ESTIMATED VALIDITY AND RELIABILITY OF ON-BOARD DIAGNOSTICS FOR OLDER VEHICLES: COMPARISON WITH REMOTE SENSING OBSERVATIONS

Anupit Supnithadnaporn and Douglas S. Noonan

School of Public Policy, Georgia Institute of Technology, Atlanta, GA

Alexander Samoylov and Michael O. Rodgers

Aerospace, Transportation and Advanced Systems Laboratory (ATAS), Georgia Tech Research Institute, Atlanta, GA

Post print; final version published as:

Supnithadnaporn, A., Noonan, D. S., Samoylov, A., & Rodgers, M. O. (2011). Estimated Validity and Reliability of On-Board Diagnostics for Older Vehicles: Comparison with Remote Sensing Observations. *Journal of the Air & Waste Management Association*, *61*(10), 996–1004. doi:10.1080/10473289.2011.596738

ABSTRACT

Based on requirements under the Clean Air Act Amendments of 1990, most state vehicle inspection and maintenance (I/M) programs have, since 2002, replaced the tailpipe emission testing with the on-board diagnostic (OBD) II testing for 1996 model and newer vehicles. This test relies on the OBD II system to give the pass or fail result, depending on certain conditions that might cause the vehicle to emit pollution 1.5 times higher than the regulated standard. The OBD II system is a computer and sensors installed in the vehicle to monitor the emission control units and signal if there is any malfunction. As a vehicle ages, its engine, pollution control units, and OBD II system deteriorate. Because the OBD II system's durability directly influences the test outcome, it is important to examine the fleetwide trend in the OBD II test results in comparison with an alternative measure of identifying high emitting vehicles. This study investigates whether the validity and reliability of the OBD II test is related to the age of the OBD II system installed in the fleet. Using Atlanta's I/M testing records and remote sensing device's (RSD) data collected during 2002 to 2005, this research establishes the convergent validity and inter-observer reliability criteria for the OBD II test based on on-road emissions measured by RSDs. The study results show that older vehicles exhibit significantly lower RSD-OBD II outcome agreement than newer vehicles. This suggests that the validity and reliability of

the OBD II test may decline in the older vehicle fleets. Explanations and possible confounding factors for these findings are discussed.

IMPLICATIONS

This research demonstrates the potential worsening validity and reliability of the OBD II test in old vehicles. If the main source of low validity and reliability comes from the OBD II system malfunction, we expect this malfunctioning OBD II fleet will continue to grow in the future. If unchecked, the deterioration of OBD II system may impair the effort of the I/M program to identify high-emitting vehicles and the ultimate objective of reducing the air pollution from automobiles. This result is especially important in a regulatory context where technological and emissions standards dominate environmental policy and yet little attention is paid to the possible degradation of environmental monitors themselves.

INTRODUCTION

To reduce emissions from vehicles, the major sources of urban air pollution, the Clean Air Act Amendments (CAAA) of 1990 mandate the use of clean fuel and clean automotive technology.[1] Under the CAAA 1990, the Environmental Protection Agency (EPA) has required automobile makers to install the second generation onboard diagnostic (OBD II) systems in all light-duty vehicles and trucks (LDVs and LDTs) starting from the 1994 model year.[2] These LDVs and LDTs include all passenger cars, trucks, vans, and sport utility vehicles (SUVs).

In essence, the OBD II is a computer and sensors system that monitors performance of the engine and other emission control components.[3] When the engine or other components malfunction and result in the emissions exceeding 1.5 times the federal test procedure (FTP) standards, the OBD II system turns on the malfunction indicator light (MIL).[4, 5] The system also stores the codes indicating the information important to repair technicians. Because the OBD II system can inform motorists earlier about the malfunctioning parts of vehicles, it can help the vehicles to operate under the optimal conditions and hence minimize emissions.[6]

In addition to mandates for clean fuel and automotive technology, the CAAA 1990 requires that states that violate the National Ambient Air Quality Standards (NAAQS) institute the vehicle inspection and maintenance (I/M) program to ensure proper maintenance of in-use vehicles[1]. Fundamentally, the program requires eligible vehicles registered in the I/M area to pass the inspection on a regular basis. In the past, vehicle inspection mainly involved tailpipe

emission test. Vehicles pass the inspection if the measured emissions are below the EPA standard cutpoints.[7] Since 2002, the EPA has changed the testing technology requirement to the OBD II test for 1996 model year or newer vehicles.[8] Vehicles pass the OBD II test if the MIL is off, no fault codes are stored, and readiness monitor is on.[9] Furthermore, OBD II test can detect the evaporative emissions leaks, which is not available with the tailpipe tests.[10]

OBJECTIVE OF THE STUDY

In the early implementation of OBD II test, concerns were related to the stringency of the OBD II test.[11] During 2000 to 2002, the failure rates of OBD II test were higher than the tailpipe test.[12-16] It should be noted that in 2002, the oldest vehicles undergone the OBD II tests were less than 7 years old. Nonetheless, as vehicles become older, their engines, parts, and OBD II systems deteriorate.[17] It is unclear whether OBD II test results that are based on the older OBD II systems still provide valid and reliable results at the same rate as they did when the systems were newer. For example, vehicles equipped with deteriorated OBD II systems may produce false pass results but keep emitting high levels of pollution on road or alternatively produce false failure outcomes, resulting in additional testing costs with no emission benefits. If these conditions become prevalent among older fleet, reliance on OBD II test for old vehicles may fail to fully achieve the I/M program goal in cost-effectively keeping high-emitting vehicles off the road.

