January 31, 2022

California Air Resources Board
1001 I Street
Sacramento, CA 95814

RE: Low Emission Diesel (LED) Study: Biodiesel and Renewable Diesel Emissions in Legacy and New Technology Diesel Engines

Renewable Energy Group, Inc. (REG) appreciates the opportunity to provide technical feedback to the Low Emission Diesel (LED) Study: Biodiesel and Renewable Diesel Emissions in Legacy and New Technology Diesel Engines (NTDE).

REG is leading the energy industry transition to sustainability by transforming renewable resources into high-quality, cleaner fuels. REG is an international producer of cleaner fuels, North America's largest producer of biodiesel, and a leading producer of renewable diesel. REG's solutions are alternatives for petroleum diesel and produce significantly lower nitrogen oxide (NOx), particulate matter (PM) and carbon emissions. REG utilizes an integrated procurement, distribution and logistics network to operate 11 biorefineries in the U.S. and Europe. In 2020, REG produced 519 million gallons of cleaner fuel, delivering 4.2 million metric tons of carbon reduction. REG is meeting the growing global demand for lower-carbon fuels and leading the way to a more sustainable future.

It is unfortunate and problematic that the research presented in the *Low Emission Diesel (LED) Study: Biodiesel and Renewable Diesel Emissions in Legacy and New Technology Diesel Engines* was not peer reviewed or, for that matter, shared with outside experts prior to its publication. We believe an appropriate review process would have informed the authors that:

- The conclusions presented in the work are not supported by the data that were generated
- The authors claimed highly significant differences between fuels for the most consistent test cycle when all values obtained for that cycle were extremely close to (if not below) the limit of quantification for the NOx measurement system
- There were significant repeatability problems with at least two of the test cycles, with one data set manipulated in an attempt to reduce the apparent variability of the test results

- The complete data set is not robust enough to draw any conclusions about the impact of fuel type on NOx emissions from NTDEs within the resolution of the measurement system

Because of these significant and numerous flaws, the potentially damaging impact in California of any claims about NOx emissions, and the fact that the research was commissioned and supported by CARB, which has far reaching authority to determine which fuels may be used in the state, we feel the only responsible action at this time is publicly retract the report.

NTDE Background

As outlined in CARB's report, New Technology Diesel Engines (NTDEs) are diesel engines that meet at least one of the following criteria:

1) 2010 CARB emissions standards for on-road heavy duty diesel engines (0.2 g/BHP-hr NOx and 0.01 g/BHP-hr PM)
2) Tier 4 emissions standards for off-road engines
3) Employs a diesel emissions control strategy which uses selective catalytic reduction (SCR) to control NOx

CARB has been advancing the penetration of these NTDEs in the California marketplace for a number of years. A strong and growing body of data, including this study from CARB, consistently shows the profound positive impact NTDEs have on both criteria pollutant emissions and fuel efficiency. In particular, NTDEs have repeatedly been shown to be effective at reducing NOx and particulate matter (PM) emissions by more than 90% for all legal diesel fuels, including those that meet California-specific specifications.

The body of work around legacy engines (i.e., non-NTDEs) has repeatedly demonstrated that fuel properties, such as cetane number, have a direct impact on criteria pollutant emissions. This is because legacy engines are relatively simple combustion devices that neither monitor nor attempt to control the criteria pollutant emissions they generate. NTDEs, on the other hand, are highly sophisticated systems comprised of a number of key elements that *both* monitor and attempt to control certain criteria pollutant emissions. These differences are essential points to consider when looking at NTDE tailpipe emissions. Because of NTDEs' mode of operation, the tailpipe NOx emissions from these new engines simply do not reflect consistent impacts based solely on the fuel, in spite of past studies with legacy engines which have occasionally shown repeatable trends for fuel composition and NOx emissions. As this study demonstrated, attempts to apply experimental approaches for determining fuel property's impacts on tailpipe NOx emissions, which were applied to legacy engines in the past, to an "out of the box" NTDE system is scientifically inappropriate .

Furthermore, the NTDE aftertreatment devices do not all operate the same way. To illustrate this point, the following paragraphs will compare, at a high level, the significant differences between the operating mode of diesel particulate filters and selective catalytic reducers, and how these differences must be considered when one interprets NTDE PM and NOx emission results.

Diesel Particulate Filtration (DPF)

The diesel particulate filter (DPF) is, as the name implies, a filtration device that captures PM by forcing the exhaust gas to flow through a filter element or barrier (Reşitoğlu, Altinişik and Keskin, 2014). Diesel particulate matter, or soot, is primarily carbonaceous material resulting from the incomplete combustion of diesel fuel. In general, the diesel particulate matter reduction efficiency is dependent only upon the total mass loading and particle size of PM in the exhaust stream and the pore size of the DPF element. It is a static device that does the same thing the same way repeatedly, no matter what else is happening with the vehicle.

Selective Catalytic Reduction (SCR)

Unlike the DPF, the selective catalytic reduction (SCR) unit is an active device governed by a complex control scheme. Generally speaking, the SCR converts approximately 90% of the NOx coming out of the engine to N2 and H2O over a catalyst bed through a reaction pathway involving ammonia (introduced to the system in the form of a urea solution called diesel exhaust fluid or DEF) (Reşitoğlu, Altinişik and Keskin, 2014). The residual NOx content of the tailpipe exhaust is influenced by a considerable number of factors (such as DEF dosing rate, exhaust gas recycle (EGR) setting, exhaust oxygen and water content, and exhaust and SCR temperatures), most of which are actively controlled by the engine control unit (ECU).

Engine Control Unit (ECU)

It is critically important to understand that the ECU is *adaptive* (i.e., the setpoints for the multitudinous parameters it controls are continually adjusted over time in response to changes in parameters like driving patterns and engine exhaust composition) and that exhaust composition varies with driving patterns, environmental conditions, exhaust gas recycle (EGR) rate, and fuel properties.

It is even more important to understand that the ECU is not programmed to minimize NOx. Rather, the ECU continuously adjusts numerous engine and exhaust system parameters to balance multiple engine performance objectives including, but not limited to: maximizing fuel economy, minimizing DPF regeneration frequency, and maintaining average NOx emissions below the applicable EPA certification limit for the engine in question. **Achieving the *lowest possible* NOx is never, at any time, a control system goal.** This means there is no technical reason to compare—nor any value in trying to compare—the impact of different fuels on a properly functioning NOx reduction system. Simply put, the ECU algorithms preclude tailpipe NOx from being an appropriate experimental output (i.e., dependent variable), at least not with any legal fuel. To help clarify this statement with two counter examples, either a NOx reduction system with a control objective of always achieving the lowest possible NOx or a system that fixed all known operational parameters that could impact NOx emissions (other than fuel) would potentially allow the kind of comparison CARB and CE-CERT have attempted. Unfortunately, neither of those is the kind of system that was used in the reported tests.

Because the ECU program's NOx reduction objective is satisfied as soon as NOx is below the EPA threshold, the ECU in effect shifts its attention from NOx reduction to fuel economy improvement or DPF protection or some other priority—i.e., it stops attempting to reduce NOx any further. Therefore, the only relevant question that could be asked regarding the impact of fuel properties on NOx emissions in the kind of NTDE exhaust systems tested in this work is - *if the ECU is allowed to fully adapt to each fuel, are the average NOx emissions under the engine's EPA certification limit*? If the answer is "no," that is evidence that at least one of the following three fundamental problems has occurred:

1. The fuel's nitrogen content is so high that the SCR's conversion capacity and/or DEF dosing system simply cannot keep up with the amount of NOx leaving the engine.

2. The ECU hasn't had a chance to fully adapt to the fuel in question (e.g., when dramatic step changes in fuel composition are made).

3. The SCR system is not functioning properly (e.g., component failure, improper engine/ECU configuration, or other inadequate human input)

Fundamental problem 1 is clearly not the case for the present study, as the nitrogen levels for the CARB ULSD, R100, R65/B35, and R50/B50 were all reported to be less than 10 ppm, which is a low level for diesel fuel. Therefore, the nitrogen content in the fuels cannot account for the differences in the NOx emissions results that were generated. Any failures to maintain NOx under an engine's EPA certification limit are caused by fundamental problems 2 and/or 3.

Before we move on to evaluate fundamental problems 2 and 3, it is important to again emphasize that the ECU/SCR control scheme never attempts to minimize NOx emissions to as close to zero as technologically feasible. The ECU/SCR control scheme is designed to keep NOx under the maximum threshold required by US EPA regulation. A NOx minimization objective for the control scheme would actually be irrational since the NOx reduction / fuel efficiency trade-off is a well-known dilemma for diesel engine control schemes. In other words, if there were an ECU control scheme directive to reduce NOx emissions any further than the requirement threshold, this would be expensive in the broadest sense of the word for both the vehicle operator (from increased consumption of both DEF and fuel) and the planet. Since both DEF production and fuel combustion produce greenhouse gases, unnecessary use of either carries a climate impact cost.

Any responsibly-developed diesel emissions inventory management program should assume that all NTDEs produce NOx on average at or just under their respective EPA certification limits and should not try to assume and/or take credit for small hypothetical reductions below the system's threshold design limit—nor should a responsible program penalize small, hypothetical, hard to measure (i.e, at the limit of quantification), and statistically meaningless increases with emissions still under the EPA threshold. For on-road

NTDEs, this threshold limit is currently 0.2 g/BHP-hr, whereas for off-road NTDEs it is 0.3 g/BHP-hr.

Challenges with Testing NTDEs

As discussed in the preceding paragraphs, NTDEs are highly sophisticated devices; the amount of sensors and control loops governing nearly every aspect of NTDE operation is staggering. Therefore, it is incredibly difficult to test an NTDE in a lab in a manner that truly reflects real-world performance. In addition, correctly and properly testing an NTDE on an engine dynamometer (i.e., with the engine separated from the vehicle for which it was designed) without the direct and continual involvement of the original equipment manufacturer (OEM) is operationally difficult and nearly impossible. As just one example of the challenges, the ECU has to essentially be "tricked" into thinking that the engine is still, in fact, mounted inside a vehicle before it will even begin to function. Perhaps most importantly, each ECU must be loaded with a "map" (essentially a software program) that governs the operation of the engine. This map is proprietary to the OEM and information about it is rarely shared outside of the OEM.

Based on our own knowledge in developing data for the ADF, we can provide other examples to show the significant challenges that a test lab must overcome when testing an NTDE. The complexity of the system in general, coupled with the fact that it was designed and programmed to work as intended in an actual vehicle, makes it infeasible to attempt to evaluate emissions control system performance without the direct and close involvement of the OEM during the design and execution of the testing. Trying to install and operate an OEM's NTDE without their full documented guidance in the setup creates more questions about the test results than could ever be answered by researchers attempting to do it themselves. And even with an OEM's full-time involvement, it is nearly impossible to confidently assume that the results obtained are representative of real-world performance. Although the report's authors acknowledge some involvement from the two NTDE OEM's whose engines were used, we believe it is telling that no OEM representatives are included as co-authors.

Impact of adaptive NOx reduction operation on emissions testing

CARB has published results from their testing of a dynamic, adaptive control system/device combination with frequent, abrupt, and dramatic fuel composition changes. CARB has suggested that the results from this testing can be utilized to draw conclusions regarding the composition and quality of NTDE tailpipe emissions for the fuels tested. We believe **this is an incorrect inference and false conclusion**.

While we agree that it is critically important that both on-road and off-road SCR systems perform to their design specification with all legal fuels, it is not correct to draw direct conclusions about the impact of a fuel on either NTDE's tailpipe emissions due to the dynamic, adaptive nature of NTDEs and their highly complex, multi-objective control schemes. As noted earlier, a multitude of factors that influence NOx emissions are constantly changing and adapting, according to the adaptive ECU control scheme, with the sole directive to maintain average NOx below a threshold, not to minimize NOx nor even to maintain a target NOx reduction percentage.

REG submits only **two** conclusions can be drawn from this data set. The first is that NOx emissions is not an appropriate dependent variable for comparing NTDE performance based on fuel properties alone, while adding the additional complexities involved in dynamometer testing and translating results into meaningful "real world" conclusions is even more challenging. For example, the on-road NTDE exceeded its NOx emissions EPA certification limit with all four test fuels in one test cycle while remaining comfortably below the certification limit with the same four fuels in another test cycle. This same conflicting data phenomenon occurred with the off-road NTDE with even more dramatic disparities between the reported NOx emission levels for the two test cycles. Any future NTDE study in which the OEM is not integrally involved throughout the test program—particularly in the engine and ECU set-up—is likely to see similarly erratic NOx emissions results.

The second conclusion is that both the on-road and off-road NTDEs were demonstrated to be capable of dramatically reducing engine-out NOx and to conform with their respective EPA certification requirements, as designed, for all the fuels tested. The fact that the researchers managed also to produce erratic and unreliable results from one test cycle for each engine reinforces the need for consistent, well-documented OEM involvement in any future NTDE test programs, but it does <u>not</u> negate the conclusion that NTDEs successfully and dramatically reduce engine-out NOx.

Evaluation of the Statistical Methods Applied

*NOTE: the following critique of the data analysis itself is not meant in **any way** to be an acknowledgment of validity for the test results or an endorsement of their appropriateness for parametric statistical analysis. This is merely intended to provide general guidance to explain further why the report's conclusions would lack validity* even *if the NOx data were appropriate for comparison. The general guidance could be helpful in improving the quality of future work since the current work violated numerous statistical norms and common standards for scientific data analysis.*

One of the first lessons taught in statistics is the importance of checking the assumptions associated with the test being performed. Parametric tests, the class of tests to which the Student's t-test used in this study belongs, have four inherent requirements for the data being analyzed:

1) The data points are independent. This means the comparison groups of data are independent from one another, as well as the individual data points within a comparison group are also independent from one another.

2) The data in each comparison group are normally distributed.

3) The data in each comparison group are approximately homoscedastic, or have similar amounts of variance.

4) The data set does not contain outliers.

If any of these assumptions are violated by the data being analyzed, then it is <u>not</u> possible to draw valid conclusions from the use of parametric tests. **The data in this study violate <u>all</u> <u>four</u> of these assumptions, the most problematic of which is the false assumption that the individual data points are independent.** Violating the requirement of independent data is a fundamental problem with the experimental design, and *cannot* be resolved after the fact through statistical methods or data exclusion. **In spite of this, the authors chose to exclude data points in the on-road NTDE data set in order to reduce the variability of the results**

based on a subjective *a priori* assumption about what values should be allowed rather than on an objective, evidence-based assessment of the validity of the results.

