Evaluation of alternative statistical procedures for the evaluation of in-service conformity emissions tests under the real-driving emissions (RDE) regulation

V. Franco, Z. Kregar, P. Dilara*

*Contact : vicente.franco@ec.europa.eu.

Contents

Executive summary

In the context of the technical preparation for the fourth regulatory package of the real-driving emissions (RDE) regulation, several proposals were made by the European Commission and other stakeholders to modify the current statistical procedure used for in-service conformity (ISC) of emissions in order to adapt it for the introduction of RDE tests. This paper presents the results of the evaluation of these alternative statistical procedures to be applied to emissions tests of light-duty vehicles in-service conformity for both the RDE and WLTP tests.

All the alternative statistical procedures that were proposed (see section 2.3) are in principle suitable for application to RDE tests, and they bring about improvements from the current procedure in key aspects such the re-balancing of manufacturer and consumer/environmental risks, a reduction of the overall testing burden, or an improved treatment of outliers. However, not all improvements could be fully realised by every procedure.

In an initial comparison of the risk curves of all the "ISO" procedures under evaluation (see section 2.3), the *ISO COM* approach was found to deliver the best compromise between the risk balance of the current, UNECE Regulation 83 approach[1]—which tends to favour the manufacturer—and approach used for emissions in-service conformity testing of heavy-duty vehicles[2]—which tends to favour the interests of the consumer/environment.

A more detailed evaluation of all the ISC procedures was performed using a simulation approach, which was implemented in spreadsheet software. A final version of the simulation spreadsheet has been uploaded to CIRCABC along with this paper for all stakeholders to examine the results presented herein.

The simulation of the ISC procedures (whose results are discussed in section 4.2) allowed an evaluation of the differences in testing burdens and treatment of outliers among the ISC procedures for a series of scenarios proposed (and described in section 2.2). Considering its strong performance according to the evaluation criteria described in section 3 of this discussion paper, **we put forward *ISO COM (with custom outlier treatment)* as the EC's proposal for a statistical procedure to be used in the evaluation of emissions in-service conformity tests.**

1 Background

According to Commission Regulation EU 2016/427[3] (first regulatory act of the real-driving emission regulation, RDE), not-to-exceed emission limit values apply throughout the normal life of the vehicles. Therefore, in preparation of the fourth regulatory package of RDE—which deals, among other issues, with in-service conformity (ISC) testing of real-driving emissions—the Commission is evaluating suitable statistical methods to complement ISC testing traditionally done up to now with the test cycle (NEDC/WLTP) in the lab, with on-road, RDE tests.

Following stakeholder input provided in the context of the RDE-LDV technical working group, a number of guiding principles for RDE ISC testing were put forward: it was agreed that the same statistical method should be applied by all actors participating in ISC (*i.e.*, type-approval authorities, manufacturers, and accredited emission testing laboratories), and that the ISC statistical procedure would follow the principles of *sequential sampling* (where the sample size is initially kept small, but can be increased if a decision cannot be reached with a sufficient degree of confidence). In order to keep the testing burden low, and in recognition of the increased logistical difficulties[1] brought about by RDE, it was decided that *the maximum sample size would be limited to ten vehicles* (instead of the value of twenty applied to ISC tests under the NEDC framework[2]). This will hold for both WLTP and RDE testing.

The majority of the proposals ("ISO" proposals) were based on sequential sampling by attributes supported by a "pass-fail" chart as described in international standard ISO 8422[4]. On the other hand, the Netherlands proposed to adapt a statistical procedure based on the one used to verify conformity of production (CoP), as this was implemented recently in the WLTP regulation.

In order to enable a meaningful comparison of the alternative procedures, the evaluation of was performed through a simulation, which was implemented in spreadsheet software. The final version of the spreadsheet file where the simulations can be run is available in the CIRCABC site of the RDE-LDV working group.

1.1 Managing the increased variability of RDE tests

There are important differences between ISC testing with the RDE procedure and with the laboratory tests (be it NEDC or WLTP) that were considered during the design and evaluation of the ISC procedures for RDE. For laboratory tests, conditions of the test are controlled, and eventual deviations from type-approval emissions during ISC are mostly due to ageing (and also the variability of the production processes of individual components of emission control systems). For RDE tests, on the other hand, several test conditions with an impact upon emissions may vary across tests. The variability of RDE ISC emissions will therefore be due to ageing and variability of production processes, but also to differences in testing conditions.

Since RDE emission requirements apply throughout the normal life of a vehicle type-approved according to Regulation (EC) No 715/2007[5], the increased variability due to real-world testing conditions needs to be managed by the vehicle manufacturer through improvements in the real-

[1] Longer time required for testing could lead to increased difficulty to get vehicles to test from the general population.

[2] According to the analysis presented in this paper, this can be accomplished without compromising the robustness of the procedures.

world emissions performance over a broad range of driving conditions in order to comply with the legislation and avoid ISC challenges. In-service conformity testing is therefore a key element of the RDE regulation because it ensures that vehicles will be tested under a wide range of conditions, throughout their normal life. As acknowledged in recital 7 of Commission Regulation (EU) 2016/427:

"An individual RDE test at the initial type-approval cannot cover the full range of relevant traffic and ambient conditions. Therefore in-service-conformity testing is of utmost importance for ensuring that a widest possible range of such conditions is covered by a regulatory RDE test, thereby providing for compliance with the regulatory requirements under all normal conditions of use."

In order to achieve the necessary emission reductions, the ISC statistical approach used in RDE needs to be robust. In particular, the statistics should not be used to absorb the variability of RDE emissions, as this would be detrimental for overall compliance levels.

2 Simulation methodology

The simulation of the ISC procedures was implemented in a single Microsoft Excel spreadsheet that represents multiple instances of the ISC procedures. The user defines the expected distribution of the emissions results and then the spreadsheet simulates 3,000 runs of each ISC procedure based on the same sequences of 10 vehicles randomly drawn from the population.

A new population of vehicles (characterised by their emissions over an RDE test) is automatically generated every time the spreadsheet is modified, but the user can change this under Formulas->Calculation options->Manual. The population is divided into the same 3,000 sequences of 10 vehicles[3], which can be viewed in each separate spreadsheet tab named after the ISC procedures.

2.1 Data input

The main data input to the simulation are the emissions characteristics of the population. These emissions characteristics are initially modelled using a generalised beta distribution (such as the one in Figure 1). Such distributions are a good basis to approximate the distribution of emissions results resulting from RDE testing, as they can be adjusted to feature increased variance, a certain level of skewness and the significant 'right-hand tail' that is expected from on-road tests.

The parameters of the beta distribution ("shape" parameters *a* and *b*, and upper and lower limits) used in the simulations can be adjusted to modify its skewness and how much the results spread across the compliance space. For the sake of simplicity, the upper and lower limits of the distributions are expressed as multipliers of the not-to-exceed limit; this means that an individual emissions result value is in compliance with the NTE limit as long as it is equal or smaller than one.

[3] Except for the ISO Reg. 83 procedure, where the same population is divided into 1,500 samples of 20 individuals by joining what in the other procedures are sample 1 and sample 1501, sample 2 and sample 1502, and so forth.

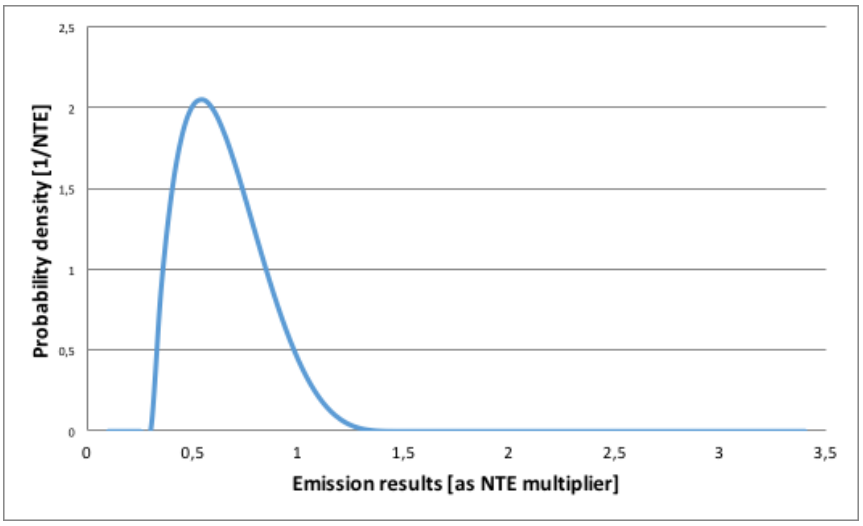

Figure 1. Probability density function (generalised beta distribution; example shown corresponds to simulation scenario 1)

There are six pre-defined scenarios (see section 2.2 below), mostly defined by the shape parameters and by the lower and upper bounds of the beta distribution (*i.e.*, the lower and higher values at either end of the probability distribution function). These are set for all example scenarios, but they can be manually changed in the spreadsheet if desired.

A very simplified model that can incorporate "additional outliers" is also included. The prevalence and importance of outliers can be adjusted by indicating the % of the highest values in the beta distribution that will be multiplied in value and a multiplier for these values (*e.g.*, in scenario 3, the 10% highest values get multiplied by 5; these additional outliers can be removed by setting the % of outliers to 0, or the multiplier to 1).

The scenario to be simulated can be changed by changing the input in the cell in the top left corner of the "Dashboard" worksheet. The main ISC simulation results, which include relevant metrics on both the simulated population (common for all ISC options) and the outcome of the simulations (different among ISC procedures), as well as some representative charts, can be found at the bottom of the "Dashboard" worksheet.

2.2 Simulation scenarios

Six 'default' scenarios are proposed in the spreadsheet uploaded to CIRCABC. These scenarios represent plausible distributions with variable degrees of noncompliance (understood as % of valid RDE tests with emissions results above the NTE limit).

Noncompliant tests results could come from insufficient performance of the emission control systems for certain driving conditions corresponding to a valid RDE test (resulting in a right-hand tail of the distribution overshooting the NTE limit value) or from a low durability of the emission control systems (which could manifest itself in the presence of emission outliers; *i.e.*, a minority of vehicles with unusually high emissions separate from the main mode of the distribution).

The default scenarios (see Table 1) do not include populations with very high proportion (close to 100%) of compliant vehicles since these would result in pass rates equal or very close to 100% for all of the statistical procedures.

The scenarios included in the final version of the simulation spreadsheet are as follows (for quantitative parameters of the underlying distributions, see Table 1):

***Scenario 1**: Standard "optimistic", no additional outliers*. This scenario represents a population with moderate levels of noncompliance (~5%). This distribution has a relatively short right-hand tail that tapers off at a maximum value of 1.5. No additional outliers are modelled in this distribution.

***Scenario 2**: Standard "pessimistic", no additional outliers*. This scenario represents a population with significant levels of noncompliance (~35%). This distribution has a longer right-hand tail that tapers off at a maximum value of 2.0. No additional outliers are modelled in this distribution.

***Scenario 3**: Compliant population with outliers (10% noncompliant rate with a multiplier of x5)*. This scenario represents a 'baseline' population with very low levels of noncompliance. This distribution has very short right-hand tail that tapers off at a maximum value of 1.1. A 10% of additional outliers with a multiplier of 5 are modelled on top of this distribution. This distribution is intended to evaluate the ability of ISC procedures to detect a significant share of outliers with a high emissions level.

***Scenario 4**: Compliant population with outliers (5% noncompliant rate with a multiplier of x10)*. This scenario uses the same 'baseline' population as scenario 3. A 5% of additional outliers with a multiplier of 10 are modelled on top of this distribution. This distribution is intended to evaluate the ability of ISC procedures to detect a moderate share of outliers with a very high emissions level.

***Scenario 5**: ACEA compliant*. This scenario represents a population with average emissions of 0.75 times the NTE limit, and a proportion of non-compliance of approximately 24%. This distribution has a significant right-hand tail that tapers off at a maximum value of 4.0.

***Scenario 6**: ACEA non-compliant*. This scenario represents a population with average emissions of 1.78 times the NTE limit, and a proportion of non-compliance of approximately 74%. This distribution has a long right-hand tail that tapers off at a maximum value of 11.0. The parameters of the distributions of scenarios 4 and 5 were proposed by the European Automobile Manufacturers Association (ACEA) in the context of the RDE-LDV technical working group, following a similar logic to the one behind scenarios 1 and 2.

Table 1. Main parameters of the underlying distributions, by scenario

Scenario ID	Generalised beta distribution parameters				Additional outlier parameters		Average emissions [NTE limit]	% of non-compliant*	Standard deviation* [NTE limit]
	a	*b*	*min*	*max*	*%*	*multiplier*			
1	2	5	0.3	1.5	0	1	0.64	5.0	0.19
2	2	5	0.5	2.0	0	1	0.93	35.0	0.24
3	2	5	0.3	1.1	10	5	0.84	10.1	1.03
4	2	5	0.1	1.1	5	10	0.72	4.9	1.56
5	2	10	0.5	4.0	0	1	0.75	23.9	0.40
6	2	11	0.1	11.0	0	1	1.78	73.8	1.04

**These values are calculated from a random sample of 3,000 results, and may vary slightly across different simulation runs*

2.3 Statistical procedures evaluated

The statistical procedures that were simulated in the spreadsheet are as follows:

- ***NL margin***: This procedure models the so-called "Dutch statistics", an adaptation of the conformity of production (CoP) WLTP statistics[6], incorporating a "statistical margin" that increases the fail threshold (especially at the beginning of the ISC process, when a low number of vehicles have been tested). Following a recommendation of TNO, the margin in the simulations takes a fixed value of 0.3[4] (meaning that, for an initial sample of 3 vehicles, the average of 3 results can be up to 30% above the NTE limit value before resulting in a fail; note that this margin decreases with sample size and is zero for 10 vehicles). Other values for the margin can be simulated by adjusting the parameter in the spreadsheet.

The equations describing the *NL margin* approach can be found in the corresponding tab of the spreadsheet. This is the only procedure with a built-in treatment of outliers based upon CoP statistics, whereas the remaining approaches are adaptations of the current approach. For this reason, the balance between manufacturer and consumer risk cannot be visualised in a pass-fail chart directly comparable to the rest of the statistical procedures under evaluation.

- ***ISO COM***: This procedure models an adaptation of the procedure from Reg. 83 (which in turn is based on ISO standard ISO 8422; hence the name), where the maximum sample size was curtailed to 10 vehicles and the producer/consumer risks were re-balanced. The pass-fail chart for this and all of the "ISO" procedures is found in Figure 2. Note that the proportion of green, 'pass' cases and red, 'fail' cases in the figure is perfectly balanced up until the maximum sample size [10 vehicles], in which case a small bias in favour of the manufacturer is introduced to close the statistical procedure. One of the distinguishing features of this procedure is that it is very balanced, with 3 out of 3 fails leading to a failed sample, while 3 out of 3 passes lead to a pass. The thresholds for the number of vehicles required for a 'pass' or a 'fail' decision (in absence of outliers) are indicated at the bottom of each pass-fail chart represented in Figure 2.

- ***ISO COM (custom outlier treatment)***: This procedure is identical to *ISO COM*, except that the parameters that characterise the treatment of outliers were modified. In this proposal, the threshold for intermediate outliers is lowered from 1.5 to 1.3 (and two of such outliers lead the sample to fail), while the threshold for extreme outliers is kept at 2.5 (but the sample would fail with a single extreme outlier). The reason for proposing an intermediate outlier threshold at 1.3 instead of 1.5 is

[4] This was the value of the margin used for the purposes of the evaluation, but it remains modifiable by the user in the spreadsheet (from the 'Dashboard' worksheet).

to increase the protection against insufficient performance of the emission control systems and to limit the situations in which the variability of emissions results due to testing conditions is 'absorbed' by the statistical procedure. The procedure with custom outlier treatment is otherwise identical to the *ISO COM* procedure, using the same balanced pass-fail chart.

- ***ISO ACEA***: This procedure also models an adaptation of the Reg. 83. As in the *ISO COM* procedure, the maximum sample size is curtailed to 10 vehicles and the producer/consumer risks were re-balanced. The producer risk for this procedure is somewhat lower than for the *ISO COM* procedure, and also lower than for either the ISO ICCT or the ISO HDV procedures (see pass-fail chart in Figure 2).

- ***ISO ICCT***: This procedure models yet another adaptation of the ISC statistical procedure of Reg. 83. As in the *ISO COM* and *ISO ACEA* procedures, the maximum sample size is curtailed to 10 vehicles and the producer/consumer risks are re-balanced. The producer risk for this procedure is somewhat higher than for either the *ISO COM* or the *ISO ACEA* procedures, but lower than for the *ISO HDV* procedure (see pass-fail chart in Figure 2).

- ***ISO HDV***: This procedure models the ISC statistics used in the heavy-duty regulation. Earlier in the evaluation of the statistical procedures, some shortcomings were identified in the *ISO HDV* procedure that rendered it unsuitable for application to RDE-LDV (namely an increase in the average testing burden, and a tendency to favour the interests of the consumer/environment that can be observed in the prevalence of red, 'fail' cases over green, 'pass' cases in the pass-fail chart; see Figure 2).

- ***ISO Reg. 83***: This procedure models the original statistical procedure of Reg. 83, where the maximum sample size is 20 vehicles. This procedure is not strictly part of the evaluation, but it is nonetheless modelled as a reference.

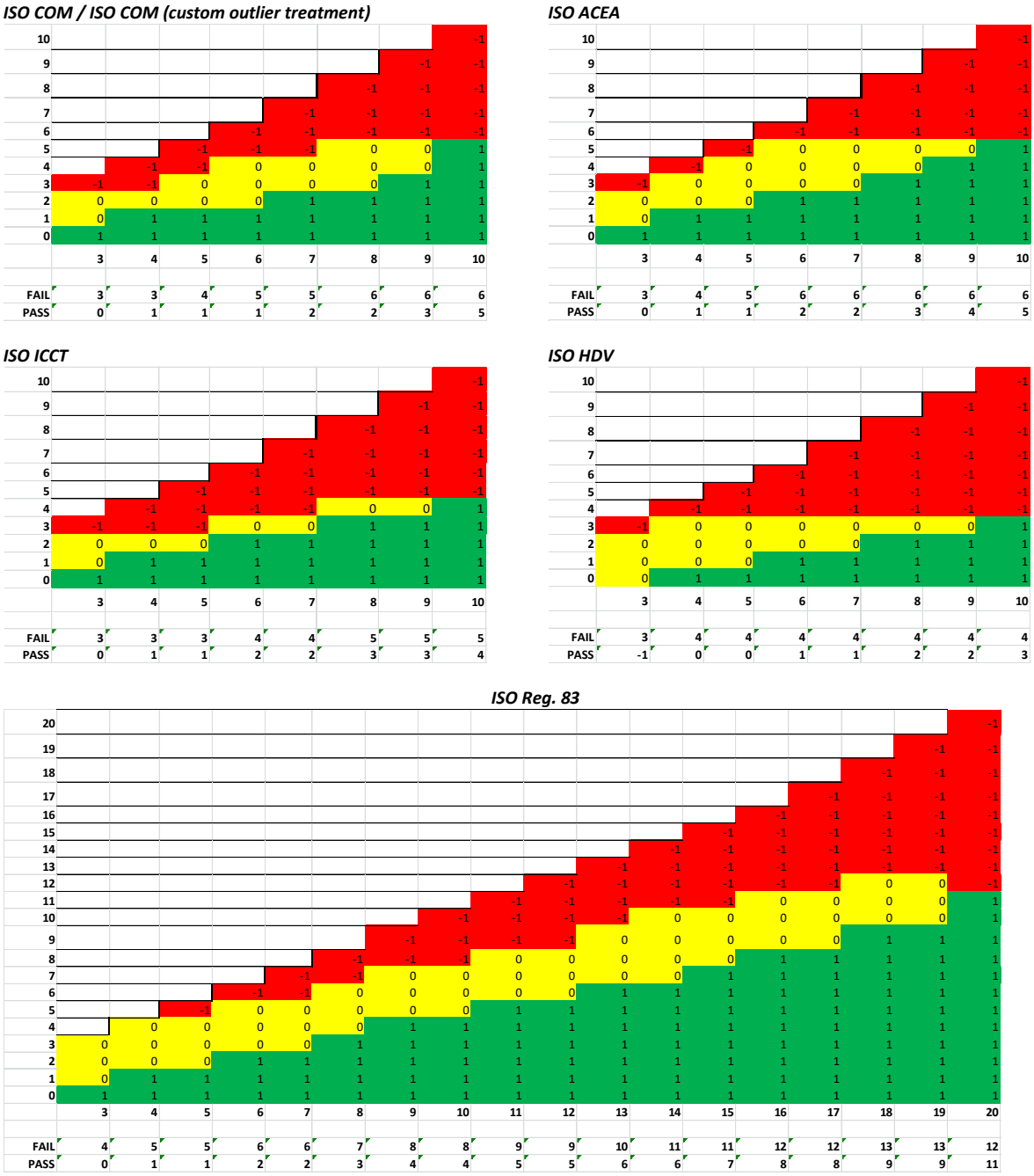

Figure 2. Pass-fail charts for the "ISO" procedures

For the "ISO" procedures, it is possible to compute the probability of a 'pass' decision at the maximum sample size as a function of the non-compliance rate of the population (excluding the effect of outliers) and to plot the corresponding so-called operating characteristic curves to visualise the distributions of producer and consumer risks underlying the pass-fail charts. This is done in Figure 3, where it can be observed that all the 'new' proposals (*ISO ACEA, ISO ICCT and ISO COM*) achieve a rebalancing of the risks that lies between ISO Reg. 83—which tends to favour the manufacturer—and *ISO HDV*—which tends to favour of the interests of the customer/environment. Please note that this chart does not reflect the effect of outlier detection thresholds, which is changed in the ISO COM (custom outlier treatment).

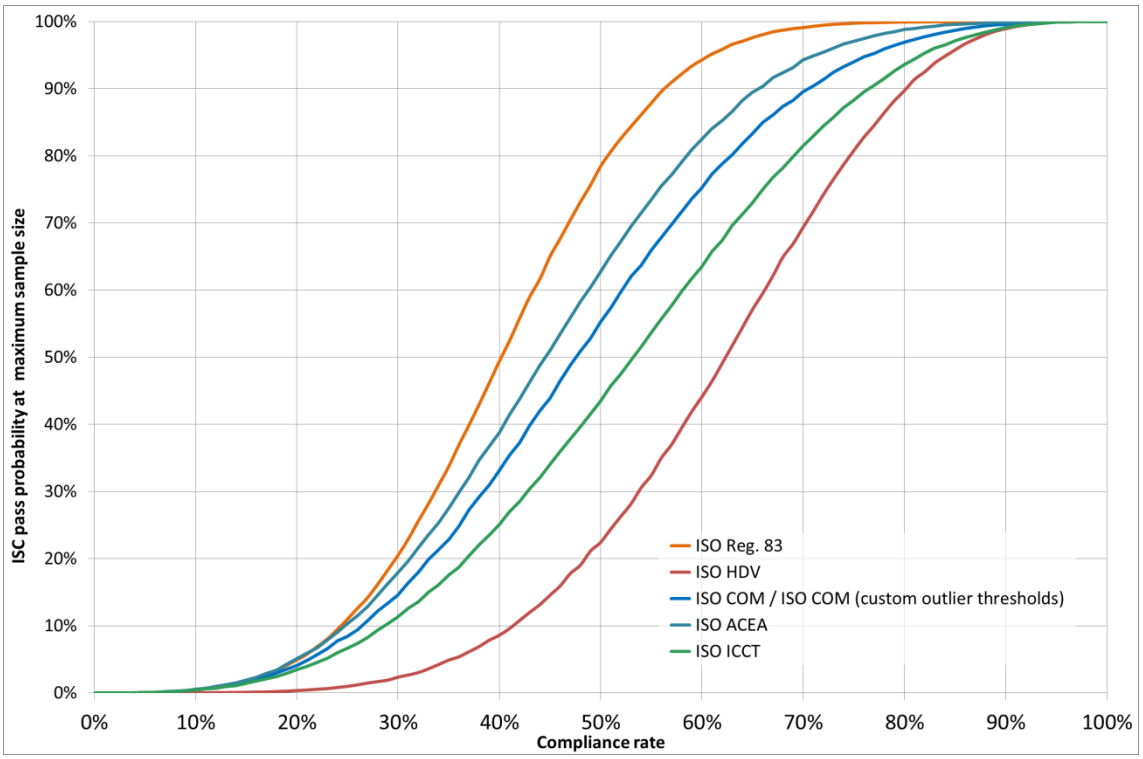

Figure 3. Operating characteristic curves for the "ISO" procedures (results based on a simulation by ICCT)

3 Evaluation criteria

Qualitatively speaking, all of the ISC procedures under evaluation compel the manufacturer to keep the proportion of valid tests resulting in non-compliance low, and to pay special attention to outliers in order to have a very low statistical chance of failing the ISC procedure. But in order to evaluate the extent, and the effectiveness with which they do so, appropriate metrics relating to compliance and to the correctness of the outcome of the ISC procedure have to be established.

3.1 Compliant / non-compliant vehicle family

During ISC, the compliance of a vehicle family is assessed on a statistical basis, trying to balance the risks of the manufacturer (notably those associated with an emissions investigation and a potential vehicle recall) and those of the customer/environment (undue excess emissions).

A vehicle family will be considered non-compliant if it fails the ISC procedure (an outcome with an associated stochastic component) *and* an authority determines—following a formal investigation—the causes of high emissions are attributable to the vehicle (*e.g.,* hardware durability issues, or insufficient performance of the pollution control systems).

3.2 'Correctness', 'fairness' and 'unfairness' of the outcomes of the ISC procedure

In order to evaluate the performance of the different ISC statistical procedures simulated in the spreadsheet, it is very important to establish a sound criterion of what constitutes a "correct" decision.

For the purposes of the simulation, it was considered that it is correct for an individual ISC process to result in a 'pass' if the average of the results considering the maximum sample size[5] (10 vehicles for all options except for *ISO Reg. 83*, where it is 20) is equal or below the NTE limit. Conversely, it would be correct for an individual ISC process to result in a 'fail' if this average is above the NTE limit. 'Incorrect' decisions give rise to two types of errors: the type I error (result "unfair" towards manufacturer) happens whenever an ISC process that should have resulted in a 'pass' according to the criterion results in a fail during the simulation, and the type II error (result "unfair" towards the environment/customer) that happens whenever an ISC process that should have resulted in a 'fail' has a 'pass' outcome in the simulation.

This choice of the "average emissions at maximum sample size" criterion for correctness and fairness of decisions has two main advantages: first, for a given population, it is possible to see both correct and incorrect decisions and, within the incorrect decisions, to discern the proportion of decisions that were "unfair" to the manufacturer and those "unfair" to the environment. The frequency of the different outcomes can be used to evaluate the performance of the method, including consumer and producer risks, which are simultaneously estimated on the basis of the same input. A second advantage of the "average emissions at maximum sample size" criterion has a direct link to the reality of the ISC process because the actual pass/fail decision is taken on the basis of a few data points. The realistic "best" decision would be the one taken using the max number of points reasonably expected to be available; *i.e.*, at the maximum sample size.