This study explores the possibility that the OBD II system may deteriorate over time and exhibit the invalid and unreliable test outcomes. Unlike a controlled experimental study, this study utilizes vehicle data from Atlanta during 2002 to 2005, which allows the actual fleetwide trends to emerge. In particular, this research compares the OBD II test results with the on-road emissions measured by remote sensing devices (RSD). This study analyzes the agreement between the OBD II test results and the RSD measured emissions and discusses the findings.

RESEARCH METHOD

Validity and Reliability of the OBD II Test

The ***validity***[18] of the OBD II test is defined as the extent to which the results (pass or fail) from the test reflect the vehicle conditions being evaluated. The principal objective of the I/M program is to identify and repair high emitting vehicles. Thus, the OBD II test should identify these high polluting vehicles so that they are repaired. For this reason, the OBD II test is valid

when it yields the 'pass' outcome to a vehicle of which the on-road emissions are within the standard and 'fail' otherwise. Specifically, the study focuses on the test's ***convergent validity***, which refers to a high degree of correlation between the OBD II test outcomes and the criterion emissions.[19] This study uses the on-road emissions measured by the RSDs as the criterion or benchmark. The convergent validity of OBD II test occurs when the results from OBD II test agree with the emissions measured by RSDs.

The ***reliability***[18] of the OBD II test is the degree to which the OBD II test yields the results consistent with the RSD measurement. Particularly, this study concentrates on the ***inter-observer*** or ***inter-rater reliability***, which evaluates the degree to which different testing methods (RSD vs. OBD II) give consistent outcomes (pass or fail) to the same vehicle. Although not generally true, the operational meaning of ***validity*** and ***reliability*** are equivalent in this application.

RSD Measured On-road Emissions as the Criterion

RSD is an instrument that measures tailpipe emissions using the principle of infrared (IR) and ultraviolet (UV) absorption.[20] The source detector module (SDM) projects the beams of IR (for carbon dioxide (CO_2), carbon monoxide (CO), and hydrocarbon (HC)) and UV (for nitrogen oxide (NOx)) across a roadway to the transfer mirror module (TMM). Then, TMM reflects the beam back into a set of detectors in the SDM, which convert the IR/UV energy into an electrical signal.[21] When a vehicle passes through the beam, the interrupted signal triggers the device to measures the ratio of three chemicals to carbon dioxide in the air in front and in the exhaust plume of the vehicle. The difference in detected IR/UV energy due to the absorption by contaminants in the exhaust plume is measured and converted into equivalent concentrations through calibration.[20] The technical details of the RSD emission measurement procedure are provided elsewhere.[22-25]

The main disadvantage of on-road emission measurement by RSD owes to substantial variability of measurement. Collecting RSD in the field poses serious challenges, especially for broken cars with high emissions variability. Previous research using field data in this context, including our study site of Atlanta, [26-27] have had to face problems like identifying control groups sample selection, and measurement accuracy. Since the RSD takes only one possible sample of the emission at a specific driving condition, the result might not reflect the characteristics of the vehicle's emissions generally. Wenzel, Singer, and Slott[28] suggest that tailpipe emission

variation within a vehicle originates from vehicle conditions, fuel quality, driving behavior, engine load, and ambient environment such as temperature and humidity. The ambient conditions tend to influence indirectly the variability via engine load resulting from different operating conditions (such as whether an air conditioner is turned on). In 1993, EPA claimed that RSD mistakenly identified clean vehicles as high emitters but failed to identify up to 90 percent of the high emitters that required repair.[21] Furthermore, EPA stated, "these results are indicative of changes in vehicle emission levels that typically occur when a vehicle is operated under driving conditions different than those observed by the RSD."[21]

During the past decade, however, RSD technology has improved significantly. A variety of studies have examined several aspects of RSD and the data collected by RSD. Walsh, Sagebiel, Lawson, Knapp, and Bishop[29] compare two emissions measurement techniques of RSD and the IM240 test and found that "fleet-wide characteristics of pollution distribution derived from the two techniques were similar in range and shape." (The IM240 test simulates transient driving conditions by using a chassis dynamometer.[30] Although the IM240 test is advanced and reliable,[30] it is inconvenient for motorists, expensive for investors, and complicated for inspectors.[31]) In addition, comparison of individual vehicles also demonstrates that "high emitters in the idle test are also high emitters on-road."[29] A number of studies report similar findings that the on-road measurements show high fleet–average correlation by model year with the IM240 emissions for CO, HC, and NOx.[32-33]

Moreover, Bishop, Stedman, and Ashbaugh[33] denote few key aspects of tailpipe emission variability based on their comprehensive study comparing the emissions measured by different methods: FTP, IM240, idle test, roadside pullover, and RSD. First, and most important, test-to-test emissions variability has similar characteristics across different testing procedures. Second, the main source of emission variability is the vehicle. Third, emissions variability increases with the increasing general emissions. Their findings reveal that vehicles with higher than average emissions tend to have higher emissions variability. Regardless of the methods, they suggest that the only way to eliminate emission variability is to test a vehicle multiple times.

In summary, although the RSD method can yield high measurement variation, the alternative methods also face the same challenge because the variability of emission depends on the vehicles. Recent study has demonstrated that RSD is capable of identifying high emitters.[34] Furthermore, the RSD emission measurement has the important advantage of minimizing the

behavioral bias from the drivers, including no self-selection, no preparation for a test, and actual driving conditions.[11] For this reason, this study will utilize the RSD measured emissions to evaluate the performance of the OBD II test. In particular, the research question examines how the agreement between RSD and OBD II results varies with the age of LDVs or LDTs.

RSD–OBD II Agreement

To demonstrate the convergent validity and inter-rater reliability, the common approach is to identify the agreement between the two measurements of the same vehicle obtained from the two testing technologies: RSD and OBD II. If the two testing methods had measured the same quantity of tailpipe emission in the continuous scale, then the comparison would be straightforward. That is, the correlation between the emissions from the two testing techniques will determine the validity of the OBD II test. Likewise, the deviation of the emissions measured by the OBD II from those measured by the RSD will inform the reliability of the OBD II test. Unfortunately, the OBD II test results yield only binary outcomes of either *pass* or *fail*. To compare the results from the two testing methods, the emission measured by RSDs must be transformed into two values of either *pass* or *fail* with respect to the EPA standard.[35] The consequence of the OBD II's nominal scale measurement (*pass* or *fail*) is that the notion of validity and reliability converge. In other words, whenever the OBD II test is valid, it is also reliable. Because the OBD II test might identify the high emitters either successfully (hit) or unsuccessfully (miss), the resulting agreement between the RSD measurement and OBD II test indicates the validity and reliability of the OBD II test.

To *fail* the RSD measurement, vehicles have to emit (CO, HC, or NOx) at least higher than the standard used in the acceleration simulation mode (ASM) testing. The ASM test measures the tailpipe emission while the vehicle is driven on a dynamometer. The ASM dynamometer includes a fixed inertia weight, a power absorption unit, a torque measurement system, a speed encoder, a warm up motor, a platform lift, and two sets of rolls to support the vehicle's drive wheels. Unlike the IM240 test, the ASM test is a steady state loaded mode test that is relatively easier to operate but performs equally well as the IM240 test.[36] In fact, there are two modes used in the ASM testing[35]: (a) high load/low speed (50 percent load/15 mph) and (b) moderate load/ moderate speed (25 percent load/25 mph). The study selects the ASM2525 standard to determine the *pass* or *fail* result from the RSD measurement because most vehicles in the Atlanta data passed the RSD with moderate speed. For each pollutant, the ASM2525

standard assigns the concentration threshold specific to each combination among 47 categories of estimated test weight (ETW) and 13 groups of vehicle model year.[35]

To *fail* the OBD II test, the vehicle has to fail at least one condition of the three steps. First, to fail the bulb checking, the MIL bulb is not illuminated during the key-on/engine-off (KOEO) status but remains dark during the key-on/engine-running (KOER) status. Second, to fail the readiness monitor checking, the vehicle has more than the allowable number of monitors not set to be ready in the KOER status: two for model year 1996 to 2000 and only one for the newer model year. Third, to fail the trouble code checking, the vehicle fails the diagnostic trouble codes (DTC) that exist in the OBD II system. In addition, the vehicle may fail the OBD II test due to the gas cap being loose or missing. The technical details are available in the EPA guidance for performing OBD system checks.[9]

Table 1 illustrates the classification of RSD–OBD II agreement. Using the information from Table 1, this study applies two statistical methods to determine the RSD–OBD II agreement: (1) raw agreement indices and (2) agreement coefficient (AC1).

Table 1

Raw Agreement Indices. Raw agreement indices include both proportions of overall and specific agreement.[37] *Overall agreement*, also known as accuracy, is the observed proportion of cases in which both RSD and OBD II methods yield the same results (both raters agree), relative to the sample size. Formally, the overall agreement (P_O) is defined in eq 1 using the terms indicated in Table 1.[37]

$$P_O = \frac{n_{00} + n_{11}}{N} \tag{1}$$

The overall agreement proportion is informative but limited in that it treats the positive agreement the same as the negative agreement. As a result, *specific agreement proportion* is suggested by Spitzer & Fleiss[38]: positive (P_{S+}) and negative (P_{S-}) ratings.

$$P_{S+} = \frac{2n_{11}}{2n_{11} + n_{01} + n_{10}} \quad \text{and} \quad P_{S-} = \frac{2n_{00}}{2n_{00} + n_{01} + n_{10}} \tag{2}$$

A positive rating focuses only on the failing category by RSD, OBD II, or both. A negative rating, on the other hand, emphasizes the passing category by RSD, OBD II, or both. Thus, specific *(*to each category) agreement proportions may be considered as estimated conditional probabilities.

To test the statistical significance of the agreement, both overall and specific, the null hypothesis is that the results of both RSD and OBD II methods are independent, meaning the expected marginal probabilities are equal to the observed ones. In the case of the 2×2 table, the test is the same as the test of statistical independence in a contingency table.[39] Accordingly, several statistics are applicable and yield similar results. The study calculates the Pearson Chi-squared, likelihood Chi-squared, and Fisher's exact test. Despite the ease of understanding, raw agreement indices do not take into account the agreement by chance. Therefore, the method that addresses this issue, such as the agreement coefficient (AC1), is necessary.

Agreement Coefficient (AC1). Agreement Coefficient (AC1), in this study, refers to the conditional probability that the results agree from two independent methods of measuring emissions, given that there is no agreement by chance.[40] The estimator of AC1 is κ_γ which is defined in eq 3.

$$\hat{\kappa}_\gamma = \frac{P_O - P_{e\gamma}}{1 - P_{e\gamma}} \tag{3}$$

$$P_{e\gamma} = 2 \cdot \pi_0 (1 - \pi_0) \tag{4}$$

$$\pi_0 = \frac{n_{\bullet 0} + n_{0\bullet}}{2N} \text{ and } \pi_1 = 1 - \pi_0 \tag{5}$$

where P_O is as defined earlier and $P_{e\gamma}$ is the chance agreement probability with a certain degree of uncertainty. Moreover, π_0 and π_1 are the probabilities that a method classifies a vehicle into passing (0) and failing (1) cases respectively. Gwet[40] also defines the variance of κ_γ estimator as shown in eq 6.

$$V(\kappa_\gamma) = \frac{1}{N-1}\left(\frac{P_O(1-P_O)}{(1-P_{e\gamma})^2}\right) \quad (6)$$

To summarize, both raw agreement indices and AC1 are measures of validity and reliability of the OBD II test. We expect that the values of these two agreement measures are lower in the older vehicle fleets, consistent with declining durability of the OBD II system as vehicles age.

DATASETS

Inspection and Maintenance Program Records 2002–2005.

I/M Program Records 2002-2005 consist of the inspection transactions that occurred at the decentralized testing stations located in Atlanta I/M areas. In each transaction, the OBD II test result is recorded automatically by the test technician. The quality and performance of the test analyzer are regulated by the enforcement agency via a biennial certification process and a widespread auditing effort.[41]

Continuous Atlanta Fleet Evaluation (CAFE) 2002–2005.

The Air Quality Group (AQG) in the Georgia Tech Research Institute has collected the CAFE database since 1993 under a contract with the Department of Natural Resources, State of Georgia. The CAFE database contains on-road emissions of vehicles, gathered by RSD, and the corresponding license plates, captured by the video cameras. The license plate information is linked with the vehicle registration database to obtain the vehicle identification number (VIN). The AQG uses the specialized software called VIN Decoder version 2002.01[42] to extract the vehicle characteristic information from VIN. In the year 2000, AQG collected on-road emission data by RSD from 43 sites throughout the Atlanta I/M program area. This study utilizes the data collected during 2002 to 2005.

Matching Identical Vehicles from CAFE with I/M Data.

This study matches VINs from the CAFE 2002–2005 with the same VIN from the I/M Program Records 2002–2005. In both datasets, there are VIN duplicates because vehicles might pass by the RSD many times or they might retest several times. In those cases, the study selects the matched pair with the shortest elapsed time between the RSD measurement and the OBD II

test to minimize any possible intervening factors. For the same reason, this study also utilizes only the observations of RSD measurement taken *before* the OBD II test. (There is some concern about RSD measurement variability affecting our results – a concern of any method testing emissions a single time.[33] Although this study's focus on agreement at the vehicle age group level rather than the individual vehicle level mitigates this concern, the results presented below are also not sensitive to analyzing only vehicles with multiple RSD measurements.) The total observations from matching are 82,523 unique vehicles, which consist of the 11,888, 19,975, 26,241, and 24,419 observations collected in the year 2002 to 2005 respectively. Table 2 shows the distribution of vehicles in the sample classified by age and model year. In this sample, vehicles of model year 1999 hold the largest share (20 percent) whereas those of age 3 account for the largest share (22 percent).

Table 2

RESULTS

Classification of RSD-OBD II Agreement Categories

This research classifies the total observations of 82,523 vehicles into four categories according to the classification scheme in Table 1, as shown in Table 3. The number of vehicles in the pass-RSD-fail-OBD and the fail-RSD-pass-OBD groups are vastly different. If the OBD II systems malfunction randomly, the numbers of observations in these two groups are expected to be comparable, given the variability in the RSD measurement is likely random (minimum behavioral bias in data collection). It is plausible that the number of vehicles in the pass-RSD-fail-OBD group is the result of measurement error in RSD data collection. The significantly larger number of vehicles in the fail-RSD-pass-OBD category, however, suggests a bias in OBD II tests that favors passing high-emissions vehicles. As emissions (measured via RSD) tend to rise with vehicle age, Table 3 is consistent with a greater bias in older vehicle fleets.

Table 3

Table 4 shows the percentage of Table 1 classification disaggregated by age groups. The agreement between RSD measurement and OBD II tests is rather high among new vehicle groups. For 3-year-old vehicles, the RSD-OBD II agreement is approximately 81 percent. Clearly, the majority of the vehicles appear to be clean because they pass both RSD

measurement and OBD II test (80 percent). Given the high RSD measurement variability, the RSD-OBD II agreement on clean vehicles reported here is reasonable when compared to the IM240-OBD II agreement reported in studies prior to 2002. For instance, studies in Colorado, Illinois, and Wisconsin show the IM240-OBD II tests agreeing as often as 97 percent of the time.[6]

Table 4

However, the compositions of the four categories within each age group are different. On average, the share of the clean vehicle category (pass-RSD-pass-OBD) is decreasing at the rate of 4.51 percentage points per year of age. Alternatively, the average increase in the share of the dirty vehicles (fail-RSD-fail-OBD) type is 0.88 percentage points. The trends when RSD and OBD II results disagree are striking. The share of vehicles classified as fail-RSD-pass-OBD is increasing at the average rate of 3.30 percentage points year to year. On the other hand, the share of vehicles classified as pass-RSD-fail-OBD is increasing at the rate of only 0.33 percentage points, which is ten times smaller than the previous disagreed group. Figure 1 illustrates the trends of all 4 categories of RSD–OBD II agreement. Evidently, the fail-RSD-pass-OBD category is the main contributor to the RSD–OBD II disagreement.

Figure 1

Raw Agreement Indices between RSD Measurement and OBD II Test

The raw agreement indices are shown in Table 5. The overall RSD–OBD II agreement ratio is smaller for the older vehicle group, indicating the divergent results between RSD and OBD II. In addition, the specific agreements of the pass-RSD-pass-OBD category are smaller for older vehicle groups, as indicated by the negative rating ratios. This simply suggests that the new vehicle has a higher chance of passing both the RSD measurement and the OBD II test than the older one. In contrast, the positive rating ratios inform that the specific agreements of the fail-RSD-fail-OBD category are higher in the older vehicle groups. This shows that the old vehicle has a higher chance of failing both the RSD measurement and the OBD II test. Moreover, for each age group, both Pearson and Likelihood-ratio Chi-squared statistics demonstrate similar results as shown in Table 4. The null hypothesis of independence between the RSD measurement and the OBD II test results is strongly rejected by both Chi-squared statistics and the Fisher exact test. The results show that the outcomes from the OBD II test are not independent of those of the RSD measurement, as expected.

Table 5

Agreement Coefficient (AC1) between RSD Measurement and OBD II Test

The AC1s in Table 6 illustrate similar results to the overall agreement ratios in Table 5. The AC1 of the older vehicle group is smaller than the AC1 of the newer one. However, the size of AC1 is smaller than the overall agreement ratio because the AC1 formulation is adjusted for the agreement by chance. The variance of AC1 is greater for the old vehicle group than the one of the new group. Figure 2 plots the overall agreement ratio in comparison with the AC1. Both statistics confirm that the RSD–OBD II agreement is lower in the older vehicle groups.

Table 6

Figure 2

DISCUSSION

This study shows that the RSD-OBD II agreement is lower in the older vehicle fleet. The results from the two different statistical methods are striking and consistent. The data selection also ensures the minimal intervening factors between the RSD measurement and OBD II test. Findings in this study strongly suggest that the OBD II system malfunction may be the main source of the low RSD-OBD II agreement among old vehicles.

Despite this careful analysis, some other possible sources of low RSD-OBD II agreement for older vehicles cannot be ruled out due to the unobservable information about vehicles. The intervening influences may exist because the numbers of vehicles classified as fail-RSD-pass-OBD are not comparable to those of pass-RSD-fail-OBD. One possible explanation of the fail-RSD-pass-OBD incident is the ineffective repair. When vehicles are older, their engines and other components usually deteriorate. It is possible that the vehicle owners may repair the vehicles to merely pass the test in the previous inspection cycle. Such repair may not last long until the current inspection cycle when the data are collected, resulting in the high on-road emission detected by RSD. Shortly before the test (for the current inspection cycle), the owners may repair their vehicles in the same manner again and then pass the OBD II test.[43-44]

The other possible explanation of the fail-RSD-pass-OBD event is fraudulent activity. The OBD II system code clearing[45] and use of oxygen simulator[6,46] are a few examples of possible frauds. These frauds would be felonies but are unobserved in the available data. (Less illegal, there may be some systematic bias in the appearance of older vehicles at particular, less

reliable RSD sensors, which could account for the results.) Even though the sources of the lower RSD-OBD II agreement in older vehicle fleets are uncertain, the findings here remain unchanged: the validity and reliability of the OBD II test is lower in the old vehicle fleets. These results are directly relevant to the effectiveness of the I/M program in terms of pollution reduction.

These results appear robust to confounding factors and RSD measurement variability. Restricting the sample to minimize time elapsed between the RSD measurement and the OBD II test minimizes the unobserved intervening factors under the assumption that owners tend to effect repairs closer to their test date. The results presented here, however, are essentially unchanged without the sample restriction. On the one hand, it is comforting that no evidence of intervening factors is found. On the other hand, this may be because effective repairs are unrelated to test dates, which is troubling for a program aiming to spur pre-test repairs. The asymmetry in test disagreements, where false-passes are much more common than false-fails, suggests some strategic behavior by owners. A malfunctioning OBD II system that is too strict may yield a false-fail and then costly repairs of the OBD II system, whereas one that is too lax yields a false-pass and no vehicle repairs of the dirty vehicle. This has serious implications for evaluations of the OBD II program, especially when this asymmetry is more pronounced in older, dirtier vehicles.

RSD measurement variability poses a challenge to establishing that OBD II deterioration accounts for decreased RSD-OBD II agreement rather than simply increased RSD measurement error. Multiple tests of the same vehicle could help if the repeat sampling was of sufficiently high frequency, but Atlanta's CAFE data offer little help here. Only 7 percent of the vehicles have multiple RSD observations, and those few vehicles may be representative of the fleet and may have experienced substantial changes in conditions between RSD measurements. Given the high within-vehicle emissions variation and with so few vehicles being observed multiple times, this approach emphasizes the use of comparisons between vehicle-age groups rather than within-vehicle comparisons to identify the deteriorating reliability. As sensitivity tests, alternative *fail* definitions for the RSD test – at multiples of two and three times the ASM standard – were tested because these higher thresholds will be less sensitive to RSD measurement variability. The main results of deteriorating RSD-OBD II agreement with vehicle age are unchanged at these less strict thresholds. Emissions testing programs that rely on only one test will encounter reliability

problems that have major impacts on effectiveness in cleaning the fleet because it is the dirtier vehicles that have greater measurement variability.

CONCLUSION

The OBD II test has replaced the tailpipe emission in the I/M program since 2002. Because of the different functionality from the tailpipe test, the OBD II test results rely on the OBD II system, which is a component of the vehicles. As vehicles become older, their engine, pollution control equipments, and OBD II system deteriorate. It is questionable whether the OBD II test results of the old vehicles are still valid and reliable. Results from this study show that on the fleetwide basis, the RSD-OBD II agreement is lower in the older vehicle groups. This suggests that the validity and reliability of the OBD II test are lower in the older vehicle fleets. The possible sources of the low agreement are the OBD II system malfunction, ineffective repair, and fraudulent activities. If the main source is the OBD II system malfunction, we expect to observe the increasing share of the fail-RSD-pass-OBD fleet. This malfunctioning OBD II fleet is likely to grow in the near future because neither owners nor inspectors are able to notice this defect. As a result, the necessary repair is less likely to take place.

Lastly, the decay of the OBD II test may impair the I/M program's ability to identify high-emitting vehicles, which may jeopardize its ultimate objective: reducing the air pollution from automobiles. Combating the deterioration of vehicles' emission control systems with an on-board testing technology that itself might deteriorate merits close attention. Unreliable OBD II tests should make the I/M program increasingly unfair as vehicles age (i.e., dirty vehicles pass while clean vehicles face costly repairs). Fair or unfair, the unreliability of OBD II tests is not neutral to overall emissions. The I/M program's reliance on an OBD II technology that appears to decay over time threatens to undermine air quality improvements because (a) disagreements between RSD and OBD II require repairs on clean cars (or remove clean cars from the fleet) and allow dirty cars to avoid repairs, and (b) most disagreements occur in dirty vehicles rather than clean ones. Reforms to emissions testing programs should consider shifting away from uniform testing frequencies across vehicles to obtain greater testing frequency – and more reliable measures – for older vehicles. These vehicles at least may also warrant a testing technology that does not degrade with the vehicle itself. Regardless, the nonrandom inaccuracy of OBD II tests should enter evaluations of the emissions testing programs.

REFERENCES

1. U.S. Environmental Protection Agency. Clean Air Act. The 1990 Amendments to the Clean Air Act. http://www.epa.gov/air/caa/ (accessed June 19, 2009).
2. U.S. Environmental Protection Agency. Control of Air Pollution from New Motor Vehicles and New Motor Vehicle Engines: Regulations Requiring On-board Diagnostic Systems on 1994 and Later Model Year Light-Duty Vehicles and Light-Duty Trucks; Final Rule; 40 CFR Parts 9468 and 9486; *Federal Register* **1993**, 58.
3. U.S. Environmental Protection Agency. *Environmental Fact Sheet: Frequently Asked Questions about On-Board Diagnostics*; EPA 420-F-97-003; Office of Mobile Sources: Washington, D.C., 1997.
4. U.S. Environmental Protection Agency. Control of Air Pollution from Motor Vehicles and New Motor Vehicle Engines: Modification of Federal On-board Diagnostic Regulations for Light-Duty Vehicles and Light-Duty Trucks; Extension of Acceptance of California OBD II Requirements; 40 CFR Parts 86; *Federal Register* **1998**, 63.
5. U.S. Environmental Protection Agency. Federal Test Procedure Revisions. http://www.epa.gov/oms/sftp.htm (accessed June 19, 2009).
6. Clean Air Act Advisory Committee. *On-board Diagnostics (OBD) Policy Workgroup: Findings and Recommendations*; Mobile Source Technical Review Subcommittee, Clean Air Act Advisory Committee: 2002.
7. U.S. Environmental Protection Agency. Cars and Light Trucks: Inspection & Maintenance (I/M) Best Practices. http://www.epa.gov/oms/im.htm (accessed June 19, 2009).
8. U.S. Environmental Protection Agency. Amendments to Vehicle Inspection and Maintenance Program Requirements Incorporating the Onboard Diagnostic Check; Final Rule; 40 CFR Parts 51 and 85; *Federal Register* **2001**, 66.
9. U.S. Environmental Protection Agency. *Performing Onboard Diagnostic System Checks as Part of a Vehicle Inspection and Maintenance Program*; EPA420-R-01-015; Office of Transportation and Air Quality: Washington, D.C., 2001.
10. Reineman, M. *Effectiveness of OBD II Evaporative Emission Monitors - 30 Vehicle Study*; EPA420-R-00-012; Office of Transportation and Air Quality, U.S. Environmental Protection Agency: Washington, D.C., 2000; p 62.
11. National Research Council. *Evaluating Vehicle Emissions Inspection and Maintenance Programs*; The National Academies Press: Washington, D.C., 2001.
12. Barrett, R. A.; Ragazzi, R. A.; Sidebottom, J. A. *Colorado OBD II Vehicle Evaluation Study: Final Report*; Air Pollution Division, Colorado Department of Public Health and Environment: Denver, CO, 2005.
13. Eisinger, D. S.; Wathern, P., Policy Evolution and Clean Air: The Case of US Motor Vehicle Inspection and Maintenance. *Transportation Research Part D* **2008,** 13, 359-368.
14. Cadle, S. H.; Belian, T. C.; Black, K. N.; Carlock, M. A.; Graze, R. R.; Minassian, F.; Murray, H. B.; Nam, E. K.; Natarajan, M.; Lawson, D. R., Real-world Vehicle Emissions: A Summary of the 15th Coordinating Research Council On-road Vehicle Emissions Workshop. *J. Air & Waste Manage. Assoc.* **2006,** 56, 121-136.
15. Cadle, S. H.; Belian, T. C.; Black, K. N.; Minassian, F.; Natarajan, M.; Tierney, E. J.; Lawson, D. R., Real-world Vehicle Emissions: A Summary of the 14th Coordinating Research Council On-road Vehicle Emissions Workshop. *J. Air & Waste Manage. Assoc.* **2005,** 55, 130-146.

16. Gardetto, E.; Bagian, T.; Lindner, J., High-mileage Study of On-board Diagnostic Emissions. *J. Air & Waste Manage. Assoc.* **2005,** 55, 1480-1486.
17. Doll, N. J.; Reisel, J. R., Catalyst Deterioration Over the Lifetime of Small Utility Engines. *J. Air & Waste Manage. Assoc.* **2007,** 57, 1223-1233.
18. Carmines, E. G.; Zeller, R. A. *Reliability and Validity Assessment*; SAGE Publication: Newbury Park, 1979.
19. Hogan, T. P. *Psychological Testing: A Practical Introduction*. John Wiley & Sons, Inc.: Hoboken, NJ, 2007.
20. Bureau of Automotive Repair. Remote Sensing Devices Fact Sheet, 2007. Bureau of Automotive Repair, California Department of Consumer Affairs Web Site. http://www.bar.ca.gov/80_BARResources/02_SmogCheck/Remote_Sensing_Devices.html (accessed March 9, 2009).
21. U.S. Environmental Protection Agency. *Fact Sheet OMS-15: Remote Sensing: A Supplemental Tool for Vehicle Emission Control*; EPA 420-F-92-017; Office of Mobile Sources: Washington, D.C., 1993.
22. Bishop, G. A.; Starkey, J. R.; Ihlenfeldt, A.; Williams, W. J.; Stedman, D. H., IR Long-path Photometry, A Remote Sensing Tool for Automobile Emissions. *Anal. Chem.* **1989,** 61.
23. Bishop, G. A.; Stedman, D. H., Measuring the Emissions of Passing Cars. *Acc. Chem. Res.* **1996,** 29, 489-495.
24. Bishop, G. A.; Stedman, D. H., Automobile Emissions, On-road. In *Encyclopedia of Environmental Analysis and Remediation*; Meyers, R. A., Ed.; John Wiley & Sons, Inc.: Somerset, NJ, 1998; pp 542-552.
25. Bishop, G. A.; Stedman, D. H., Automobile Emissions - Control. In *Encyclopedia of Energy Technology and the Environment*; Bisio, A.; Boots, S., Ed.; John Wiley & Sons, Inc.: Somerset, NJ, 1995; pp 359-369.
26. DeHart-Davis, L.; Corley, E.; Rodgers, M. O., Evaluating Vehicle Inspection/Maintenance Program Using On-Road Emissions Data: The Atlanta Reference Method. *Evaluation Review* **2002,** 26, (2), 111-146.
27. Corley, E.; DeHart-Davis, L.; Lindner, J.; Rodgers, M. O., Inspection/Maintenance program evaluation: Replicating the Denver step method for an Atlanta fleet. *Environ. Sci. Technol.* **2003,** 37, (12), 2801-2806.
28. Wenzel, T.; Singer, B. C.; Slott, R. S., Some Issues in the Statistical Analysis of Vehicle Emissions. *Journal of Transportation and Statistics* **2000,** 3, 1-14.
29. Walsh, P. A.; Sagebiel, J. C.; Lawson, D. R.; Knapp, K. T.; Bishop, G. A., Comparison of Auto Emission Measurement Techniques. *The Science of the Total Environment* **1996,** 189/190, 175-180.
30. Faiz, A.; Weaver, C. S.; Walsh, M. P., *Air Pollution from Motor Vehicles: Standards and Technologies for Controlling Emissions*; World Bank Publications: Washington, D.C., 1996.
31. Ando, A.; Harrington, W.; McConnell, V. *Estimating Full IM240 Emissions from Partial Test Results: Evidence from Arizona*; Resources For the Future: Washington, D.C., 1998.
32. Stedman, D. H.; Bishop, G. A.; Aldrete, P.; Slott, R. S., On-road Evaluation of An Automobile Emission Test Program. *Environ. Sci. Technol.* **1997,** 31, (3), 927-931.
33. Bishop, G. A.; Stedman, D. H.; Ashbaugh, L., Motor Vehicle Emissions Variability. *J. Air & Waste Manage. Assoc.* **July 1996,** 46, 667-675.
34. Cadle, S. H.; Ayala, A.; Black, K. N.; Fulper, C. R.; Graze, R. R.; Minassian, F.; Murray, H. B.; Natarajan, M.; Tennant, C. J.; Lawson, D. R., Real-world Vehicle Emissions: A

Summary of the 16th Coordinating Research Council On-road Vehicle Emissions Workshop. *J. Air & Waste Manage. Assoc.* **2007,** 57, 139-145.
35. U.S. Environmental Protection Agency. *Acceleration Simulation Mode Test Procedures, Emission Standards, Quality Control Requirements, and Equipment Specifications: Final Technical Guidance*; EPA420-B-04-011; Office of Transportation and Air Quality: Washington, D.C., 2004.
36. Patel, D.; Carlock, M. A. *The Relative Benefits from the IM240 and ASM Tests Performed on Vehicles Tested in CARB's I&M Pilot Program*; California Air Resources Board: El Monte, CA, 1997; p 19.
37. Von Eye, A.; Mun, E. Y., *Analyzing Rater Agreement: Manifest Variable Methods*; Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, 2006.
38. Spitzer, R.; Fleiss, J. L., A Re-analysis of the Reliability of Psychiatric Diagnosis. *British Journal of Psychiatry* **1974,** 125, 341-347.
39. Uebersax, J., Statistical Methods for Rater Agreement, 2008, http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm (accessed April 19, 2009).
40. Gwet, K., *Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement between Two or Multiple Raters*; STATAXIS Publishing Company: Gaithersburg, MD, 2001.
41. Georgia Environmental Protection Division, *Rules for Enhanced Inspection and Maintenance Chapter 391-3-20*; Department of Natural Resources: Atlanta, GA 2007.
42. Eastern Research Group, VIN Decoder Version 2002.01; Eastern Research Group, Inc.: Atlanta, GA
43. Cadle, S. H.; Ayala, A.; Black, K. N.; Graze, R. R.; Koupal, J.; Minassian, F.; Murray, H. B.; Natarajan, M.; Tennant, C. J.; Lawson, D. R., Real-World Vehicle Emissions: A Summary of the 18th Coordinating Research Council On-road Vehicle Emissions Workshop. *J. Air & Waste Manage. Assoc.* **2009,** 59, 130-138.
44. Glazer, A.; Klein, D. B.; Lave, C., Clean on Paper, Dirty on the Road: Troubles with California's Smog Check. *Journal of Transport Economics and Policy* **1995,** 29, (1), 85-92.
45. Lawson, D. R., Passing the Test: Human Behavior and California's Smog Check Program. *J. Air & Waste Manage. Assoc.* **1993,** 43, 1567-1575.
46. Cadle, S. H.; Ayala, A.; Black, K. N.; Graze, R. R.; Koupal, J.; Minassian, F.; Murray, H. B.; Natarajan, M.; Tennant, C. J.; Lawson, D. R., Real-world Vehicle Emissions: A Summary of the 17th Coordinating Research Council On-road Vehicle Emissions Workshop. *J. Air & Waste Manage. Assoc.* **2008,** 58, 3-11.

About the Authors

Anupit Supnithadnaporn is a Ph.D. candidate and Dr. Douglas Noonan is an Associate Professor at the School of Public Policy, Georgia Institute of Technology. Alexander Samoylov is a Research Scientist and Dr. Michael Rodgers is a Principal Research Scientist at the Aerospace, Transportation and Advanced Systems Laboratory (ATAS), Georgia Tech Research Institute.

Table 1. Classification of results from RSD measurement and OBD II test

		RSD Measurement Fail (1)	RSD Measurement Pass (0)	Total
OBD II Test	Fail (1)	Fail RSD Fail OBD n_{11}	Pass RSD Fail OBD n_{10}	$n_{1\cdot}$
	Pass (0)	Fail RSD Pass OBD n_{01}	Pass RSD Pass OBD n_{00}	$n_{0\cdot}$
	Total	$n_{\cdot 1}$	$n_{\cdot 0}$	N

Table 2. Distribution of vehicles categorized by age and model year in the logistic model

Age	**Model Year of Vehicle**							**Total**
(year)	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	
3				3,292	5,014	5,337	4,874	18,517
4			3,049	4,429	5,306	4,213		16,997
5		2,804	4,062	4,811	4,244			15,921
6	2,459	3,447	4,103	3,823				13,832
7	2,958	3,625	3,066					9,649
8	2,988	2,567						5,555
9	2,052							2,052
Total	10,457	12,443	14,280	16,355	14,564	9,550	4,874	82,523

Table 3. Number of vehicles in the sample classified by the results from RSD measurement and OBD II test

		RSD Measurement		
		Fail (1)	Pass (0)	Total
OBD II Test	Fail (1)	2,409	3,336	5,745
	Pass (0)	19,821	56,957	76,778
	Total	22,230	60,293	82,523

Table 4. Percent of vehicles grouped by age in different agreement categories

Age (year)	Pass RSD Pass OBD (n_{00})	Fail RSD Pass OBD (n_{01})	Pass RSD Fail OBD (n_{10})	Fail RSD Fail OBD (n_{11})
3	79.94	15.60	3.47	0.99
4	75.55	18.87	3.89	1.69
5	68.59	24.53	4.02	2.85
6	62.20	28.72	4.65	4.44
7	58.69	32.33	4.15	4.83
8	54.76	36.02	4.27	4.95
9	52.88	35.38	5.46	6.29
Average Annual Time Trend	-4.51	3.30	0.33	0.88

Table 5. Raw agreement indices and the statistical tests by vehicle age group

Age (year)	Overall Agreement Ratio[a]	Specific Agreement[a]		Pearson Chi-squared[b]	Likelihood-ratio Chi-squared[b]	Fisher Exact[c]
		Positive Rating Ratio	Negative Rating Ratio			
3	0.809	0.094	0.893	20.155	18.663	
	(0.003)	(0.002)	(0.035)	(0.000)	(0.000)	(0.000)
4	0.772	0.129	0.869	57.740	52.702	
	(0.003)	(0.003)	(0.039)	(0.000)	(0.000)	(0.000)
5	0.714	0.166	0.828	117.675	108.937	
	(0.004)	(0.003)	(0.046)	(0.000)	(0.000)	(0.000)
6	0.666	0.210	0.789	153.611	145.848	
	(0.004)	(0.004)	(0.052)	(0.000)	(0.000)	(0.000)
7	0.635	0.209	0.763	112.889	108.928	
	(0.005)	(0.004)	(0.056)	(0.000)	(0.000)	(0.000)
8	0.597	0.197	0.731	37.843	37.167	
	(0.007)	(0.005)	(0.060)	(0.000)	(0.000)	(0.000)
9	0.592	0.235	0.721	15.804	15.573	
	(0.011)	(0.009)	(0.062)	(0.000)	(0.000)	(0.000)

[a] Standard errors in parentheses for overall and specific agreement ratios
[b] p-values in parentheses for Pearson and Likelihood-ratio Chi-squared
[c] Fisher exact tests report only p-values

Table 6. Agreement coefficients (AC1) and their variances by vehicle age group

Age (year)	Agreement Coefficient (κ_γ)	Variance (V_{κ_γ})
3	0.765	0.000013
4	0.705	0.000017
5	0.601	0.000025
6	0.500	0.000036
7	0.434	0.000058
8	0.354	0.000111
9	0.329	0.000318

List of Figures

Figure 1. Percent of vehicles by age group, classified by the results from RSD measurement and OBD II test

Figure 2. Overall agreement index and Agreement Coefficient (AC1) of different vehicle-age groups