The data exclusion approach is doubly problematic: first, it reduces the already small comparison group sample sizes and thereby decreases the likelihood that the affected data sets will constitute normal distributions; and second, the researchers chose to apply a subjective rejection standard to only *one* of the data sets. This kind of data manipulation is highly questionable and unusual – and would raise serious concerns if the manipulated results are used to draw any impactful conclusions.

We believe it is very damaging for CARB to allow publication of such a cursory, speculative, and inappropriately analyzed report. It is also confusing why CARB did not apply the same scientific rigor to this study that they require for emissions data submissions for commercial stakeholders. As noted previously, we believe the report does not meet basic standards for publication and should be formally withdrawn.

The additional incorrect assumptions of normal distribution, equal variance, and lack of outliers are addressed below. We hope this critique might prove useful in future work if it helps the researchers develop more thorough and rigorous experimental methodology.

In appropriate circumstances, the Central Limit Theorem can be applied to allow an assumption of normal distribution for sample sizes greater than 30 (Ross, 2014). However, the largest sample size in this study is 18 (CARB ULSD for legacy engine cycles), necessitating a statistical evaluation of normality for all sample groups before one can assume a normal distribution, as the researchers in the present study did. Table 1 shows the results of the Shapiro-Wilks test for normality for all sample groups from the study. The Legacy/NRTC engine/cycle combination was the only one for which there is 95% confidence all four of the fuel subsets are normally distributed. This means the Legacy/NRTC engine/cycle combination is the only set of data that passes the normality assumption of the Student's t-test parametric test.

Table 1: Results from evaluation of normality of data samples using the Shapiro-Wilks test.

Engine-Cycle Combination	**Fuel**	**Sample Count**	**Shapiro-Wilks p-value**	**95% Confidence of Normality?**
Legacy/D2	CARB	18	0.07	Yes

Legacy/D2	R100	9	0.01	No
Legacy/D2	B35	9	0.92	Yes
Legacy/D2	B50	9	0.57	Yes
Legacy/NRTC	CARB	18	0.96	Yes
Legacy/NRTC	R100	9	0.45	Yes
Legacy/NRTC	B35	9	0.07	Yes
Legacy/NRTC	B50	9	0.53	Yes
On-Road NTDE/FTP	CARB	15	0.00	No
On-Road NTDE/FTP	R100	10	0.00	No
On-Road NTDE/FTP	B35	9	0.01	No
On-Road NTDE/FTP	B50	11	0.00	No
On-Road NTDE/RMC	CARB	15	0.00	No
On-Road NTDE/RMC	R100	9	0.00	No
On-Road NTDE/RMC	B35	8	0.00	No
On-Road NTDE/RMC	B50	9	0.00	No
Off-Road NTDE/C1	CARB	11	0.00	No
Off-Road NTDE/C1	R100	9	0.00	No
Off-Road NTDE/C1	B35	9	0.00	No
Off-Road NTDE/C1	B50	6	0.11	Yes
Off-Road NTDE/NRTC	CARB	11	0.09	Yes
Off-Road NTDE/NRTC	R100	12	0.00	No
Off-Road NTDE/NRTC	B35	12	0.28	Yes
Off-Road NTDE/NRTC	B50	12	0.01	No

Analogous to how the Shapiro-Wilks test can be utilized to evaluate normality, the non-parametric Levene's test can be used to evaluate equivalence of variance in non-normally distributed data. Table 2 shows the results of applying Levene's test to the six engine/cycle combinations. The variances within the fuel subsets of four of the six engine/cycle combinations were found to be statistically equivalent. However, the variances of the fuel subsets for the off-road NTDE engine were found to be significantly different from one another, therefore violating the Student's t-test assumption of equivalent variance.

Table 2: Results from evaluation of equivalence of variance between fuels, within each engine/cycle combination. Levene's test was used to perform the evaluation.

Engine-Cycle Combination	**Levene's p-value**	**95% Confidence of Homoscedasticity?**
Legacy/D2	0.34	Yes

Legacy/NRTC	0.86	Yes
On-Road NTDE/FTP	0.95	Yes
On-Road NTDE/RMC	0.44	Yes
Off-Road NTDE/C1	0.00	No
Off-Road NTDE/NRTC	0.00	No

Lastly, we can look at the whiskers of a box and whisker plot as a non-parametric way to evaluate outliers in each data set. Table 3 shows the count of data points in each engine/cycle/fuel combination that fall outside the whiskers for that data set. The whiskers extend 1.5*(interquartile range) beyond the first and third quartiles. Decisions to include or exclude outliers in a statistical analysis depends on many factors, including the context and intended application of the test program results. Regardless of that decision, the presence of statistical outliers violates the fourth requirement for the use of the Student's t-test.

Table 3: Results of initial outlier analysis using quartiles and interquartile range.

Engine-Cycle Combination	**Fuel**	**Sample Count**	**Number of Outliers**
Legacy/D2	CARB	18	1
Legacy/D2	R100	9	0
Legacy/D2	B35	9	0
Legacy/D2	B50	9	0
Legacy/NRTC	CARB	18	0
Legacy/NRTC	R100	9	1
Legacy/NRTC	B35	9	0
Legacy/NRTC	B50	9	0
On-Road NTDE/FTP	CARB	15	3
On-Road NTDE/FTP	R100	10	1
On-Road NTDE/FTP	B35	9	1
On-Road NTDE/FTP	B50	11	0
On-Road NTDE/RMC	CARB	15	0
On-Road NTDE/RMC	R100	9	2
On-Road NTDE/RMC	B35	8	1
On-Road NTDE/RMC	B50	9	0
Off-Road NTDE/C1	CARB	11	0
Off-Road NTDE/C1	R100	9	0
Off-Road NTDE/C1	B35	9	0
Off-Road NTDE/C1	B50	6	0

Off-Road NTDE/NRTC	CARB	11	0
Off-Road NTDE/NRTC	R100	12	1
Off-Road NTDE/NRTC	B35	12	0
Off-Road NTDE/NRTC	B50	12	0

It is our view that, in this context, the existence of outliers and the absence of normal distributions are important evidence of an unreliable experimental system with inadequate sample sizes and, most important, a dependent variable that depends on a myriad of uncontrolled factors, including the intricacies of the multi-objective ECU algorithms as well as apparently inadequate NTDE system setups and operator inexperience. With this much uncertainty in the measurement system, the extreme data points cannot be excluded simply because they are extreme. Until they can be rejected based on an objective explanation, these data points must be acknowledged as evidence that the experimental system was not "in control" under the chosen operating conditions and experimental approach. This means the data generated are not repeatable enough to be subjected to a meaningful statistical analysis at the level of distinction desired by the researchers. To put it another way, no meaningful conclusions can be drawn from "out of control" test results because the potential variability of repeated tests of the NTDE systems is too large to achieve the desired level of distinction between fuels.

The fact that the NTDE NOx emissions results in this study are not truly independent presents additional unaddressed obstacles to any meaningful analysis of the data. In addition, the fact that the NOx emissions results from two of the NTDE test cycles are clearly out of control and not repeatable fully invalidates them from inclusion in any statistical comparison of fuel impacts on NOx emissions from these two engines.

Any analysis of this data and any resulting conclusions will be meaningless due to the lack of sample independence. On top of that, the application of the Student's t-test is inappropriate for the vast majority of the data from the study. The Legacy/NRTC engine/cycle combination is the only subset of data from this study that shows potential for being appropriate for the Student's t-test analysis. More thorough and more careful experimental methods and strategies must be developed and implemented to even attempt to publish a meaningful, useful, and credible report on fuel impacts on NOx emissions from an NTDE.

Experimental Data Are Unreliable Because Experimental Systems Were Not "In Control" for All Test Cycles

The Executive Summary and full text portions of this report could reasonably be interpreted by a well-intentioned reader as implying that the researchers' experimental approach yielded NTDE NOx emissions values that can be anticipated by fuel composition, while the report's statistical summaries state outright that the authors believe significant conclusions can be drawn. As discussed in the section above, the statistical evaluation conducted by the authors was, in effect, meaningless and the conclusions presented were inappropriate. More critically, the report's implication that the study acquired meaningful data for the impact of fuel composition on NTDE NOx emissions is simply incorrect.

To their credit, the researchers included the entire data set in their report, including a test identification code comprised of date and time that allows the chronology of the tests to be evaluated as an influential factor. The inclusion of the full data set allows others to provide potentially appropriate and meaningful evaluations of the test results, for which we are grateful.

> *Note: because CARB is a regulatory agency with the ability to directly control which fuels are allowed in the marketplace, the full data set should always be published when the agency endorses experimental work. The fact that CARB has broad governmental authority places a higher standard for scientific integrity on any experimental work published by the agency than on an independent researcher. Great power must be tempered by even greater discretion.*

Presenting the complete collected data set, as collected, in graphical form, as we do below, clearly demonstrates the unreliability of the experimental design and execution and of the data collected from the NTDEs. We will start with the off-road legacy engine results, which do not suffer from all of the problems described above. They provide a helpful counter example for the NTDE results. Unlike the NTDEs, the legacy engine appears to meet a reasonable standard for repeatability for both cycles tested (see Figure 5 and subsequent discussion).

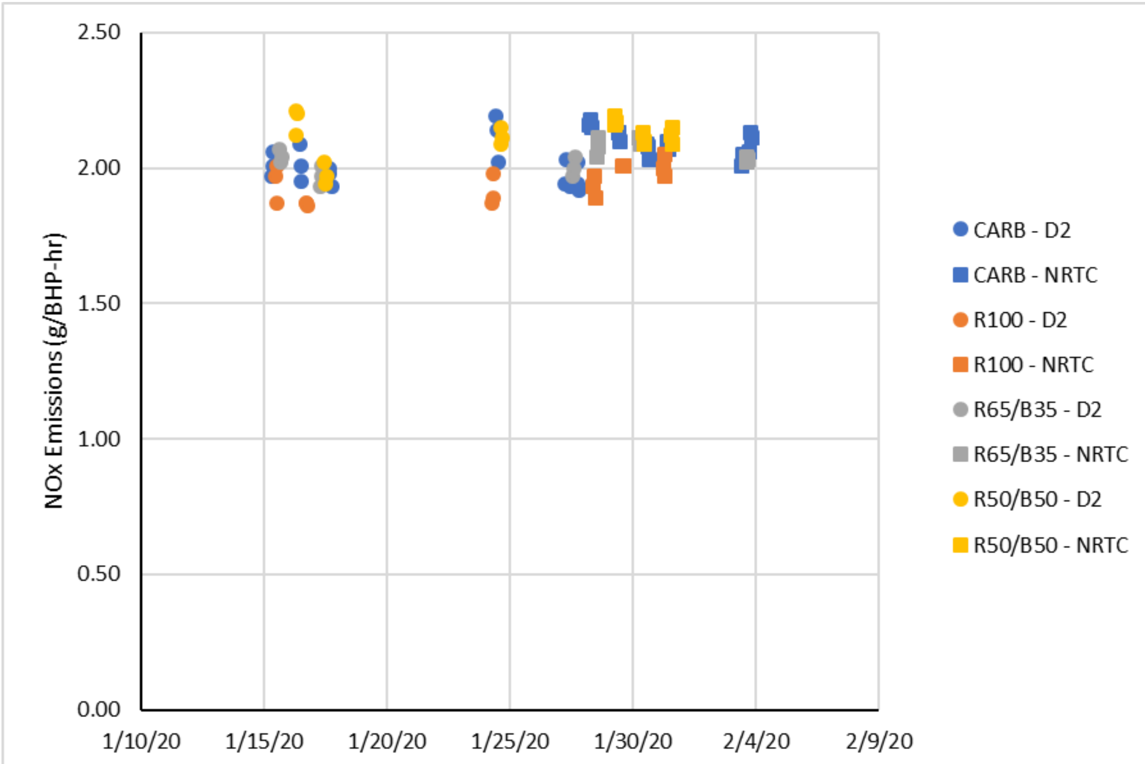

Figure 1: Complete data set in chronological order of off-road legacy engine tailpipe NOx emission results

The data in Figure 1 show expected results with reasonable relative variability, as noted in the report. This was the only engine with a full data set that could be considered a conditional success for evaluating the impact of fuel type on NOx emissions. It is also the only data set generated without an ECU-controlled SCR system, which is not a coincidence. No data for this engine were excluded from the report. The relative repeatability of the NRTC results offers an opportunity to determine a reasonable variability threshold for assessing the reliability of NOx data generated with this engine, if desired.

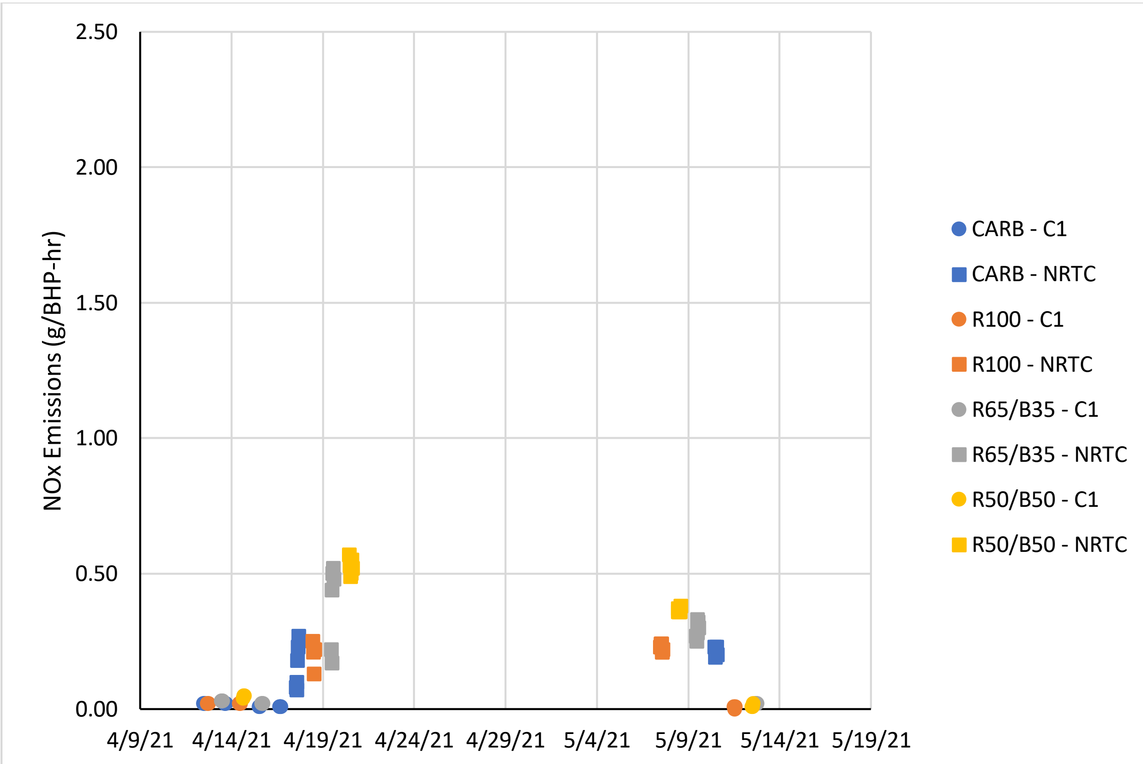

Figure 2: Complete data set in chronological order of off-road NTDE tailpipe NOx emission results