3.3 Performance evaluation of the ISC procedures

The simulation spreadsheet was used to evaluate the alternative ISC procedures on the basis of three main performance criteria:

- *High "correct" decision rate.* A key desirable characteristic of an ISC procedure would be to consistently come to the "correct" decision for a high percentage of all simulated ISC procedures. As secondary related sub-criteria, we also evaluated the *balance in 'unfair' decisions* (i.e., for a given proportion of 'unfair' decisions, it is preferable to have a balanced distribution between type I and type II errors and the *robustness to input* (*i.e.*, a proportionate response to both 'optimistic' and 'pessimistic' simulated populations).

- *Low testing burden.* From the point of view of the test performer, it is important to be able to reach a decision after a small number of tests. This can be evaluated by looking, *e.g.*, at the average sample size at the time of reaching a decision.

- *Treatment of outliers.* Even a small proportion of emission outliers in a vehicle family can account for a significant share of total emissions. An adequate treatment of outliers is an important characteristic of any ISC procedure. In particular, a good ISC procedure should be able to detect and qualify outliers.

[5] An alternative valid criterion would have been to compare the decision reached by each individually simulated ISC process to the decision that would have been reached at the maximum sample size. However, when comparing different ISC procedures, the column of "decision at maximum sample size" will not be the same for all procedures (even with the same input for all procedures) because the way a decision is reached is method-specific. This would introduce a bias in the comparison.

We note that high (or low) 'pass' or 'fail' rates were not considered a performance metric for the ISC procedures under evaluation. However, we paid attention to these values in comparison to the share of non-compliant results in the input population. As a general principle, we considered it acceptable for the actual pass rate to be somewhat higher than the "compliance" rate (*i.e.*, for a compliance rate of 80%, the simulated pass rate should be at least 80%), as long as the population under study had reasonable compliance characteristics (average emissions below the NTE limit and a low proportion of outliers).

4 Results and discussion

In this section we present and discuss the outcomes of the simulation exercise. The reader is invited to open the simulation spreadsheet to explore the results provided by the spreadsheet in full detail. We note that many of the numeric results quoted in this section are calculated from a random sample of 30,000 results, and may therefore vary slightly across different simulation runs (upon a recalculation of the spreadsheet leading to the generation of a new random sample).

4.1 On the use of simulation

The simulation approach used allows the simultaneous comparison of several ISC procedures on the basis of the same input and using the same metrics. This is particularly useful considering that the comparison includes ISC procedures with significant differences in their basic approach. Using synthetic, plausible distributions for RDE emissions results instead of relying on (scarce) measured data increases the flexibility or the evaluation. The proposed spreadsheet remains fully editable and can in any case be modified by the user to simulate the outcome of the ISC procedures to any input.

4.2 Simulation of ISC statistical procedures: results by scenario

In this section, we qualitatively evaluate the performance of the different ISC statistical procedures by examining the outcomes of the simulation process for each one of the scenarios proposed in section 2.2. Fully detailed, quantitative results can be explored in full detail by opening the final version of the simulation spreadsheet and running a few simulations.

Scenario 1: Standard "optimistic", no additional outliers

For this scenario, all of the ISC procedures have a pass rate of approximately 99.8-100.0%. These pass rates practically coincide with "correct" decisions because there are very few or no instances at all in which the average emissions at the maximum sample size exceed the NTE limit (as expected from a distribution with moderate levels of noncompliance).

The global average testing burden (considering both 'pass' and 'fail' outcomes) is low for all ISC procedures (<3.5 vehicles tested at the time of decision) except for the *ISO HDV* procedure (>4 vehicles). This is in line with expectations, considering that the *ISO HDV* procedure can only reach a 'pass' decision after a minimum of four vehicles tested (see Figure 2).

The results of this scenario indicate that all the ISC procedures under evaluation appear to deal adequately with distributions with good levels of compliance, i.e.*, by making the 'fail' decision a statistically very unlikely occurrence*.

Scenario 2: Standard "pessimistic", no additional outliers

For this scenario, pass rates drop from the results of scenario 1, as expected from the increased percentage of non-compliant results in the distribution, now standing at ~35%. "Correct" decision rates also drop, ranging from approximately 72% (for *ISO HDV*) to 91% (for *ISO Reg. 83*). The *ISO HDV* approach has the lowest pass rate at ~57%. Looking at this result, and considering that the population's "compliant" rate is about ~65%, it appears that *the ISO HDV procedure is markedly more stringent than the others*. This is further indicated by the imbalance between type I and type II errors for this procedure, which shows a clear bias towards 'unfair' results for the manufacturer (~27%, vs. just ~1% of 'unfair' results for the environment). The *ISO ICCT* approach also shows a similar behaviour under this scenario, but its results are less extreme (which is likely a consequence of its more balanced distribution of risks, as discussed in section 2.3).

The *ISO ACEA* procedure achieves slightly better "correct" decision rates under this scenario than *e.g. ISO COM* or *NL margin* (~87% vs ~84-85%), but it does so at the expense of a high imbalance against the environment in the share of type I and type II errors (~3% "unfair" to the manufacturer vs. ~10% of "unfair" to the environment). Finally, the testing burdens range from ~4.3 vehicles of the *ISO ICCT* procedure to ~6 vehicles of *ISO Reg. 83*.

The results of this scenario indicate that a redistribution of risks (in the case of ISO procedures) has a significant effect upon the "correct" decision rates. Imbalances in the customer/manufacturer risk balance will manifest themselves in 'fail' rates higher than the 'compliant' rate of the input population, and in a bias in the distribution of type I and type II errors. A trade-off between a high rate of "correct" decisions and increased testing burden is also observed.

Scenario 3: Compliant population with outliers (10% noncompliant rate with x5 multiplier)

In this scenario, the 'baseline' population has overall better compliance characteristics than the population of scenario 1, but a 10% of outliers with a multiplier of 5 is added to evaluate the response of the ISC procedures to outliers in isolation. These outliers do not push the average emissions of the distribution above the NTE limit, but they increase average emissions by ~60%.

For this scenario, "correct" decision rates lie consistently around ~76-79% for all procedures, except for *ISO HDV*, which has a significantly higher "correct" decision rate (of around ~84%), but at the expense of a substantially higher testing burden (~4.5 vehicles vs. 3-3.5 vehicles for the other procedures).

The 'pass' rates of the *ISO COM, ISO ACEA and ISO ICCT* procedures for scenario 3 lie around ~95%, but they drop significantly (to ~73%) for both *NL margin* and *ISO COM (custom outlier treatment)*. These differences highlight the increased ability of the *NL margin* to detect and qualify outliers; since the 'pass' and 'fail' thresholds for this procedure are dynamically adjusted on the basis of the variability of the results in the sample, the likelihood of a 'fail' outcome is increased by the presence of one or more outliers. The *ISO COM (custom outlier treatment)* procedure also shows improved outlier detection capabilities for this scenario with respect to the *ISO COM* procedure, yielding results that are similar to those of *NL margin*. Moreover, both the *NL margin* and the *ISO COM (custom outlier treatment)* achieved a good balance between type I and type II errors for this scenario.

Scenario 4: Compliant population with outliers (5% noncompliant rate with x10 multiplier)

In this scenario, the 'baseline' population has overall better compliance characteristics than the population of scenario 1, but a 5% of outliers with a (high) multiplier of 10 is added to evaluate the response of the ISC procedures to outliers in isolation. As in scenario 3, these outliers do not push the average emissions of the distribution above the NTE limit; they do however increase average emissions by ~88%.

For this scenario, "correct" decision rates lie consistently around ~64% for procedures *ISO COM, ISO HDV, ISO ACEA* and *ISO ICCT*. For *ISO HDV*, the average testing burden was highest, with an average number of ~4.3 vehicles tested at the time of decision vs. ~3.1 vehicles at most for the other procedures). For *ISO COM (custom outlier treatment)* and *NL margin*, "correct" decision rates are somewhat improved (to ~75%) with no apparent trade-offs with testing burden.

The 'pass' rates of the *ISO COM, ISO ACEA* and *ISO ICCT* procedures for scenario 4 lie around ~98%, but they drop somewhat for NL margin and for *ISO COM (custom outlier treatment)* (to ~85%). The outlier detection power is somewhat diminished for this scenario with respect to scenario 3 ('fail' rates do not increase, in spite of the fact that the additional outliers have in this case a larger impact upon average emissions). Moreover, both *NL margin* and *ISO COM (custom outlier treatment)* no longer achieve a good balance between type I and type II errors for this scenario: the 'incorrect' decisions appear consistently biased against the environment.

The results of this scenario—and also those of scenario 3—show the importance of an adequate treatment of outliers. The NL margin *procedure has a built-in capability to qualify the outliers and dynamically adjust the 'pass' and 'fail' thresholds accordingly. The* ISO COM (custom outlier treatment) *procedure can improve the outlier detection capabilities of the* ISO COM *procedure by modifying the outlier treatment parameters (modifying the 'intermediate outlier' threshold, and reducing the number of allowed 'extreme outliers' from two to one). These improvements in the treatment of outliers have no apparent trade-offs with testing burden.*

Scenario 5: ACEA compliant

In this scenario, the population has ~24% non-compliance rate and average emissions of ~0.75 times the NTE limit. Approximately 5% of the population would qualify as an 'intermediate' outlier under current ISC provisions, and there would not be a significant proportion of 'extreme outliers' (<0.1%). In the simulations, *ISO Reg. 83, ISO ACEA* and *ISO COM* achieve the highest rates of "correct" decision (~96%, ~95% and ~94%, respectively). On the other hand, *ISO HDV, NL margin, ISO COM (custom outlier treatment)* and *ISO ICCT* achieve somewhat lower rates of "correct" decision (~83%, ~88%, ~89% and ~89%). The 'pass' rates are also lower for the four aforementioned procedures (~80%, ~87%,~88% and ~90%); compare this number to a 'compliant' rate of ~76%).

The balance between type I and type II errors is consistently biased against the manufacturer for this scenario—especially for *ISO HDV* and *NL margin*, but significantly less so for *ISO ACEA* and *ISO COM*. The testing burden is lowest for *ISO ICCT*, and highest for *ISO HDV* (with an average testing burden of

>6 vehicles tested at the time of decision). Interestingly, the testing burden for the NL margin approach was especially high for individual ISC runs with a 'fail' outcome (average vehicles tested at the time of decision: ~7; that is roughly double the average for runs with a 'pass' outcome) [this result (particular to the 'NL margin' procedure can be inspected in the "NL margin" tab of the worksheet].

The results of this scenario show once again that the ISO HDV *procedure leads to a significant increase in the testing burdens for populations with good or moderate compliance levels. They also show the previously observed trade-off between a high rate of 'correct' decisions and the testing burden. The results would also provide some indication that, for certain distributions where the compliance profile is not clear-cut (*e.g.*, where the average of the population is close to the NTE limit, and a significant proportion of it is moderately above the NTE limit), the* NL margin *procedure could lead to an increase in the testing burden before a* 'fail' outcome*.*

Scenario 6: ACEA non-compliant

In the sixth and final scenario, the population has ~74% non-compliance rate and average emissions of ~1.78 times the NTE limit. This is consistent with the intention to model a population with a clearly poor compliance profile. The results for this scenario show consistently low 'pass' rates (ranging from ~1% to ~7%) coupled with consistently high 'correct' decision rates (ranging from ~93% to ~99%).

The results of this scenario indicate that all the ISC procedures under evaluation deal adequately with distributions with clearly inadequate levels of compliance, i.e.*, by making the 'pass' decision a statistically very unlikely occurrence*.

5 Conclusions

A simulation approach was used to evaluate six different procedures for the statistical treatment of the results of emissions in-service conformity testing with RDE. The current approach for ISC testing was also included in the simulation.

All the alternative statistical procedures that were proposed were in general suitable for the application to RDE tests, and they brought about improvements from the current procedure in key aspects such the re-balancing of the manufacturer/consumer risks, the overall testing burden, or the treatment of outliers. In general, however, not all improvements could be fully realised at the same time. Most notably, a trade-off between the testing burden and the accuracy (*i.e.*, the share of 'correct' decisions in the simulations) was consistently observed.

A comparison of the pass-fail charts and the operating characteristic curves of the ISO procedures (section 2.3) showed that all three 'new' ISO procedures (*ISO ACEA, ISO COM* and *ISO ICCT*) achieved a re-balancing of the manufacturer and consumer/environmental risks. Of these three, ISO COM delivered the most satisfactory compromise by keeping a perfect balance between manufacturer and consumer/environmental risk up to the maximum sample size of 10 vehicles. **Of all the "ISO" procedures, *ISO COM* is therefore the preferred option.**

Because the *NL margin* procedure has a built-in treatment of outliers and is fundamentally different from the "ISO" approaches, it is not possible to meaningfully compare it to them on the basis of operating characteristic curves (Figure 3). However, the simulation approach used in our evaluation allows the simultaneous comparison of several ISC procedures on the basis of the same input and using the same evaluation metrics, and it crucially allows an evaluation of the differences in testing burdens and treatment of outliers among ISC procedures for a series of scenarios proposed to that avail (and discussed in section 2.2). On the basis of this evaluation, we reach the following conclusions:

- All of the procedures achieve satisfactory results for populations with clear-cut compliance characteristics (*i.e.*, populations with high levels of non-compliance will consistently achieve low 'pass' rates, and conversely populations with high compliance levels will consistently achieve very low 'fail' rates).

- The *NL margin* procedure can achieve generally low testing burdens. It also has very good outlier detection power, and an adequate balance between manufacturer and consumer/environmental risk as evaluated with the criteria described in section 3.2. The main shortcoming of this procedure is that it represents a strong conceptual depart from the known “ISO” framework, and to the lack of objective data to establish an appropriate margin (which may have to be different for RDE and WLTP tests).

- The power to detect outliers of "ISO" approaches can be increased by tuning the outlier treatment parameters. This was done for the *ISO COM (custom outlier treatment)* approach, in which the threshold for intermediate outliers is lowered from 1.5 to 1.3, and the threshold for extreme outliers is kept at 2.5, but it takes only one extreme outlier before the sample is failed. These modifications reinforce the early detection of outliers and encourage sound engineering practices for vehicle aftertreatment design regarding durability and real-world performance, therefore delivering the desired environmental benefit. Furthermore, this approach can in principle be used for WLTP tests without further adaptation.

In light of all of this, **we put forward *ISO COM (custom outlier treatment)* as the EC's proposal** for a statistical procedure to be used in the evaluation of emissions in-service conformity tests under the real-driving emissions (RDE) regulation, as well as for WLTP tests.

Acknowledgements

We thank Norbert Ligterink (TNO) for proposing the spreadsheet simulation as a way to perform the evaluation of the ISC procedures, and producing a proof of concept of the spreadsheet. We gratefully acknowledge the support of the group of ACEA experts led by Elsa Malki (Ford Motor Europe) in improving and correcting errors in earlier versions of the spreadsheet. Last but not least, we thank Yoann Bernard and Uwe Tietge (ICCT Europe) for producing the operating characteristic curves for the "ISO" procedures.

References

[1] Regulation No. 83. Uniform provisions concerning the approval of vehicles with regard to the emission of pollutants according to engine fuel requirements. In: Addendum 82: Regulation No. 83, Revision 4. UNECE. United Nations Economic Commission for Europe, Geneva, Switzerland.

[2] Regulation (EC) No 595/2009 of the European Parliament and of the Council of 18 June 2009 on type-approval of motor vehicles and engines with respect to emissions from heavy duty vehicles (Euro VI) and on access to vehicle repair and maintenance information and amending Regulation (EC) No 715/2007 and Directive 2007/46/EC and repealing Directives 80/1269/EEC, 2005/55/EC and 2005/78/EC. Official Journal of the European Union. *http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32009R0595*

[3] Commission Regulation (EU) 2016/427 of 10 March 2016 amending Regulation (EC) No 692/2008 as regards emissions from light passenger and commercial vehicles (Euro 6). Official Journal of the European Union. *http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0427*

[4] International standard ISO 8422. Sequential sampling plans for inspection by attributes. Second edition 2006-10-01. International Organisation for Standardisation, Geneva, Switzerland.

[5] Regulation (EC) No 715/2007 of the European Parliament and of the Council of 20 June 2007 on type approval of motor vehicles with respect to emissions from light passenger and commercial vehicles (Euro 5 and Euro 6) and on access to vehicle repair and maintenance information. Official Journal of the European Union. *http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32012R0064*

[6] NL statistics: using average and spread for a well-informed decision. Presentation to the RDE-LDV technical working group delivered in May 2017. *https://circabc.europa.eu/d/d/workspace/SpacesStore/07bb5240-9895-49ea-bcbf-933d05cc1479/CoPversusISC.pdf*