was interpreted as a leak that grew over time (0.0, 4.7, 9.9 and 55.5 g/Qhr). However, a specific physical cause for the observed behavior was not diagnosed.

[2] The first bullet point under Summary of LSHED and PSHED states that vehicles with low hot-soak values have PSHED and LSHED results that "are very similar". I think this statement is misleading and may not be correct. The similar scatter shown by the data across three orders of magnitude on a log-log plot suggests that variation at low values was indeed less than at high values. But it is not clear that the data could be considered nearly the same. This assertion requires further justification from the data analysis.

RESPONSE:

The comment is well taken. The log-log analysis suggests that, as typical for emissions, the logarithmic variance is similar across the observed range of 3 orders of magnitude. This result suggests that relative variability is similar across the range, but as the reviewer notes, absolute variability increases with increasing emissions levels.

[3] The last paragraph in . . . section [3] providing relevance to the on-road fleet requires clarification, further explanation and a review of the underlying assumptions. I believe the author is saying that because there is high scatter and a small number of samples available, the upper bound on extrapolating this data to the on-road fleet is necessarily high; higher than it would be if there were either a larger number of sample or a smaller variation in the data. If this is his message, it needs to be stated more clearly and with a more definitive confidence level. Also, is a normal distribution being assumed? If so, state it and explain why such an assumption is valid. If not, then what distribution is assumed and why?

RESPONSE:
The discussion in the revised report estimates sampling error for the estimated frequencies of "elevated" evaporative emissions, which accounts for the sizes of samples of measured vehicles. An analysis is also presented that assesses the sensitivity of the estimated frequencies to the range of expected degree of measurement variability. As typical for emissions measurements, we assumed normal distributions for logarithms of emissions, but not for emissions measurements in their native units (i.e., g/Qhr).

[4] Paragraph 2 of Section 4: The last sentence of this paragraph suggests that two influence factors complicate extrapolation of the Ken Caryl dataset to the Denver-wide fleet. What exactly those two reasons are, however, is not clear from the paragraph text. My interpretation is summarized in the following bullets. Text of the paragraph should more clearly support the thesis statement given at the end of the paragraph.

- The sample of vehicles that visit I/M stations likely has higher emissions than the fleet at-large. The Denver-wide "clean screening" program exempts about 40% of registered vehicles based on low RSD readings. Consequently, the 60% of vehicles that go to I/M stations are the higher emitting fraction of the total Denver fleet. Using this sample population for emissions projection to the Denver-side fleet will likely skew the overall population estimate. However, there is no reason to believe that high tailpipe emissions vehicles are necessarily correlated with high evaporative emissions vehicles. So the real effect of this bias is not clear.
- The Ken Caryl I/M station is located in a higher income part of Denver. Consequently, the population of vehicles visiting this I/M station is likely to comprise newer and therefore cleaner vehicles than the Denver fleet as a whole. As far as I can tell, this bias has no mitigating factors.

RESPONSE:
The discussion in the revised report addresses both of these factors. We do not conclude that either of these factors necessarily cause bias, although both require consideration. With respect to "clean-screen," it is not clear that road-side remote sensing screens vehicles preferentially with respect to their evaporative emissions, if for no other reasons than that evaporative and exhaust hydrocarbons need not be correlated, and that

vehicles pass the remote-sensing units at speeds too high for the instruments to reliably detect evaporative emissions as distinct from exhaust emissions. With respect to the affluence of the population visiting Ken Caryl station, we suggest that emphasizing the importance of model-year group in the analyses should largely neutralize potential bias stemming from differences in the socio-economic status of vehicle owners. While the differences in fleet structure between areas of differing affluence has not been clearly defined, it is plausible that such differences would be apparent primarily in the model-year structures of the fleet.

[5] Accurate application of the Monte Carlo simulation method assumes a random distribution and a large number of samples. This paragraph should include a statement regarding the limitations of this method for analyzing the current dataset. The author does provide later in this report adequate justification that the sample population truly is random. This was well thought-out and well reported. Including some statement in this paragraph, however, would be helpful. I do not believe the author addressed the limitation of population size. This limitation should be mentioned here. Some comment regarding the potential impacts of this limitation should also be stated.

RESPONSE:
We agree that there are limitations in the data set with respect to its utility in a Monte Carlo simulation. The proposal to conduct Monte Carlo simulation did not contribute to the draft report. Consequently, it has been removed from the revised report.

[6] In Section 4.4, Table 4-6: It is not clear how the fourth and fifth columns are calculated from columns 2 and 3. This should be explained.

RESPONSE:
In the revised report, the method used to estimate leak fractions has been changed fundamentally from that presented in the draft report. Thus, the "stratified" calculation presented in Table 4-6 is not relevant in the revised report.

[7] [S]ection [5] of the report

- goes on to discuss additional data that is now available for further investigation. Limitations of the additional data are also identified. For example, the PSHED data from Summer 2010 are identified as not being selected using a stratified random design. As such, these data are not suitable to the Denver-wide fleet.
- leaves the estimation of the high-PSHED fraction of the Denver-wide fleet incomplete. No estimation is provided because the data are identified as inadequate.
- provides no basis for extrapolating the results obtained to an estimate of the nationwide fleet as is needed by EPA. For EPA to apply this dataset to the nationwide fleet (via MOVES), additional justification would be necessary.

RESPONSE:
Unfortunately, the 2010 data were valuable for further index development but not for estimation of representative fleet fractions. With respect to the Denver fleet, the analysis in the revised report is limited to estimating fractions of "elevated emissions" for the fleet sampled. No attempt is made to apply the results more broadly, as such

extrapolation is beyond the scope of the project. We agree that additional analysis is necessary, particularly for the emission rates, to interpret the results more broadly in applications such as the MOVES model.

Responses to Section 3.2, "General Comments"

Each of the reviewers provided general comments on the *High Evaporative Emissions Field Study*. Among these general comments were evaluations of the report's strengths, suggestions for improving and strengthening certain of its elements, and queries for further information.

Frey: [1] What is the main contribution of this report? What are the key limitations? What additional work is needed? If the purpose is to estimate the fraction of vehicles with evaporative emissions exceeding a threshold, the method described in this report using EI23 Bins and a 'stratification' approach may be reasonable; however, the uncertainty in the estimates made using this method are unknown. Such uncertainties should be estimated as the next step. Without quantification of uncertainty, the utility of this approach is unclear.

RESPONSE:
With respect to methodology, the main contribution of this project is to demonstrate successful use of a screening index to improve sampling efficiency, reduce measurement burden, and to estimate a specified fleet characteristic more quickly and at lower cost than would otherwise be possible. With respect to content, the contribution of the report is to present the results of a fleet survey following a rigorous sampling design which allowed estimation of the prevalence of a relatively rare fleet characteristic, e.g., "elevated" hot-soak emissions. In the revised report, we also provide estimates of the effects of sampling error and measurement variability (repeatability) on the estimation of frequencies of "elevated" evaporative emissions.

[2] Some key issues that should be addressed in the conclusions:

- Is PSHED a useful surrogate for SHED?

RESPONSE: see response to this comment above under Specific Technical Comments (Frey).

- Can RSD measurements, if appropriately interpreted, provide an indicator of evaporative emissions?
- Is EI23 a useful indicator?

RESPONSE: see response to this comment above under Specific Technical Comments (Frey).

Are the trends in the results for high evaporative emissions fractions in the vehicle fleet consistent with model year? What results developed here provide some confidence that EI23 is operationally useful?

RESPONSE:

Trends in results by model-year group are very consistent, in that the estimated fractions of "elevated" hot-soak emissions decline markedly with model year group, as expected. Additionally, the efficiency of the screening index also declines with model-year group, also as expected, as the positive predictive value of an index is related to the actual prevalence of the screened trait in different sub-populations. Nonetheless, results indicate that use of an index such as EI23, is operationally useful in reducing the level of effort and expense needed to conduct measurements supporting estimation of fleet characteristics.

- What are limitations of EI23? What other indicators should be explored?

RESPONSE:
As mentioned, the efficiency of EI23 declines for more recently manufactured vehicles. The EI23 index itself is the result of a lengthy development process in which a number of possible indices were proposed and tested. However, the results suggested that relatively simple indices such as EI23 perform as well or better than more complex alternatives involving methods such as principal components analysis applied to sets of absorbance values for HC, CO, CO_2 and NO.

Other indicators have potential value, such as results of OBD system scans performed during maintenance inspections. A separate report on this topic, "Evaluation of the Effectiveness of On-Board Diagnostic (OBD) Systems in Identifying Fuel Vapor Losses from Light-Duty Vehicles", has also been drafted and peer reviewed.

- What uncertainties have been quantified? What uncertainties have not yet been quantified?

RESPONSE:
Uncertainties have been quantified for the utility and effectiveness of the screening index, based on parameters used in epidemiology for this purpose, including sensitivity, specificity, negative predictive value and positive predictive value. The revised analysis also estimates sampling error for the estimated frequencies of "elevated" emissions, using Clopper-Pearson "exact" binomial confidence intervals. In addition, an uncertainty analysis was performed to assess the sensitivity of estimated fractions of vehicles with "elevated" evaporative emissions to systematic over-or under-estimation of vehicles' "true" hot-soak emissions levels. Replication of these results in other fleets would be helpful in further characterizing the utility of the screening index for vehicles manufactured since 1996, and especially for vehicles manufactured since 2004.

- Need for further evaluation of uncertainties prior to making a decision on acceptance of this approach?

RESPONSE:
The approach presented in the draft report has been considered as unacceptable for application. Further assessment has shown that remote-sensing measurements alone lack precision sufficient to reliably characterize evaporative emissions below levels of 10 g/Qhr. However, with respect to use of measures such as EI23 as sampling indices, replication of these results in other fleets would be helpful in further characterizing the

utility of the screening index for vehicles manufactured since 1996, and especially for vehicles manufactured since 2004.

- Application of this or other approaches to fleets that are more representative of the U.S. fleet.

RESPONSE:
The scope of this report is limited to estimating fractions of vehicles with "elevated" evaporative emissions for the fleet sampled. This approach may be applicable elsewhere, but that issue was not addressed in this work. Approaches to be applied to using these data to estimate evaporative emissions in broader contexts are discussed in other sources such as technical documentation for the MOVES model.

Fujita: [1] The experimental approach and methods are adequately documented in the report and accompanying background document. Presentation of the results, including tables and figures, are generally clear except as noted .

RESPONSE:
We appreciate the reviewer's comment.

[2] P. 4-24, 1st paragraph, last sentence: Are the quantifications of uncertainties and bias part of a follow-up report? When is this expected?

RESPONSE:
In the revised report, estimates of uncertainty attributable to both sampling error and measurement repeatability are presented.

Knoll: [1] The analysis relating RSD measurements to SHED results appears valid and well thought out. Uncertainties were investigated and sensitivity analyses were conducted. Use of RSD appears to provide considerable promise for determining high evaporative emissions vehicles from the in-use fleet.

RESPONSE:
We agree in principle, but our current experience suggests that remote-sensing measurements are more suited for identifying candidate vehicles for further evaluation than as a direct predictor of elevated evaporative emissions.

[2] The limited set of vehicles (175 total) that received both RSD and PSHED measurements was used to develop a correlation between RSD readings and measured evaporative emissions. This correlation was applied to the larger set of vehicles (5830 total) that visited the Ken Caryl I/M station during the summer of 2009. In this way, an estimate was made of the percent of vehicles visiting Ken Caryl over the study period that had high evaporative emissions. This projection was well justified based on results presented in the report. Speculation was also made regarding projecting these results to the Denver-wide fleet. Limitations associated with such a broad projection were given. Specifically it was noted that the existing dataset from the Ken Caryl I/M station was limited in relevance to the Denver-wide fleet for two reasons: 1) Colorado exempts about 40% of all registered vehicles from I/M

inspection based on RSD measurements and 2) the Ken Caryl I/M stations is located in an affluent section of the Denver metro area. The first caveat means that the study sample (5830 vehicles) is likely to contain a disproportionate percentage of vehicles with high emissions – either evaporative or tailpipe. As such, the study sample is likely to be biased towards those vehicles with high evaporative emissions and is therefore ***not*** a random representation of the Denver fleet. The second caveat means that the study sample is likely to be composed of newer, properly functioning vehicles. Again, this introduces a bias in the database preventing it from being a random representation of the Denver fleet. Speculation was also made regarding projecting these limited results to the nationwide fleet. Limitations associated with this larger projection were not discussed.

RESPONSE:
See response to a similar comment by Mr. Knoll above under Specific Technical Comments.

Responses to Section 3.3, "Editorial Comments"

Each of the reviewers to varying degrees assessed the narrative of the report and suggested improvements for accuracy, clarity, and consistency. One of the reviewers undertook a thorough critique of the report in this regard, providing significant editorial suggestions and stressing the need for a thorough re-organization, rewrite, and technical editing . To this end, all of the reviewers highlighted typographical and formatting errors, incorrect word choice, and omissions, including missing references.

RESPONSE:
In response to editorial comments, the report has been fundamentally reorganized and redrafted. Detailed attention was given to upgrading the quality of the presentation as well as the scope and depth of needed explanations. Terms and acronyms used in the report were either removed and replaced with clearer terms or defined at first use in the text. In addition, a glossary was added. Informal terminology and metaphors were replaced with more formal terms.

However, we have maintained use of active voice in the document. The consensus that active voice was to be avoided in scientific and technical writing has shifted in recent years in favor of active voice, to allow the clear and direct expression that it provides. At this point, even first-line peer-reviewed publications such as *Nature* and *Science* encourage authors to use active voice[1,2]. In addition, Federal Agencies are under direction to conform to the concepts of "Plain Writing," which encourages the use of active over passive voice where appropriate, even in "technical support" documents[3].

---

[1] For the journal *Nature*: *"Nature journals prefer authors to write in the active voice ("we performed the experiment...") as experience has shown that readers find concepts and results to be conveyed more clearly if written directly."* See http://www.nature.com/authors/author_resources/how_write.html.

[2] For the journal *Science*: *"Use active voice when suitable, particularly when necessary for correct syntax (e.g., "To address this possibility, we constructed a λZap library . . .," not "To address this possibility, a λZap library was constructed . . .").* See http://www.sciencemag.org/site/feature/contribinfo/prep/res/style.xhtml.

[3] See: http://www.epa.gov/plainlanguage/faqs.html#h.

The following are comments which were found in this section which warranted a more substantive response:

Fujita:

[5] The current Chapter 3 should be rewritten as "Assessment of Concordance Between Portable and Fixed Location Evaporative Emissions Measurements." This chapter needs technical editing. The basic information is useful and interesting. The technical analysis should include quantification of the statistical significance of each parameter in the regression equation, the standard error of the estimate, the distribution of the residuals, a normality check for the residuals, the coefficient of determination, and other basic information that would commonly be reported as diagnostic goodness-of-fit indicators when developing a regression model. To what extent are results such as in Figures 3-4 and 3-5 actually providing an indication of repeatability of the test – are the conditions really the same in each test? If the repeatability is really this poor, what are the implications for selecting a threshold for what constitutes a 'high evap' vehicle? It is more common to report 95% probability ranges, not 68% probability ranges.

RESPONSE:
The discussion referred to in the comment has been moved from the body of the draft report to Appendix D in the revised report.
The log-log regression of PSHED on LSHED values has been removed. It was judged to not contribute meaningfully to the revised report, which focuses on PSHED results, independently of LSHED results. However, the revised report provides a more rigorous reanalysis of the paired PSHED results, as a basis for reassessing the effects of repeatability on the estimation of the prevalence of "elevated" evaporative emissions. The basis of the analysis is a simple regression of 2nd replicate on 1st replicate PSHED results, with goodness of fit parameters reported and applied to estimate a prediction interval for subsequent replicates based on initial replicates. Despite efforts to maintain uniformity of test conditions during the sequences of repeat tests, it is not possible to be certain that all vehicle and test conditions were in fact uniform throughout. Nonetheless, we interpreted the paired tests as providing the best estimates of repeatability available for the PSHED procedure on a practical basis, incorporating unknown as well as known sources of variability. To assess the implications of this uncertainty in comparing measured results to a set of discrete thresholds, we performed a sensitivity analysis to assess the expected differences estimated prevalence of "elevated" evaporative emissions under assumptions that measured values both systematically underestimated and overestimated the "actual" emissions status of sampled vehicles.

[4] P. A-2: Add a description of the origin of the constants used in equations shown at the bottom of the page. Explain how this reduces dependence of EI23 on exhaust HC concentrations.

RESPONSE:
These equations are based on results of experimental work in which the calculation of EI23 values was related to the levels of simulated exhaust and evaporative emissions. The constants in the equations represent intercept and slope terms. The constants taking values of 0.0 and 15.0 represent minimum and maximum propane release rates, as scfh,

which represented evaporative emissions in the experiment. The HC concentration value is included to account for the presence of exhaust emissions in the plume and their effect on the values of EI23 obtained. However, having reconsidered the issue, we no longer attempt to make a case that the use of a "method B" value reduces the "dependence on exhaust emissions" in the plume. The values we used were calculated by method B, although this selection is not a necessary one. Values calculated by method A could also have been used; the inclusion of the term is not based on assumptions concerning differences between methods A and B in distinguishing exhaust and evaporative emissions. The double application of the natural log transformation was applied as a variance stabilizing measure in the regression. Material has been added to Appendix A to the revised report to clarify the development and application of these equations.

Knoll: [1] Elsewhere in the literature, estimates are made providing comparison of PSHED results with EPA's Tier 2 requirements for evaporative emissions.[4] It would be helpful to include that here for context.

RESPONSE:
In this reference the 0.3 g/15 min level for the PSHED mistakenly refers to Tier 2 standards, in reality this 2 gram standard is for Enhanced Evaporative Emissions Vehicles Standard. Assuming the hot soak portion accounts for 20%, which is 0.4 g, and then 75% of those emissions occur in the first 15 minutes, arrive at a 0.3 g/15 min "standard". This is discussed in the report in Section 4.3.3 "Interpretation".

[4] 1 "Evaluation of Evaporative Leaks using RSD and Inventory Implications," D. Hawkins, C. Hart, C. Fulper, J. Warila, D. Brzezinski, et al., Presented at the 19th Annual International Emission Inventory Conference, San Antonio, TX, Sept 27-30, 2010.