Conversely, the NRTC cycle test results from the off-road NTDE shown in Figure 2 demonstrate extremely high variability and trends indicating an adaptive system that is not in control. From an experimental system troubleshooting perspective, there is also an interesting and potentially meaningful "bimodal" data set for the R65/B35 testing in April. The C1 cycle data, on the other hand, appear to be substantially more reliable than the NRTC data and are also in the expected range for a properly functioning SCR system. The most important observation from Figure 2 is that the out-of-control and unreliable NOx data from the NRTC run were obtained with the exact same fuels and the exact same NTDE configuration as the reasonably repeatable C1 cycle results.

A reasonable conclusion would then be to assume that the in-control, repeatable results from the C1 cycle are more representative of SCR performance than the out-of-control, unrepeatable results from the NRTC cycle. However, the authors have presumed all results for this engine are valid and reliable by treating all data equally in their flawed statistical analysis. Because high variability inevitably produces higher average values in a measurement system with a lower bound of zero and with expected values near the limit of quantification, this presumption of validity effectively overweights the out-of-control results.

Treating out-of-control and in-control data equally would have been acceptable if the stated purpose of the report had been to provide insight into the experimental challenges of trying to evaluate NOx emissions with a dynamometer-mounted NTDE, but that was not the subject of this report. Instead, the authors have incorrectly presumed they are able to provide a report with meaningful NOx emissions results for all combinations of engine, fuel, and test cycle (with one critical exception, which will be discussed below in the "Inappropriate Data Exclusion" section). And, rather than give more credibility in their discussion and conclusions to the in-control, repeatable results the authors chose to emphasize the out-of-control, unrepeatable results.

One substantial benefit of the relative repeatability of the C1 cycle results is that they also offer an opportunity to estimate a reasonable variability threshold for the off-road NTDE NOx emissions. This is discussed below in the "Comparison of Variability" section.

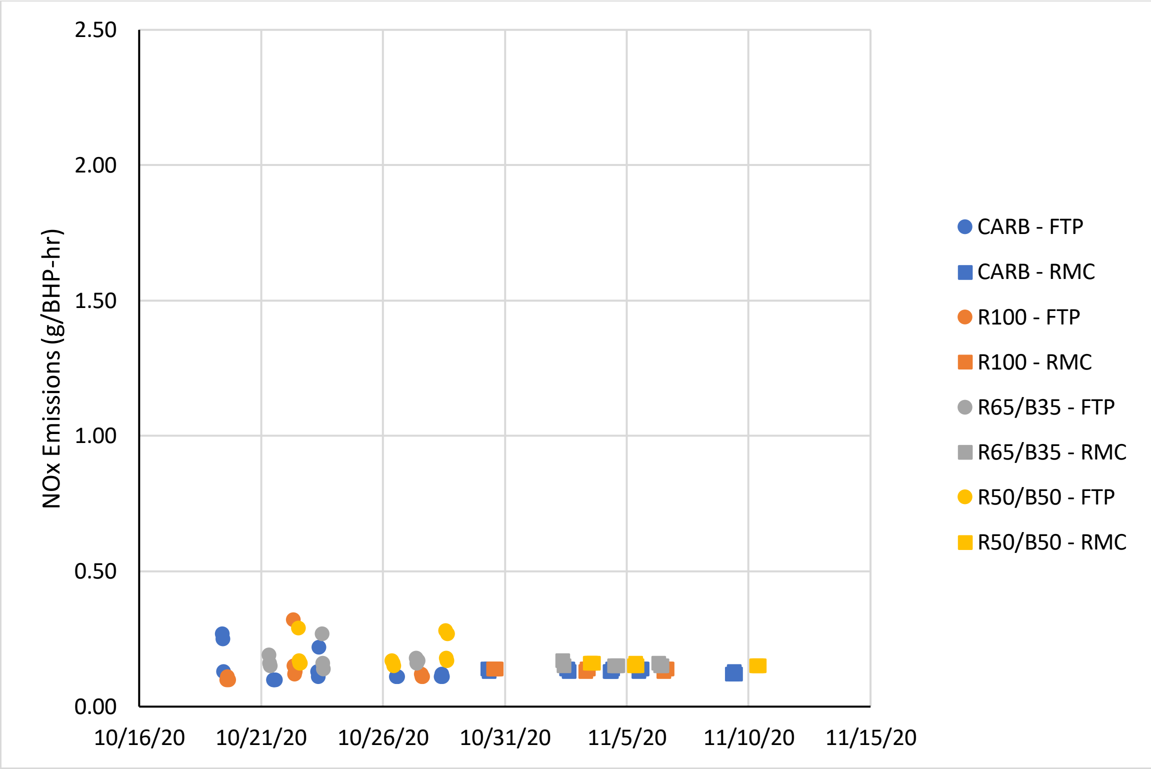

Figure 3: Complete data set in chronological order of on-road NTDE tailpipe NOx emission results

The FTP cycle test results in Figure 3 show extremely high variability and the same bimodal distribution of the test results that was observed for R65/B35 in the off-road NTDE NRTC data (Figure 2), but this time for all four fuels. The high variability with the FTP cycle indicates out-of-control operation and the manifestation of the bimodal results artifact for all fuels is strong evidence of systematic problems with the experimental design and/or the equipment. However, what is truly problematic is that the highest test results for this engine were excluded from the report without appropriate justification. The researchers' decision to exclude only the high test results for this specific engine/test cycle combination represent a major breach of common standards of scientific integrity whose consequences will be discussed further below.

The fact that the RMC results are relatively self-consistent indicates that repeatability is possible and that the engine's SCR system can reliably stay under the current on-road certification limit of 0.2 g/BHP-hr. As with the C1 cycle for the off-road NTDE, the relative repeatability of the RMC results offers an opportunity to determine a reasonable variability target for the on-road NTDE engine.

