The attached document represents CTP's then-current thinking on certain aspects of tobacco regulatory science. The information contained herein is subject to change based on advances in policy, the regulatory framework, and regulatory science, and, is not binding on FDA or the public. Moreover, this document is not a comprehensive manual for the purposes of preparing or reviewing tobacco product applications. FDA's review of tobacco product applications is based on the specific facts presented in each application, and is documented in a comprehensive body of reviews particular to each application.

Given the above, all interested persons should refer to the Federal Food, Drug, and Cosmetic Act, and its implementing regulations, as well as guidance documents and webinars prepared by FDA, for information on FDA's tobacco authorities and regulatory framework. This document does not bind FDA in its review of any tobacco product application and thus, you should not use this document as a tool, guide, or manual for the preparation of applications or submissions to FDA.

DEPARTMENT OF HEALTH & HUMAN SERVICES

Food and Drug Administration
Center for Tobacco Products
Office of Science

MEMORANDUM

FROM: Todd L. Cecil, Ph.D.
Branch Chief, Chemistry II
Division of Product Science, Office of Science

Todd L. Cecil -S
Digitally signed by Todd L. Cecil -S
DN: c=US, o=U.S. Government, ou=HHS, ou=FDA, ou=People, cn=Todd L. Cecil -S, 0.9.2342.19200300.100.1.1=2001734216
Date: 2017.02.24 14:05:24 -05'00'

Qian Li
Team Lead, Statistics
Division of Population Health Science, Office of Science

Qian H. Li -S
Digitally signed by Qian H. Li -S
DN: c=US, o=U.S. Government, ou=HHS, ou=NIH, ou=People, cn=Qian H. Li -S, 0.9.2342.19200300.100.1.1=1300136765
Date: 2017.02.24 14:21:26 -05'00'

THROUGH: Matthew Holman, Ph.D.
Director
Office of Science

Digitally signed by Matthew R. Holman -S
Date: 2017.02.24 14:35:35 -05'00'

TO: David Ashley, Ph.D. , RADM Public Health Service (Ret.)
Director
Office of Science

SUBJECT: Equivalence Testing for SE Evaluations

Introduction

The evaluation of chemistry information provided by an applicant in support of an SE Report or PMTA is a multi-step process. However, the ultimate goal is to allow the chemist to confidently state that a measured value of a constituent or product of a tobacco product (measurand) accurately represents the true value of that analyte. All measurements are approximations of the true value and some are better than others. Because analysts know that their measurements have inherent and often unknown uncertainty, replicate measurements are taken. However, once multiple measurements exist, an analyst has a series of estimations of the true value to deal with. If the source of the error is random and not systematic, then more measurements will reduce the error of the mean estimation of the true value. Analysts will typically turn to some form of statistics to provide a better estimation of the true value. In their simplest form statistics like averages and medians may provide a picture of the true value. However, it is important for analysts to consider the quality of the estimation, which leads to many of the more complex forms of statistics.

Such is the case in the evaluation of SE Reports. In an SE Report, the analyst needs to determine; (1) if the procedure used to measure the measurand is capable of providing a meaningful estimation of the

true value, (2) if the measured values of new product and the predicate product are the same, and (3) if they are different, is the difference sufficiently large to raise a concern. These decisions are often described using the following three questions (1) is it validated? (2) is it statistically significantly different? (3) is it importantly different, such that those differences cause the new product to raise different questions of public health? Validation is discussed at length in a number of different FDA guidances, international standards (ICH, Codex Alimentarius, ISO/Eurachem), excellent books on the topics[1], and is the subject of a working group (VRAM). So, Validation will not be discussed further in this document. There is also a tremendous amount written about statistics, but for most chemists the differences in approaches may seem esoteric, complicated, and just plain mysterious. The goal of this document is not to describe all of statistics, but rather to describe two approaches for determining whether the measured values of new and predicate products provided in an SE Report describe a substantially equivalent product.

It is important to understand that in SE evaluation of the new and predicate products, the differences in HPHC levels observed between products are due to product differences and error in the analytical measurement. For the purpose of comparing two products, the ideal approach is to eliminate the all possible error in the analytical measurement. Unless proper data are collected, it will not be easy to tease out the analytical error from the product difference. When the analytical error cannot be teased out of the data, the product comparison for SE evaluation may contain inherent bias.

Discussion

Student's T-Test

Depending upon how SE data are collected, the choice of the test statistics can be different. In the case of most SE Reports, the data consist of replicates from a single batch of each product, the Student's T-test, is the most commonly used statistical treatment. Often t-test is used to test if there is a statistically significant difference between the two means of the new and predicate products. The calculation (see Appendix A) requires the calculation of the mean of the replicates from each product, and the standard deviation of the replicates.

- If the number of replicates in each group is the same and the standard deviations are similar, use Equation A1
- If the number of replicates is different, but the standard deviations are similar, use Equation A2
- If the number of replicates is different and they have different standard deviations, use Welch's T-test – Equation A3

In the case where there are equal number of replicates and equal standard deviation, Equation A1 is calculated as the difference of the two means divided by the common standard deviation which is multiplied by the square root of twice the inverse of the number of replicates:

$$t = \frac{difference\ in\ the\ means}{standard\ deviation * \sqrt{\frac{2}{replicates}}}$$

After the t test statistics is calculated, it is compared to the t-distribution with 2(n-1) degrees of freedom to find the p-value of the test statistics. If the p-value (2-sided) is smaller than 0.05, the null hypothesis of equal mean is rejected. (see *Chemistry Reviewers Guide* for applications).

There are several *caveats* and considerations associated with the use of this approach:

1. Increasing replicates (↑) leads to an increased likelihood of rejecting the null hypothesis(↑). Conversely, a fewer number of replicates (↓) leads to an increased likelihood that the null hypothesis will not be rejected
2. A smaller measurement error(↓), may lead to an increased likelihood of not rejecting the null hypothesis(↑)

Therefore, a well-controlled, precise analytical procedure with large numbers of replicates is more likely to show statistically significant difference than a poor procedure with highly variable results and few replicates. This is counter-intuitive and can lead to difficult discussions in reviews. One alternative to the evaluation of the sameness of two means is to test their equivalence. The test of equivalence requires several changes in how we set up the question and knowledge of what an acceptable result would be.

When the collected data structure is complex (includes multiple batches that may come from different manufacture facilities and multiple laboratories for analytical analyses), statistical models to consider the effect among batches, manufacture facilities and laboratories t should be applied. The statistical branch would like to encourages reviewers to submit statistical consultations in such cases for proper statistical analyses of the data. The link of statistical consultation request can be found at http://sharepoint.fda.gov/orgs/CTP-OS/PopHealthSci/Statistics/SitePages/Statistics%20Consultation.aspx

Equivalence/ Non-inferiority testing

Instead of testing statistically significant difference, the SE evaluation should implement the equivalence or non-inferiority (or no-worse testing) hypothesis testing. The evaluation of SE does not require the new and predicate products to be exactly the same. Instead, the presumption is that the new and predicate product should be similar or equivalent. However, before one can determine that two results are equivalent, it is necessary to define what can be considered acceptable differences (Decision Rules) and what are not. The student's t-test and equivalence tests are similar, but require a different set-up for the null and alternative hypotheses and selection of pre-specified decision rules. The statistical underpinnings of the most popular equivalence test are the "two one-sided test" procedure or TOST. The results of the TOST calculations are compared using an equivalence margin, which is associated with the decisions rules, both of which are discussed further below.

To use the t-test to perform the equivalence test described in this document we will need:

1. The data provided in support of an SE Review consists of two data sets, one each for the predicate and the new product, or their surrogate(s).
2. Each data set represents a single lot of product
3. Each data set represents the results of a single brand of product
4. When the number of replicates is small, the data should be either normally or log-normally distributed;

In order to minimize potential bias in the equivalence test, it is preferable that the replicates are collected over a short time frame as t-test cannot handle covariates such as time period. In order to gain sufficient power for the equivalence test, sufficient independent replicates should be included in each data set. It would be preferable for the sample size justification to be included in SE submissions.

Where the above conditions are not met, a reviewer should consult statistics to valid evaluation and interpretation. In cases where the applicant has not specifically provided this information, the conditions may be assumed to be met for the purposes of data comparison while clarification concerning the data sets is sought (through a deficiency).

Decision rules

In order to set the null hypothesis, we first need to understand what amount of difference should be allowable to establish equivalence. For the purposes of this document, this allowable difference is termed "importantly" different. The selection of an importantly different value will depend upon our perspectives, our understanding of the measurements, and our intended use of the evaluation.

It would theoretically be possible to determine the level of change in the analyte value necessary to result in a change in public health outcomes and use this value as the important difference. However, this is beyond the scope of a chemistry review and will change depending upon the predicate product value and a number of other variables, making this approach inappropriate for application by DPS Chemists.

The best way to identify what an important difference would be is to test the entire population of a product for the analyte of interest. The spread of the data would define the boundaries of the variability of the products and will also include the variability of the analytical method. However, all analytical methods used for the evaluation of tobacco are destructive, thus rendering knowledge of the population unattainable. Whenever the evaluation of a fraction of a product is considered, careful selection of sampling plans that are intended to create a representative fraction should be developed. This sampling plan and the applicability of the sampled fraction to the totality of the batch or lot of the product is an important statistical discussion. Where these types of data are presented in an SE Report, the Chemistry reviewer should consult with the statistics branch. These types of considerations are the responsibility of the manufacturer, but, for purposes of this memo, we are presuming that the data provided as a part of an SE Report are representative of the product as a whole and not simply a sample of convenience. Using this presumption, then we are still faced with the comparison of two sample means, each having variability associated with the product and the method. Therefore, to determine if the products means are equivalent, we must remove the underlying method variability from the means.

This variability is convoluted within the sample means and may be deconvoluted if special efforts are taken to collect and evaluate the data. However, this type of information is rarely available and is an approximation at best. An alternative approach is to determine the important difference as a function of the method variability (independent of the product) and use this variability as an extension of our confidence intervals. Thus, method variability can be used to describe the important difference, thereby offsetting the method variability from the means and allowing the comparison of the product means. An important difference, so derived, could be termed "important analytical difference (IAD)". The term "analytical" is inserted to indicate that this determination of equivalence is not intended to reflect any public health considerations, but only that the results are analytical equivalence. Depending on specific public health considerations, the use of TOST may require specification of an important difference that is different than IAD.

The selection of an IAD will depend on the procedure, the product being measured, and the concentration of the analyte. Once chosen, the IAD will be used to define the Equivalence Margin (EM), discussed more fully below. There are several standard sources that could be referenced in the determination of the IAD. A discussion of the rationale for a recommended IAD is included in Appendix B. The recommend IADs are 10% for Tar and CO, 15 % for Nicotine, and 20% for other HPHCs, including B[a]P.

The next step in setting up an equivalence test is to set two boundaries – upper and lower limits (also known as the Equivalence Margins) for the differences between the new and predicate products. The equations for the calculations of the Equivalence Margin (EM)[2] are included in Appendix A (A4)[i].

Null & alternative hypothesis

The null hypothesis for the equivalence test is that the "Mean Range" between the new and predicate products is below the lower equivalence margin or above the higher equivalence margin and thus the products are not equivalent. The alternative hypothesis is that the Mean Range is within the high and low limits, and therefore are equivalent. The Mean Range (MR) is the 2-sided 90% Confidence Intervals[ii] for the difference of the means. The equations for the calculations of the Mean Range (MR)[2] are include in Appendix A (A5).

When either of the lower and upper bounds of Mean Range falls outside the range defined by the equivalence margins (illustrated by numbers 3, 4, 5, 6 and 7 in Figure 1), the null hypothesis is not rejected.

---

[i] