Attachment 2

State of California
AIR RESOURCES BOARD

Supplemental Analysis Supporting the Test for Demonstrating Equivalence between Primary and Secondary Methods for Measuring Formaldehyde Emissions from Composite Wood Products

January 2008

Background

On April 26, 2007, the Air Resources Board (ARB) adopted a new airborne toxic control measure (ATCM) to reduce formaldehyde emissions from composite wood products. The regulation specifies technology-forcing limits for formaldehyde emissions from composite wood products. The limits vary by product type, and will be reduced in two phases between 2009 and 2012.

The regulation requires manufacturers to certify their products before they are shipped to ensure they meet the emission limits. Manufacturers must have their routine formaldehyde emission testing verified by a third party certifier (TPC). The primary method used by third party certifiers for testing formaldehyde emissions from composite wood products is the so-called large chamber test, which requires a test chamber of at least 22 m^3 (22,000 liters). As part of the rulemaking process, staff received comments from several parties sharing concern about the lack of large chamber testing capabilities, both domestically and in foreign countries. Staff evaluated the concern and concluded that testing flexibility is warranted. However, staff also believes that adding flexibility should not come at the expense of reducing the accuracy, precision and integrity of the third party certification program.

In order to allow flexibility in certifying products, the regulation also allows a secondary method to be used. The secondary method is essentially a scaled down, "bench top" version of the large chamber test. In order to ensure that using the secondary method does not compromise the certification process, the regulation requires that manufacturers and TPCs wishing to use the secondary method demonstrate that their implementation of the secondary method yields results that are equivalent to the primary method. The regulation specifies the test for demonstrating equivalence in detail. We believe the secondary method is easier and less costly to implement by allowing chambers as small as 20 liters to be used.

This document discusses the rationale used by ARB staff to develop a statistical test for demonstrating equivalence between the primary and secondary methods. It discusses the statistical performance of the test on realistic simulated data sets.

This document is intended for technical readers with an understanding of statistics.

Statistical theory and methodology

Introduction: why not use the Student *t* test?

Before describing the test used in the regulation, we begin by discussing another type of test which was *not* used: the conventional one-sample Student *t* test encountered in elementary statistical textbooks. This subject is of interest partly because in preliminary discussions several participants asked why the Student *t* test was not used, and partly because it helps motivate the discussion that follows.

In the classic one-sample Student *t* test we seek to determine whether the mean of a random variable is different from zero. In the context of comparing large versus small chamber methods, the variable would be the bias, or mean difference between paired samples by both tests. We form the statistic

$$T = \frac{\overline{X}}{S / \sqrt{N}}$$

where $\overline{X}$ is the mean difference between paired samples, N the sample size, and S the standard deviation of the differences. We compare T with a percentile of the Student t distribution with $N - 1$ degrees of freedom and deem the bias different from zero when t exceeds the percentile.

The choice of percentile controls the probability of judging the methods *not* equivalent when they actually are equivalent; that is, judging the bias different from zero when it is actually zero. This probability is the false failure rate.

The test does not explicitly control the probability of judging the methods equivalent when they are actually not equivalent, or false pass rate.

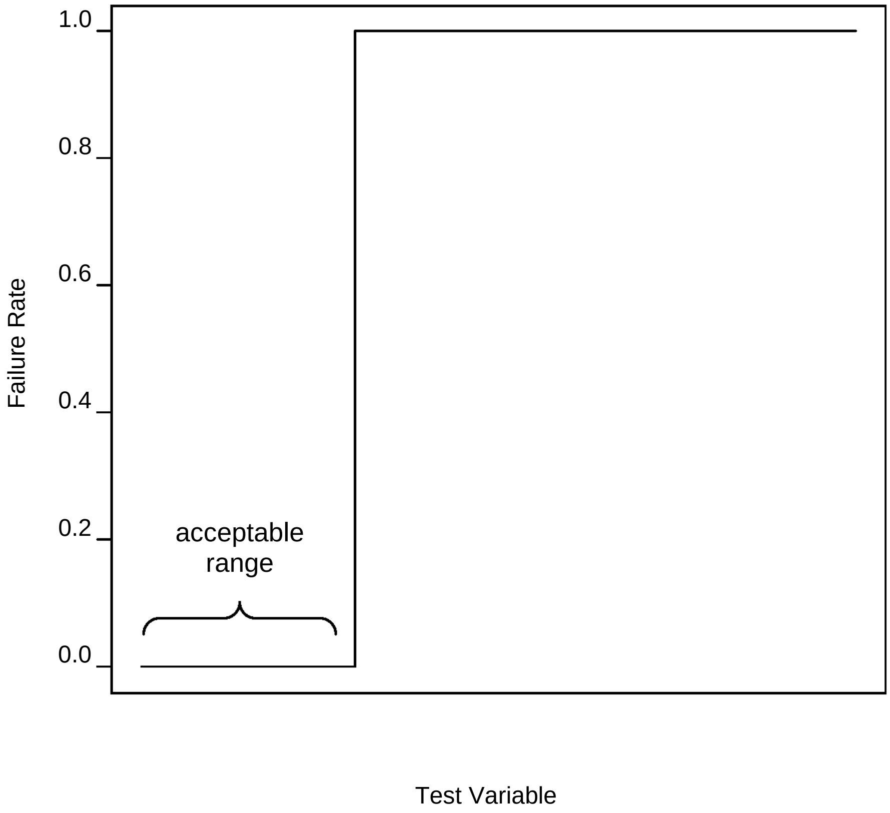
The one-sample Student *t* test is not suitable for proving equivalence between methods for several reasons:

- For an equivalence test to be meaningful, applicants must demonstrate equivalence with a high degree of probability (low false pass rate). The Student *t* test is designed to control the false failure rate, not the false pass rate. With the low sample sizes encountered in practice, its false pass rate may be unacceptably high.
- A well-designed test should reward the applicant for being more precise. The Student *t* test does the opposite: the smaller S is, the larger T becomes, and the greater the probability of failing the test.
- A well-designed test should reward the applicant for collecting more samples. Again, the Student *t* test does the opposite: the larger N is, the larger T becomes, and the greater the probability of failing the test.

How an ideal test behaves

An ideal equivalence test has false failure rate and false pass rate equal to zero. An example is shown below.

Behavior of an ideal equivalence test



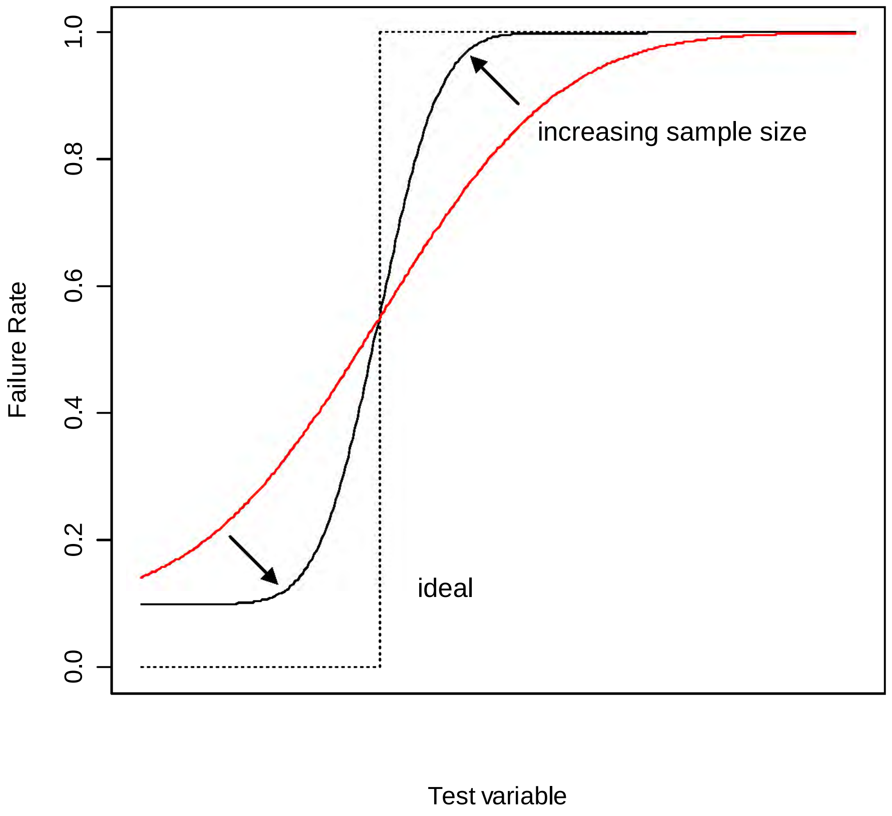
The failure rate is the probability of a sample failing the test. The "test variable" could refer to the bias, the standard deviation, or a test statistic constructed from both of these. For an ideal test, the variable passes the test with probability 1 when it is in the acceptable range, and fails with probability 1 when it is not.

How should an acceptable range be chosen? For bias, ideally the acceptable range would consist of a point at zero. Of course, in reality, the bias between two methods will always be nonzero, but one can require that it be small relative to the precision of the methods. For precision, we can deduce a reasonable value from interlaboratory studies, ASTM method repeatability, etc.

How realistic tests behave

More realistic test behavior is illustrated by the chart below.

Comparison of a realistic equivalence test with an ideal test



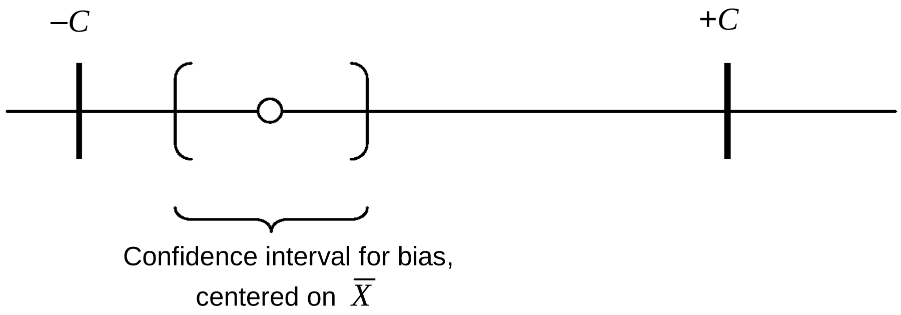
Realistic tests differ from the ideal in that the cutoff is not sharp; sometimes the test shows equivalence when the test variable is not in the acceptable range, and vice versa. However, we would like the test behavior to approach the ideal behavior as sample size increases, as indicated by the arrows.

Choosing the form of the test

Several possible forms for the test were considered. All were of the general form

$$a_1|\overline{X}| + a_2 S \leq C$$

where $\overline{X}$ is the mean difference between methods, S the standard deviation of the differences, a_1 and a_2 constants or functions of the sample size, and C a constant. This form is motivated by the following reasoning. Suppose an applicant is required to demonstrate that a confidence interval for the bias of the secondary method falls entirely within specified limits, ±*C*, as illustrated below.



The confidence interval will take the form $\overline{X} \pm pS$, where the expression p involves a percentile of a normal or Student t distribution, and perhaps a function of the sample size N. Requiring that the confidence interval fall within ±*C* leads to the criterion $|\overline{X}| \pm pS$. The more general form above allows us to adjust the test statistic so that it better approximates the ideal behavior.

Statistics of this form

- Can be designed to control the false pass rate as well as the false failure rate
- Reward the applicant for smaller bias *and* better precision by increasing the probability of passing the test
- Reward the applicant for collecting more samples by increasing the probability of passing the test

The following variants were considered:

Normal version $$|\overline{X}| + \frac{Z_\alpha}{\sqrt{N}} S$$

where Z_α is a percentile of the normal distribution, Student t version $\quad \left|\overline{X}\right| + \frac{T_{N-1,\alpha}}{\sqrt{N}} S$

where $T_{N\text{-}1,\alpha}$ is a percentile of the Student t distribution with $N - 1$ degrees of freedom, and

Constant coefficient version $\quad a_1\left|\overline{X}\right| + a_2 S$

where a_1 and a_2 are constants.

Since closed form computations with the test statistics given above can be difficult if not impossible, staff assessed their performance using Monte Carlo simulations. Random data sets were generated using normal error distributions, choosing typical standard deviations based on sample data sets, as discussed in the following section. For each version of the test, the limit C was selected to yield a failure rate of 0.10 for a sample size of 5, assuming zero bias.

Data analysis and performance assessment

Realistic values for bias and standard deviation

Staff received two sets of test data comparing the primary and secondary methods. These data sets were used to estimate the bias and precision at various ranges of formaldehyde concentration. The data sets are labeled A and B to keep the names of the data suppliers confidential.

The regulation requires that measurements given directly by the respective methods be used to compare the methods. The secondary method may not be calibrated against the primary method.

According to the regulation, the average of three replicate samples by the secondary method is to be compared against a single sample by the primary method. Data set A included 4-6 replicates by the secondary method for each primary method sample. An estimate of the sample standard deviation S_{12} of the differences between a primary method sample and the average of the three corresponding secondary method samples was obtained by resampling. In addition, a separate estimate of the precision standard deviation S_2 of the secondary method was computed by pooling the standard deviations of replicate samples. The value $S_2 / \sqrt{3}$ may be taken as an independent estimate of a lower bound on S_{12}. The following table summarizes the results in units of parts per million (ppm).

Summary of data set A

Range	**Number of primary method samples**	**Number of secondary method samples**	$\bar{X}$	$S_2/\sqrt{3}$	S_{12}
Low 0 – 0.07	2	10	+0.012	0.013	0.012
Mid 0.07 – 0.15	13	72	–0.006	0.014	0.024
High 0.15 – 0.25	7	40	–0.008	0.012	0.022

The value S_{12} = 0.012 for the low range is very unreliable, because it is based on only two primary method samples. The value $S_2 / \sqrt{3} = 0.013$, based on ten secondary method samples, is a more realistic lower bound.

Data set B only included a single secondary method sample for each primary method sample. Therefore, it was not possible to estimate S_{12} and S_2. However, the standard deviation of the differences between the primary method sample and the single secondary method sample was computed as an upper bound on S_{12}.

Summary of data set B

Range	**Number of samples**	$\overline{X}$	S_{12}
Low 0 – 0.07	10	+0.012	< 0.012
Mid 0.07 – 0.15	14	+0.011	< 0.011
High 0.15 – 0.25	0	--	--

Repeatability of the primary method, ASTM E 1333-96(2002), and secondary method, ASTM D 6007-02, indicated a precision of within:

Primary	0.03 ppm
Secondary	0.01 – 0.02 ppm

With these values in mind, a rough estimate for the standard deviation of the differences between one primary method sample and the average of three replicate secondary method samples is

$$(0.02^2 + 0.03^2 / 3)^{1/2} \approx 0.032 ,$$

slightly higher than the standard deviation for the mid and high range in data set A.

The following values were chosen as typical standard deviations for the differences:

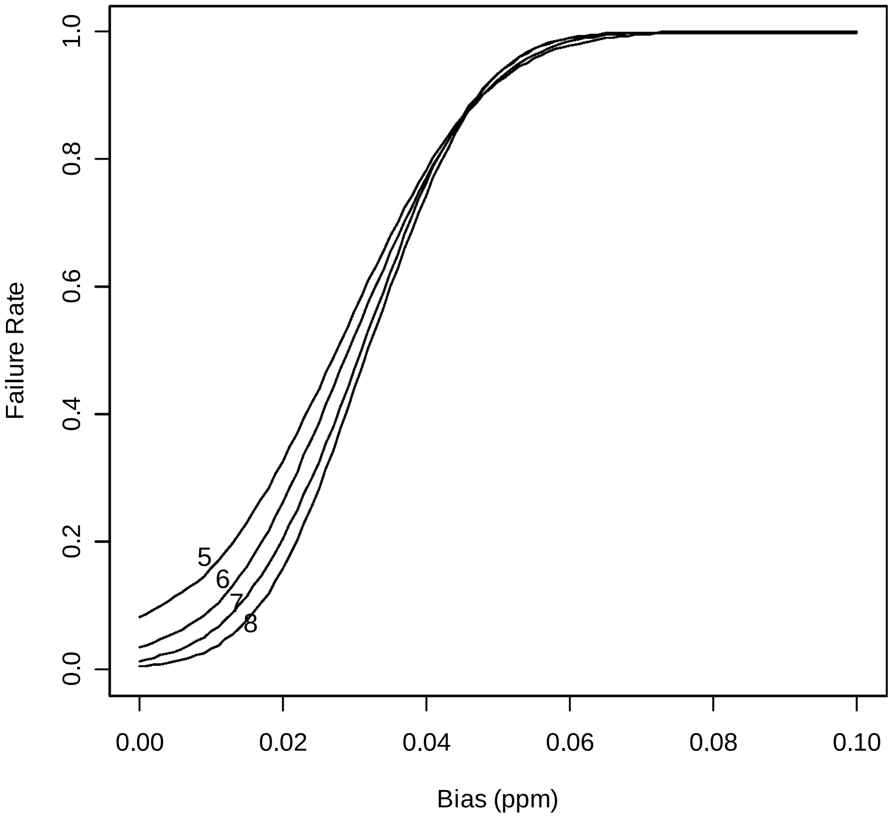
Low range	0.015
Mid range	0.022
High range	0.030

The low value is a conservative estimate based on data sets A and B. The high value represents a compromise between data set A and the value suggested by the ASTM repeatability. The mid value is halfway between the two. The standard deviations increase with concentration, which makes the test more stringent at lower concentrations. It is also consistent with the typical behavior of many analytical methods for air contaminants. These standard deviations were used in the Monte Carlo simulations to assess the performance of the different candidate versions of the test.

Simulation results: normal version

The graph below shows curves of bias versus failure rate for the normal version. The curves are based on Monte Carlo simulations, assuming a normal error distribution with a standard deviation of 0.030. Numbers indicate sample sizes. At zero bias, the failure rate decreases exponentially with increasing sample size. However, with sample sizes of 5 – 10, the failure rate curves almost overlap when the bias is high, so the failure rate at high bias increases very slowly as sample size increases, an undesirable characteristic.

Failure rate versus bias for normal version
C = 0.052, sd = 0.030

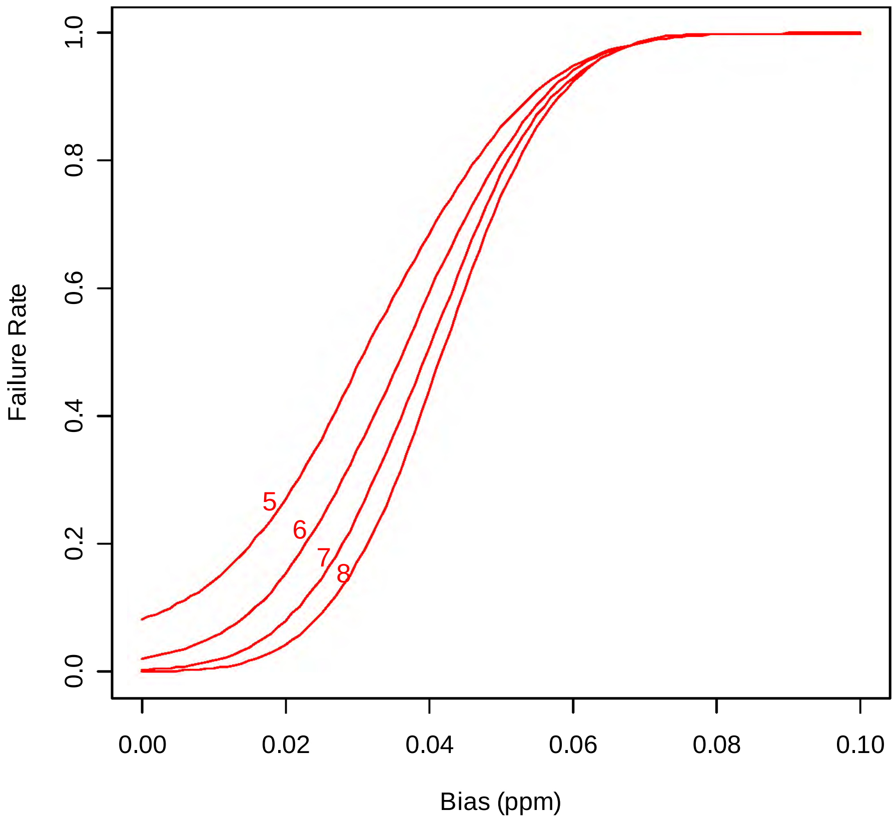


Since $\overline{X}$ and S converge to their respective population parameters the asymptotic behavior of the constant coefficient test as $N \to \infty$ is easily ascertained. The failure curves for the bias, with standard deviation held constant at 0.030, approaches the ideal step function described above, with a maximum acceptable bias of C, or 0.052. While the asymptotic value does not accurately reflect the behavior at low sample sizes, it is useful to know in that the failure curves for different sample sizes intersect at that value. Below this value, the failure rate decreases as sample size increases; above it, the failure rate increases. The failure curve for the standard deviation with the bias held constant at zero (not shown) does not converge to a step function. As N increases, the acceptable standard deviation increases without limit.

Simulation results: Student *t* version

The Student *t* version of the test displays the same characteristics as the normal version, to an even higher degree, as shown in the graph below.

Failure rate versus bias for Student *t* version
C = 0.066, sd = 0.030



The asymptotic behavior of the Student *t* version is identical with that of the normal version. As sample size increases, the failure curve for the bias converges to a step function with the step occurring at C, or 0.066, while the failure curve for the standard deviation (not shown) does not converge to a step function.

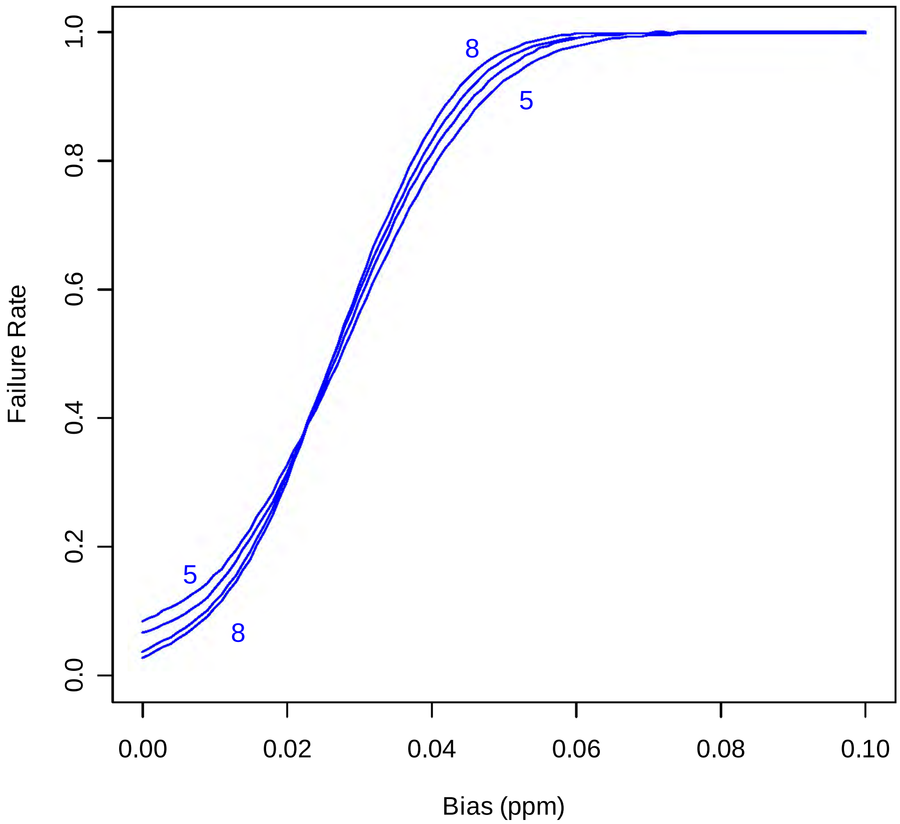
Simulation results: constant coefficient version

Coefficients were chosen to match the values of the normal version when $N = 5$, namely, $a_1 = 1$ and $a_2 = 1.96/\sqrt{5} \approx 0.88$. The C value is identical with that of the normal test, namely 0.052. Thus, the criterion for the methods to be equivalent is:

$$\left|\overline{X}\right| + 0.88\,S \leq 0.052$$

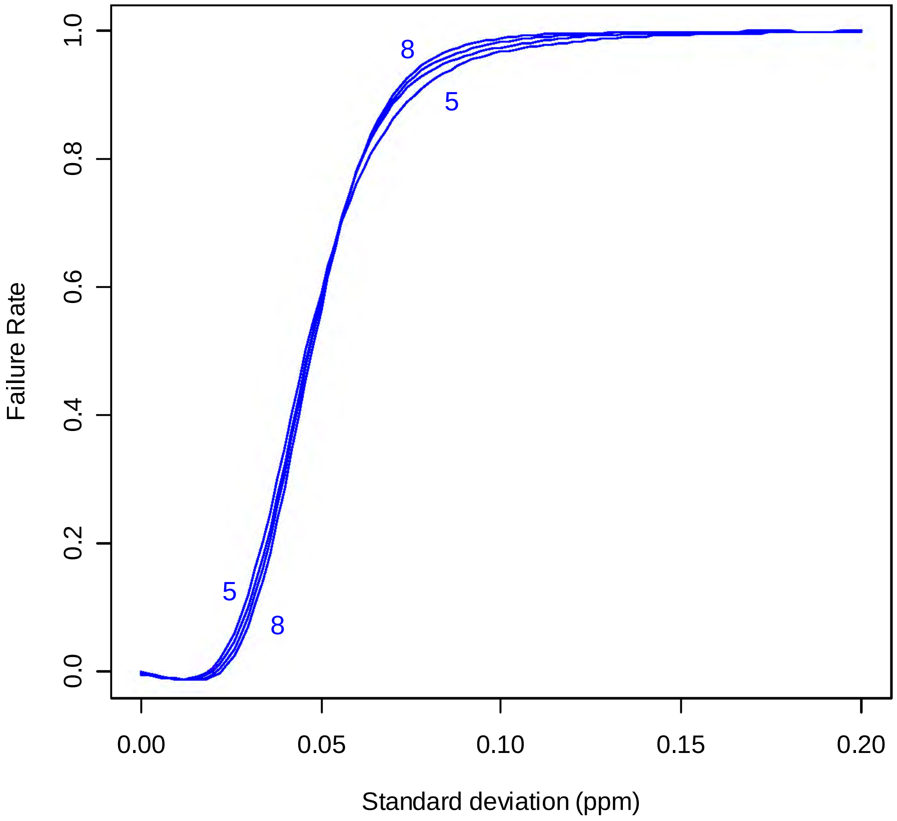
For $N = 5$ the failure rate curve matches that of the normal version. However, unlike the normal and Student *t* version, at high bias the failure rate increases steadily as sample size increases, as shown below.