Comparison of Variability

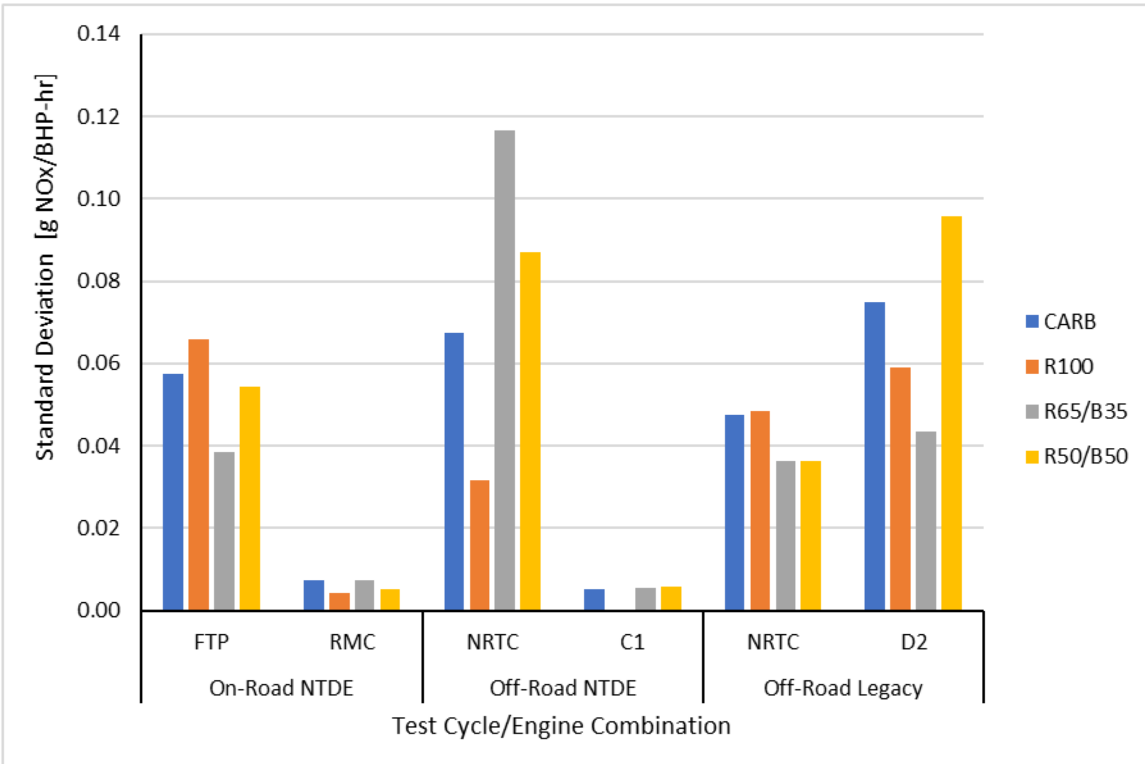

Figure 4: Standard deviation comparison for all fuels and test cycle/engine combinations.

It is remarkable that the absolute variabilities of the results from the FTP cycle for the on-road NTDE and the NRTC cycle for the off-road NTDE were in the same range as the variability of the results for the off-road legacy engine, which has tenfold higher NOx emissions. On the other hand, the absolute variabilities of the RMC and C1 cycle results were, as one would expect from an in-control experimental system, an order of magnitude smaller than those of the FTP and NRTC cycle results

The RMC and C1 cycle results are much more consistent with real-world expectations for SCR control system performance with all fuels tested, which indicates that both the off-road and on-road SCR systems are capable of controlling NOx emissions as designed regardless of fuel type. <u>The low variabilities of the RMC and C1 cycle data sets also confirm that the</u>

experimental systems with the FTP cycle for the on-road NTDE and the NRTC cycle for the off-road NTDE were not capable of producing reliable results with an SCR system that was confirmed to be properly functional.

In the absence of a common standard for ECU/SCR performance variability, the study's own data can be used to provide an estimate of a reasonably-achievable "internal standard" for acceptable variability of repeated tests for each engine. The lowest average variability that was actually achieved for each engine across all fuels provides an upper limit for a reasonable variability expectation for "in-control" operation. While this value is unlikely to be as stringent as the OEMs' own performance standards for their respective NOx control systems, it at least provides an objective indication of the capability of the system as installed and therefore the extent to which other test cycles were or were not performing in an "in-control" and reliable manner in the same installation. The greater the average variability for a test cycle (as represented by the average of the standard deviations of the NOx results for all four fuels) compared to the lowest variability attained for that engine, the more "out-of-control" the testing procedure was for that cycle. For the higher variability cycle, an "excess variability" parameter could then be calculated as the percent increase above the lower variability (the aforementioned "internal standard" for acceptable variability for that engine in this installation).

Table 4: Averaged standard deviations and relative standard deviations for each engine-test cycle combination. Rows with bold text indicate the lowest values achieved for each engine-test cycle combination.

Engine-Cycle Combination	**Average Standard Deviation (all fuels) [g NOx / BHP-hr]**	**Excess Variability**	**Average Relative Standard Deviation (all fuels)**
Legacy/D2	0.068	62%	3.4%
Legacy/NRTC	**0.042**	--	**2.0%**
On-Road NTDE/FTP	0.054	790%	34.8%
On-Road NTDE/RMC	**0.006**	--	**4.2%**
Off-Road NTDE/C1	**0.004**	--	**17.0%**
Off-Road NTDE/NRTC	0.076	1700%	26.4%

The bold-faced rows in Table 4 indicate that the off-road legacy engine is capable of achieving a standard deviation of 0.042 g NOx/BHP-hr or better (from the D2 cycle), the on-road NTDE is capable of a 0.006 standard deviation or better (from the RMC), and the off-road NTDE is capable of a 0.004 standard deviation or better (from the C1 cycle). Test cycles with data sets with averaged standard deviations that are significantly higher than these may therefore be considered out of control—specifically the FTP test cycle for the on-road NTDE (almost 800% excess variability) and the NRTC for the off-road NTDE (1700% excess variability). As discussed earlier, the lack of normality of these data sets is additionally compelling and independent evidence that these experiments were not adequately in control. Additionally, we question if it is even possible to achieve in-control NTDE emission results when switching fuels as frequently as was done in this study given their adaptive control systems and multiple control directives, and whether such extreme fuel switching is even remotely reflective of real world applications and use.

Useful Insights from Relative Standard Deviation

Relative standard deviation (RSD) is included in Table 4 because it is a highly valuable parameter for comparing the relative variability of data sets with significantly different sample means, such as comparing the variability of NOx emissions from a legacy engine to a NTDE. Using RSD as the variability estimation parameter allows side-by-side comparison of the experimental reliability of all six engine-test cycle combinations (Table 4) and the 24 engine-test cycle-fuel combinations (Figure 5) even with engines that produce NOx emissions at different orders of magnitude. A high RSD for a set of repeated tests of the same test cycle-engine-fuel combination is compelling evidence that the experimental system is not in control, with one critical caveat: the two conditions discussed in the "Limits on Assigning Difference" below must be met for RSD to have any relevance.

If the two conditions below are met and a system running the exact same engine-fuel combination that produced a high RSD with one test cycle is capable of producing a low RSD with a different test cycle, the obvious conclusion is that the experimental system was out of control for the test cycle that produced the high RSD results. This exact situation occurred for the two on-road NTDE cycles and would have occurred for the two off-road NTDE cycles except the C1 cycle data violated both of the conditions for the relevance of the RSD. Finally, as in the section above, this conclusion is independently supported by the clear failure to pass the conditions required for parametric analysis that were discussed in the Statistical Methods Evaluation section of these comments.

