Response to Puget Sound Clean Air Agency "Preliminary Review of Analysis of NSPS Test Method Variability (Curkeet, 2010)" (Dr. Phil Swartzendruber, 2012)

Rick Curkeet, PE

The Puget Sound Clean Air Agency issued a letter dated December 5, 2012 to Stephan D. Page, Director of EPA Office of Air Quality Planning and Standards. Attached to the letter was an analysis and critique apparently by Dr. Phil Swartzendruber of our paper on the variability of the EPA Wood Stove NSPS test method based on the proficiency test program conducted by EPA from 1987 through 2005.

Our analysis was conducted following the procedure specified in ASTM E691-09, "Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method." Dr. Swartzendruber critique appears to substantially disagree with the process specified in ASTM E691. Thus, the critique is primarily a disagreement with the procedures outlined in the ASTM Standard Practice and not directly a critique of our analysis and conclusions.

ASTM E691 is the established consensus procedure for evaluation test method measurement precision on the basis of Interlaboratory Studies. It is commonly applied by ASTM committees and other organizations for this purpose. Since it is ASTM policy to include precision and bias information in measurement standards, this procedure is applied very broadly. As we discussed in our paper, the database obtained through the EPA Wood Stove Emissions proficiency test program is far from ideal, but it is the only such database that exists and is certainly adequate to indicate that variability in test results is a very significant issue.

The Puget Sound Clean Air Agency letter and attached critique of our variability analysis are clearly an attempt to discount the substantial problem that a large variability in wood burning appliance emissions tests presents in the development of workable and useful regulatory limits. Every quantitative measurement process that does not include an understanding of potential variability in results leaves the validity of the measurements in doubt. We stand by our conclusions that the EPA NSPS wood stove emissions measurement process has a large variability relative to the specified limits. None of the individual "certified" emissions rates can therefore be considered a definitive or reliable measure of the appliance's true emissions performance. We also stand by our conclusion that the variability in results observed is due primarily to actual difference in the product's performance from test to test and not to significant error in the measurements themselves.

Dr. Swartzendruber makes three basic charges regarding our analysis.

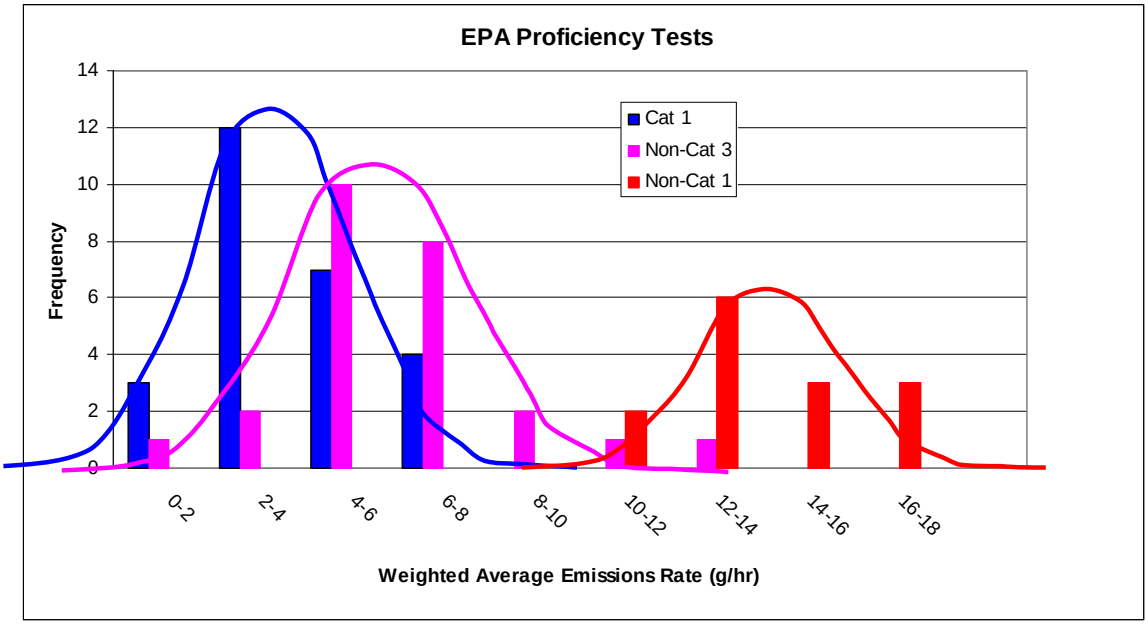
1. The data is not "normally distributed".
2. We improperly divided up the underlying data.
3. We did not make an adequate comparison to the EPA certified values.