Failure rate versus bias for constant coefficient version
C = 0.052, sd = 0.030



The following chart shows the failure rate as a function of standard deviation, with the bias held constant at zero, for various sample sizes

Failure rate versus standard deviation bias for constant coefficient version
C = 0.052, bias = 0



As N increases, the failure curve for the bias converges to a step function, the step occurring at a value of $(C - a_2 S)/a_1 \approx 0.026$. Since the step occurs at a lower value than the normal or Student t versions, the constant coefficient version enjoys a smaller asymptotic acceptable region than the other version, and its failure rate at high bias increases faster than the other versions.

Unlike the other versions, the failure curve for the standard deviation converges to a step function. The step occurs at $C / a_2 \approx 0.059$, roughly twice the typical value of 0.030.

Detailed performance results for the constant coefficient version

Staff chose the constant coefficient version because

- the failure rate at high bias increased faster as sample size increased than the other versions;
- the coefficients can be chosen so that for a given failure rate at zero bias, the failure rate at high bias is higher than that of the other version;
- unlike the other versions, the failure curve for the standard deviation converges to the ideal behavior as sample size increases.

The tables below summarize the constant coefficient version's performance in the three ranges, based on Monte Carlo simulations using the realistic standard deviations discussed above. Units are parts per million.

Low range

Sample Size	Failure rate at bias = 0	Bias at which failure rate = 0.95	Standard deviation at which failure rate = 0.95 when bias = 0
5	0.10	0.027	0.046
6	0.07	0.026	0.044
7	0.05	0.025	0.043
8	0.04	0.024	0.042

Standard deviation used in simulation	0.015
Value of constant C	0.026
Asymptotic acceptable bias	0.013
Asymptotic acceptable standard deviation	0.030

Mid range

Sample Size	Failure rate at bias = 0	Bias at which failure rate = 0.95	Standard deviation at which failure rate = 0.95 when bias = 0
5	0.10	0.039	0.066
6	0.07	0.037	0.063
7	0.05	0.036	0.060
8	0.04	0.035	0.058

Standard deviation used in simulation	0.022
Value of constant C	0.038
Asymptotic acceptable bias	0.019
Asymptotic acceptable standard deviation	0.043

High range

Sample Size	Failure rate at bias = 0	Bias at which failure rate = 0.95	Standard deviation at which failure rate = 0.95 when bias = 0
5	0.10	0.053	0.096
6	0.08	0.050	0.090
7	0.06	0.048	0.088
8	0.04	0.047	0.084

Standard deviation used in simulation	0.030
Value of constant C	0.052
Asymptotic acceptable bias	0.026
Asymptotic acceptable standard deviation	0.059

Conclusion

Using the constant coefficient version provides a test for equivalence between primary and secondary methods which minimizes the probability that methods with high bias or poor precision will qualify as equivalent. The following is a concise summary of the test.

- An applicant must test a minimum of five paired samples in at least two of the following formaldehyde ranges. The ranges must include the high and low ends of the concentrations over which the laboratory seeks to demonstrate equivalence:

Low	0 – 0.07 ppm
Mid	0.07 – 0.15 ppm
High	0.15 – 0.25 pmm

- Each paired sample consists of one measurement by the primary method, and the average of three samples by the secondary method. All of these samples must be on material from the same batch.

- The measurements are those given directly by the respective methods. The secondary method may not be calibrated to the primary method.

- The differences between the primary method measurement and average of secondary method measurements are computed.

- The mean $\overline{X}$ and sample standard deviation S of the differences are computed as follows:

$$\overline{X} = \sum_{i=1}^{n} D_i / n$$

$$S = \sqrt{\sum_{i=1}^{n} (D_i - \overline{X})^2 / (n - 1)}$$

- To demonstrate equivalence between primary and secondary methods, the following criterion must be met:

$$\left|\overline{X}\right| + 0.88\,S \le C$$

where C is equal to

Low range	0.026
Mid range	0.038
High range	0.052

References

ASTM Test Method D 6007-02, Standard Test Method for Determining Formaldehyde Concentrations in Air from Wood Products Using a Small Scale Chamber, 2002.

ASTM Test Method E 1333-96 (2002), Standard Test Method for Determining Formaldehyde Concentrations in Air and Emission Rates from Wood Products Using a Large Chamber, 2002.