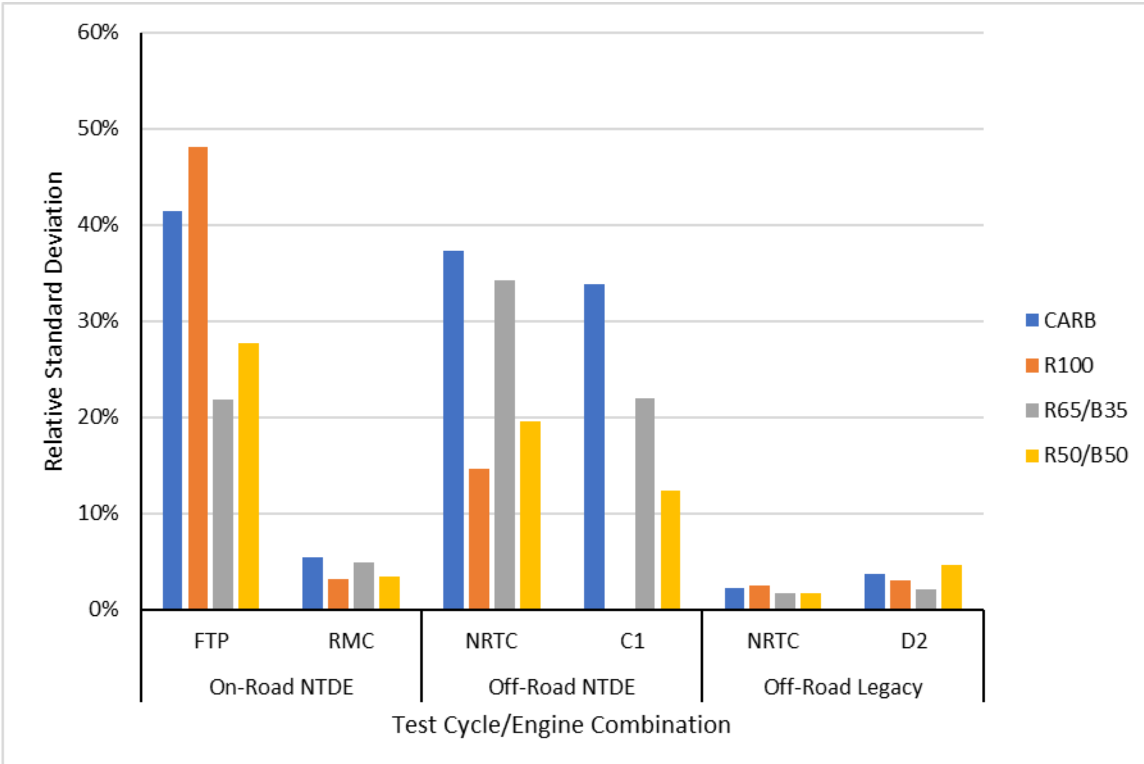

Figure 5: Relative standard deviation comparison for all fuels and test cycle/engine combinations

In addition to providing a valuable visual representation of the relative variability of the 24 data sets, Figure 5 provides an opportunity to assign a reasonable RSD threshold for an "in-control" test cycle using the study's own data. Based on the performance achieved with the lower variability test cycles for each NTD and with both of the test cycles for the legacy engine, an RSD threshold of approximately 5% appears to be a reasonable limit for considering a particular data set to be comprised of potentially repeatable results. RSDs that are substantially greater than 5% thus indicate the potential for a test cycle to be "out of control."

However, as the RSD values for the C1 cycle show, Figure 5 also suggests another problem with the researchers' inappropriate attempt at a comparative statistical analysis of the NTDE data. The off-road NTDE C1 cycle results demonstrated the lowest absolute standard deviation among the data sets, yet they have a highly disparate set of RSD values, with three of the four greater than 10% and one at zero. This type of apparent contradiction can be reconciled by considering the fundamental limitations of the experimental setup, which we do in the next section.

Disqualifying Issue #1: Limits on Assigning Difference

One of the valuable uses of the RSD metric is to inform a researcher of how close a set of measurements is to the limit of quantification for the experimental measurement system. The limit of quantification is the smallest measurement that can be made by a given measurement system with a sufficient amount of confidence that the recorded measurement is meaningful. Every measurement system has inherent sources of variability that, along with the inherent variability of the control variable(s) that are being analyzed, *limits the resolution to which one can confidently make a statistical determination of significance.* Typically, as the magnitude of the value that is being measured decreases, the variability of the measurement value becomes less dominated by the thing that's being measured and more dominated either by other sources of variability in the system or by limitations in the precision of the measurement system.

Two additional concepts should be considered when evaluating whether apparent differences in test results are meaningful or not. The first is the idea of a "minimum measurement interval" of a measurement system. In this discussion, we are considering the minimum measurement interval to be the resolution of the output data (this can also be referred to as "precision" when applied to analytical test results). The data set provided by CARB has a resolution of 0.01 g NOx/BHP-hr, so the minimum measurement interval for this work is, at best, 0.01. The second helpful concept is the idea of a "limit of quantification," which is commonly applied to analytical test methods and indicates the smallest value that can confidently be claimed to represent a non-zero result. At best, the limit of quantification for the measurement system is also 0.01, although this is highly optimistic.

The two requirements for obtaining a meaningful comparison of an actual RSD to an RSD target or threshold (as proposed in the section above) are: (1) the individual measurements must be comfortably above the limit of quantification of the experimental system; and (2) the differences between individual measurements must be significantly greater than the minimum measurement interval. If ***either*** of these conditions is violated, RSD loses its relevance. ***Both*** of these conditions were violated by the C1 data set, and the RSDs for the

C1 data set are therefore meaninglessly large for such a consistent set of measurements (on an absolute basis).

The already invalid practice of drawing conclusions about differences near the minimum measurement interval becomes even more inappropriate when the individual measurements also approach the limit of quantification of the measurement system, as the C1 results do. We can confidently conclude that the limit of quantification for the measurement system itself has been approached because the C1 test results are all between 0.01 and 0.05, and the NOx limit of quantification is at best 0.01 g /BHP-hr and more likely 0.02 or higher. The C1 data set contains results that are consistently low, which is an excellent indication of the capability of the NOx reduction system, but these results approach the limit of quantification of the test system, which results in large and meaningless RSDs and, more important, completely negates the ability to distinguish between fuels. In this situation, it is completely invalid to draw conclusions about the differences between the C1 data sets.

Yet somehow, in Table ES-4 of the report and the preceding discussion, the authors emphasize as significant a 55.1% difference between NOx emissions for R65/B35 and CARB ULSD with reported NOx emissions of 0.021 and 0.015, respectively, for a difference of 0.006—or less than the minimum measurement interval of 0.01. In a science-based regulatory environment, it is difficult to accept how the spurious conclusions about the C1 cycle data could have been allowed into an officially sanctioned study from an agency with the ability and authority to control the use of specific fuels. Worse, this move could provide fodder for an individual or entity to claim institutional bias against a particular fuel.

For the other two engine-cycle combinations with high RSDs, the individual measurements are well above the measurement system's limit of quantification (approximately 0.02), which means the limit of distinction for the results in these two situations is dominated by systematic variability and appears to be at least 0.2 g NOx/BHP-hr. This means the limit of distinction is at least ten times the magnitude of the measurement system's limit of quantification, but this also means greater differences in test results are required to draw meaningful conclusions about differences in system performance.

Note: the definition of "measurement system" in this discussion means everything that influences the measured value. For NTDEs, there are a significant number of components that include the engine itself, the SCR, the test configuration and procedure, the NOx measurement device and other instruments required to calculate NOx emissions in grams / brake-horsepower-hour, and, lastly, the fuels.