We find all three of the criticisms to be baseless. First and foremost, as pointed out already, he ignores the fact that our analysis was conducted in accordance with the controlling ASTM method for test method precision evaluations; his criticisms ignore this fundamental principle, and should be ignored for that reason alone. Nevertheless, we respond below to each of these concerns individually, and follow that discussion with additional comments on other issues.

A. HIS THREE BASIC CONCERNS:

1. The data does appear to be normally distributed for each of the 3 stoves analyzed. Since it can be presumed that each stove involved would be likely to have its own emissions test result distribution (i.e. mean and standard deviation) it is appropriate to view the data distribution on a stove-by-stove basis. The following histogram shows that the data fits a normal distribution quite well. In addition, standard statistical tests for normality do not reject the hypothesis that the data comes from a normal distribution at a 0.05 significance level.

Figure 1.



We accept Dr. Swartzendruber's criticism regarding the use of a 2.8 standard deviation multiplier when discussing deviation from the mean at a 95% confidence. We should have use a 1.96 factor in this discussion. However, we stand by the use of the 2.8 factor for evaluation of the 95% confidence for the potential difference between 2 individual test results. Our conclusion was based on this factor as it is the value specified in ASTM E691. Even Dr. Swartzendruber's analysis arrives at a standard deviation of 1.8 g/h which translates to a reproducibility of 5 g/hr (1.8 x 2.8) which is roughly the same as our conclusion.

2. The data was analyzed on a stove-by-stove basis which is completely in keeping with the procedures specified in ASTM E691. It is quite possible that variability in results is a function of the technology employed in specific designs. For example, we know from other experience that pellet burning appliances tend to show much more consistent results from test to test than cordwood burning stoves. The following is the discussion related to this process from the standard.

   10.2.2 An ILS of a test method should include at least three materials representing different test levels, and for development of broadly applicable precision statements, six or more materials should be included in the study.

   10.2.3 The materials involved in any one ILS should differ primarily only in the level of the property measured by the test method. When it is known, or suspected, that different classes of materials will exhibit different levels of precision when tested by the test method, consideration should be given to conducting separate interlaboratory studies for each class of material.

   The data did indicate that r & R are not consistent from one appliance to another. For example the low emissions catalytic stove showed a reproducibility of 4.5 g/h while the low emissions non-cat stove (3) had a reproducibility of 6.4 g/h. The high emissions non-cat stove (1) had a slightly better reproducibility at 5.1 g/h. This simply indicates that the level of precision is somewhat variable stove-to-stove and does not show a trend of being better or worse at different levels of emissions.

3. A comparison of the proficiency data to EPA certified values has no significance in the analysis of test procedure variability. The purpose of the proficiency test program was not to verify or validate the certified test results. We are particularly concerned with the statement that our analysis "ignores the key questions, which are the accuracy of the EPA certification value, and the skill that accredited laboratories have in reproducing it." Since all the participating laboratories were EPA accredited and the point of a proficiency test program is to evaluate a laboratories skill in conduct of the test, this criticism makes no sense. In fact, no laboratory should ever conduct a test of this nature with the goal of reproducing some previously established result. This would introduce an obvious bias. Each test must be an essentially independent measurement of the property of interest to be considered a valid sample.

   Since EPA modified the stoves provided for the program, there was a presumption that the emissions performance would be affected. In fact, manufacturers who provided stoves where assured that the results would not bring their certification status into question. For 3 of the 5 stoves involved the certified emissions rates were lower than any of the individual PTP results.

   It is also inappropriate to compare the average of multiple tests to a single test result which the certified emissions rate represents. It would be just as invalid to compare the average result to any of the individual results from a stove data set selected at random. Repeatability and reproducibility describe the potential difference between two individual measurements. The certified value represents one measurement. Thus, we could say that there is only a 5% chance that a second result from another laboratory that is within 5 g/hr of the certified value indicates a real difference. Conversely, individual results that are significantly more than 5 g/hr from the certified value indicate that the stove is not likely the same as the original certification test unit. Of the five stoves involved in the EPA PTP program only one had all new results within 5 g/h of the certified values. Two had no PTP results within 5 g/h of the certified emissions rate.

B. OTHER COMMENTS:

   Dr. Swartzendruber's paper includes the following paragraph.

"The combination of these three incorrect analytic steps: applying normal statistics to non-normal data, dividing up the underlying data and implying a subset could represent the population, and using a difference statistic (absolute difference) and conflating it with a confidence interval, produces an uncertainty estimate which is likely unrepresentative and biased high, and disagrees with the clearly observable relationship in Figure 1, and the well behaved residuals in

Figure 2. It is likely that a value of 1.5-2 g/hour is a more accurate representation of the uncertainty of mean test results. If this value can be decreased to about 1 g/hour, then with only four repetitions, a stove that performs at a mean of 2.5 g/hour can be said to be below a standard of 3.5 g/hour with about a 95% level of confidence."

This is an absurd argument. First, he criticizes the use of normal statistics and difference statistics to evaluate precision when that is precisely how precision is defined in ASTM E691 and E177. He then refers to the uncertainty of the mean as 1.5 to 2 g/h when the "certified" emissions rates in question are not determined from the mean of multiple tests. He then assumes that somehow this uncertainly will be "*decreased to about 1 g/h"* but provides no proposal as to how this could be accomplished. Finally, he appears to propose that each stove be tested through the process four times ("*only four repetitio*ns"). This implies a four fold increase in testing costs which is unrealistic as the cost of testing is already very high. In addition, the existing EPA certified stove database does not provide means of 4 repetitions and is thus subject to the high uncertainly related to single test results. We, of course, did not argue that the proficiency test results were based on a representative or statistically significant sample of the certified stoves in the EPA program. If they are, one would have to conclude that there would be a high probability that many of the certified stoves would not comply with the EPA NSPS regulatory limits if they were each tested multiple times and the mean of these tests was used to determine compliance.

There is reason to believe that the certified stove emissions ratings are biased toward low values versus results obtained from the proficiency test process. Certification test results typically are based on very specific start-up procedures provided by the manufacturer. In certification tests labs may have imposed significantly tighter controls of fuel selection and arrangement than required by the standard and additional tests may have been conducted to eliminate high emissions results under the "outlier" provision of the EPA procedure. Many of these procedures were not a factor in the proficiency tests.

PRECISION AND COMPLIANCE DETERMINATIONS

In virtually all areas of compliance testing and certification it is very important to know both the applicable compliance limit and the potential variability of results. It is well recognized that when compliance specifications are created they must be based on measurements with sufficient precision to reliably distinguish between compliant and non-compliant products. In general, the precision relationship to the compliance limit is evaluated by the P/T ratio (Precision to Tolerance). In most cases a P/T ratio of 0.1 is desired, but P/T ratios up to 0.25 are

considered acceptable when the consequences of non-compliant products being accepted or compliant products being rejected are not too costly or harmful. P/T ratios above 0.3 are generally considered unacceptable due to the high probability that compliant products will be rejected and non-compliant products will be accepted.

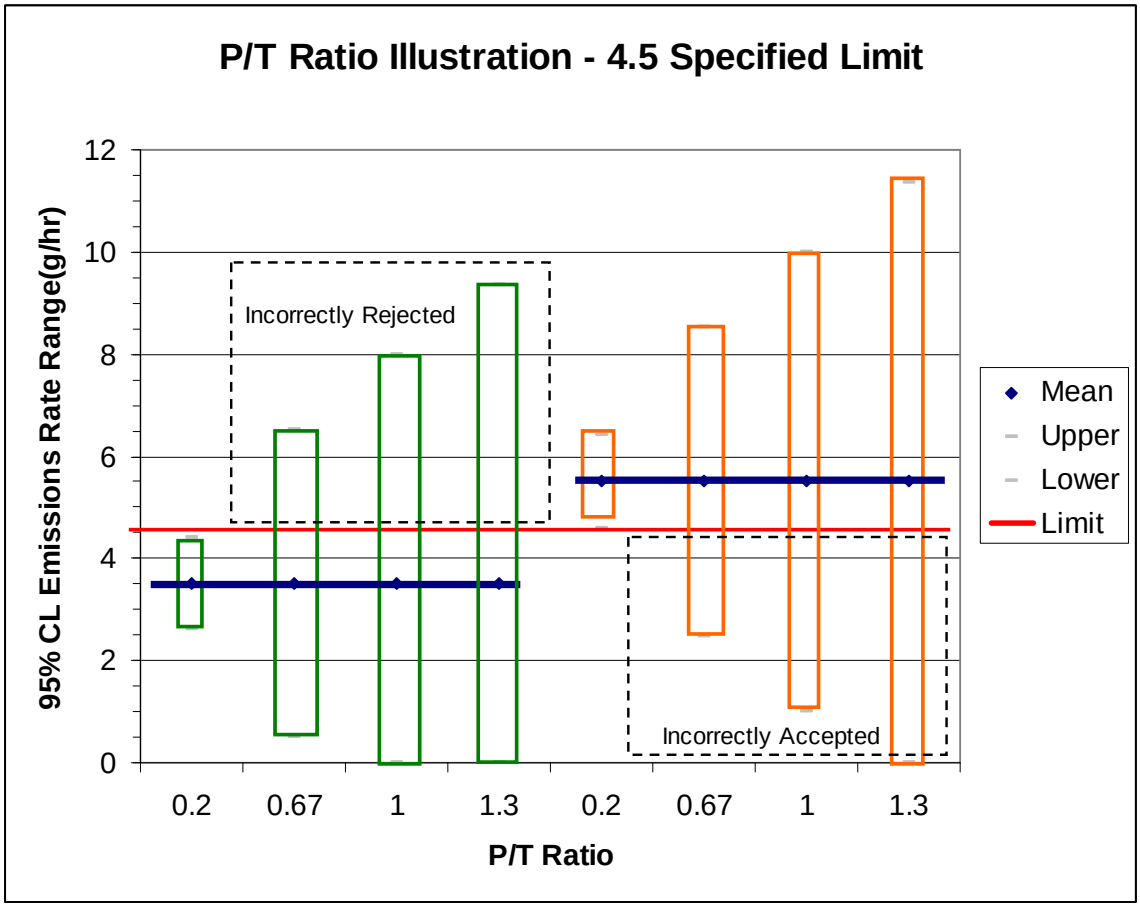
As an example of this concern, consider a specification for the diameter of a machined part of 2" ± 0.005". If we attempt to measure the diameter with an instrument with an uncertainty of ± 0.01" (P/T ratio = 2), we can never be sure if the part meets the specification or not. In fact, we need an instrument with a precision of at least 0.001" (P/T ratio = 0.2) to have a reasonable chance of identify compliant and non-compliant parts and we will still have a significant potential to improperly classify parts that have a true diameter of 1.994" to1.996" or 2.004" to 2.006".

Given that the data indicate a precision of at least 3.5 g/hr for wood stove emissions, a compliance limit in the range of 2 to 4.5 g/hr results in a P/T ratio of 1.3 to 0.8. The compliance limit would need to be at least 11 g/hr to achieve even a barely acceptable 0.3 P/T ratio.

In the following illustration the effect of various P/T ratios are shown for two base cases. Both cases assume that an emissions limit is set at 4.5 g/hr. The green boxes (left half of chart) show the 95% probability range for an appliance with a true mean emissions rate of 3.5 g/hr. The orange boxes (right half of chart) show the 95% ranges for an appliance with a true mean emissions rate of 5.5 g/hr.

It can be easily seen that higher P/T ratios result in substantial probabilities that appliances will be incorrectly accepted and rejected. It can also be seen that if the P/T ratio is high the manufacturer cannot design an appliance with a low enough mean emissions rate to approach a high confidence of passing in a single test series. Nor can the regulator specify a limit low enough to assure that a passing result will not occur even if the unit's mean emissions rate is substantially higher than anticipated.

Only a reduction in the P/T ratio can produce a system that provides a reasonable probability of accepting and rejecting products as intended. To do this one must find a way to either reduce the inherent variability of the process or one must set limits that are high enough to account for the effects of the variability.



Equivalent Precision Values

P/T Ratio	Precision (±g/hr)
0.2	0.9
0.67	3.0
1	4.5
1.3	5.85

It should be emphasized that precision is not the same as accuracy. Accuracy is defined as the difference between the measured result and a ***known*** true value. There is no known true value for wood burning appliance emissions rates. There is absolutely no reason to think that any wood burning appliance would actually produce the same quantity of emissions each time it is tested. Indeed the proficiency test data shows us that the production of emissions when the test method is followed is highly variable. We did analyze the measurement uncertainty of the process and concluded that it is small relative to the dispersion in the data. But, it is improper to use the term “error” to describe the differences in observed test results. Error is defined as the difference between the

measured result and the true value and thus cannot be quantified when the true value is unknown.

Returning to the target analogy used in our paper, the current situation is rather like shooting at a blank piece of paper and declaring that wherever the bullet hits is the bull's-eye. However, if we take multiple shots and end up with a wide dispersion, it becomes clear that we cannot claim to know where the bull's-eye is and certainly not to have hit it.

There is no simple way to resolve this concern as EPA considers revised emissions limits for wood burning appliances. There are certainly ways of reducing the effective variability such as conducting multiple tests and evaluating performance based on the average of results. It is also possible, and perhaps likely, that newer test methods and product designs would exhibit less dispersion in results. However, any such revisions to the process will create a discontinuity with the existing database and this would require an extensive research project to obtain data sufficient to determine what emissions level would represent BDT. If the proficiency test program data is at all indicative of a general relationship, one would expect that all emissions ratings would tend to increase by approximately 25 to 100% under a method that would reduce the variability in results.

Stove	Certified	PTP Mean	% Increase
Cat 1	3.1	3.9	26%
Non-Cat 1	7.5	14.32	91%
Non-Cat 2	4.5	8.6	91%
Non-Cat 3	3.6	6.74	87%
Non-Cat 4	3.1	12.96	318%

It is not scientifically defensible to look at the existing population of EPA certified products within a category (e.g. cordwood burning stoves) and conclude that some particular subset of products (e.g. Catalytic Stoves) represents superior technology to the rest without considering the variability issue. It is clearly inappropriate to claim that a single emissions test produces a result that represents some true value within less than about ± 4 grams/hour. Thus, it is incorrect to conclude that a result of 3 grams/hour means that a specific appliance is substantially better than one that gets an emissions rating of 7grams/hour. Another test of each product might well reverse these results or show no actual difference. Without proper consideration of the inherent variability in the process, there is a high risk of making incorrect conclusions. It is a basic principle of science that data that cannot be reproduced is simply not valid.