Disqualifying Issue #2: Inappropriate Data Exclusion

Rather than appropriately interpreting the high variability of their output parameter as evidence that two of their test systems were not in control, the authors made the decision to keep their preferred random data points (the low ones) and reject the random data they didn't prefer (the high ones); they further chose to do this for one of the out-of-control engine-test cycle (on-road NTDE, FTP) but not the other (off-road NTDE, NRTC). They then proceeded to make conclusions from the manipulated data set. This is poor science and needlessly opens future possible regulatory action to potential litigation. For this reason alone the report in its current state should be retracted and removed from the public record.

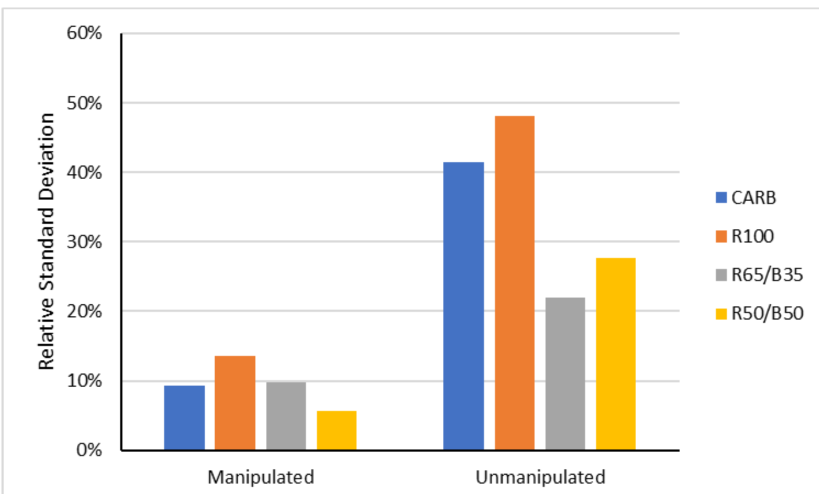

Figure 6: Comparison of Relative Standard Deviations for manipulated (i.e., some data rejected) and unmanipulated (i.e., whole data set) on-road NTDE NOx emissions from the FTP cycle.

Figure 6 compares the variability of the data presented and discussed in the report itself (the "Manipulated" data) versus the variability of the actual data set that was collected for the FTP cycle with the on-road NTDE. The difference is striking. If CARB desires CE-CERT to re-issue an appropriate report on their experimental work, they should present all their data in the report and provide a scientifically-acceptable evaluation of their full set of results with proposed explanations for their high variability data sets.

Review of Possible Explanations for NOx Excursions

As discussed in the opening section of these technical comments, there are three fundamental explanations for the average NOx emissions in a data set to exceed an NTDE's EPA certification limit. Exceeding the engine's designed NOx limits is evidence that at least one of these three fundamental problems occurred:

1. **The fuel's nitrogen content is so high that the SCR's conversion capacity and/or DEF dosing system simply can't keep up with the amount of NOx leaving the engine.**

The nitrogen content of the test fuels was measured and clearly does not exceed the capacity of an SCR system. Fundamental problem 1 does not appear to have contributed in any meaningful way to the high variability of the test results.

Test Fuel	Nitrogen Content (ppm)	Reference
CARB	4.9	Table 3-3
R100	<1.0	
R65/B35	4.8	Table 3-3
R50/B50	6.8	

2. **The ECU hasn't had a chance to fully adapt to the fuel in question (e.g., after dramatic step changes in fuel composition are made).**

Previous researchers pursuing something similar to the objectives of this study at least acknowledged the dynamic nature of the ECU/SCR control schemes and allowed their systems to equilibrate after each fuel change before beginning the test cycles they collected results from (McWilliam and Zimmermann, 2010). It is important to note that even this approach can't eliminate the problem of non-independent data points, but it would substantially improve the reliability of the individual test results. However, even with the aforementioned step changes in fuel composition, one test cycle for each engine was able to produce reasonably consistent results for repeated tests with all four fuels. Therefore, it seems that, while fundamental problem 2 likely contributed to the test result variability, it did not do so in a universal way. Without knowing more about what was done between tests or prior to initiating measurements for each test, it isn't possible to conclude any more about the relative impact of fundamental problem 2.

3. The SCR system is not functioning properly (e.g., component failure, improper engine/ECU configuration, or other inadequate human input)

When all the experimental data are considered in a scientifically and statistically appropriate manner, fundamental problem 3 appears to be the cause of the overwhelming majority of the variability in the two highest variability test cycle-engine combinations (the NRTC cycle with the off-road NTDE and the FTP cycle with the on-road NTDE). Because the same engine, ECU, and SCR components were also able to produce repeatable test results before and after the high variability results, the components either somehow exhibited intermittent failure only during the NRTC and FTP test cycles or were not a significant contributor to the high variability of these two test cycles. Assuming intermittent component failure is highly unlikely in such a controlled operating environment, this leaves only human input as the primary contributor to the high variability for the two most out-of-control test cycles (most likely via improper engine/ECU configuration or otherwise). It seems clear that something(s) went awry when the researchers tried to collect NOx emissions data using the NRTC cycle with the off-road NTDE and the FTP cycle with the on-road NTDE but not with the other two NTDE cycles tested, which produced reasonably consistent results.

Unfortunately, rather than realizing and acknowledging the foregoing, the report's authors went ahead and published not just the out-of-control test results but, most critically, their invalid statistical analyses and unsupported conclusions. It should not be overlooked that in the process they artificially improved the results for one of the out-of-control test cycles by arbitrarily excluding meaningful test results (i.e., data manipulation).

Conclusion

While the full data set provides public interest as a cautionary tale about what can go wrong when researchers try to operate an unfamiliar and highly-complex system for the first time, the report's conclusions are inappropriate, unsupported and could run the risk of opening a

Pandora's box of reactions by those pursuing their own agendas or less than civic-minded pursuits.

This flawed study's publication highlights the unintended consequences of a rush to action and ultimately poses a true risk. Regardless of intent, it provides a public forum potentially enshrining false negative conclusions about fuels with positive impacts to climate change. It also raises the specter of unfounded questions on the effectiveness of NTDEs, potentially impugning a proven technology that dramatically and incontrovertibly reduces criteria pollutants compared to legacy engines.

Even after considering all of the issues with the report in question, one point can be made with confidence: NTDE aftertreatment systems **are** highly effective at reducing tailpipe NOx and PM. As noted by the authors, the tailpipe NOx emissions were typically at least 90% lower than the NOx levels leaving the engines. Furthermore, NTDE PM emissions are at least 99% lower than what would be expected from a non-NTDE. The positive impact that NTDEs, while operating on any of the diesel fuels evaluated in this report, have on California air quality cannot be understated. For this reason, REG fully supports CARB's efforts to advance the penetration of NTDEs in California's on-road and off-road sectors.

ADDENDUM

In the CARB Notice that accompanied the report, CARB asked a number of questions around NOx and NTDE's for further discussion and evaluation. We note with equanimity that CARB made these questions essentially moot when it adopted the Heavy-Duty Engine and Vehicle Omnibus Regulation and Associated Amendments last year.

References

McWilliam, L. and Zimmermann, A., 2010. Emissions and Performance Implications of Biodiesel Use in an SCR-equipped Caterpillar C6.6. *SAE Technical Paper Series*,

Reşitoğlu, İ., Altinişik, K. and Keskin, A., 2014. The pollutant emissions from diesel-engine vehicles and exhaust aftertreatment systems. Clean Technologies and Environmental Policy, 17(1), pp.15-27.

Ross, S., 2014. Introduction to probability and statistics for engineers and scientists. 5th ed. Amsterdam: Elsevier, AP.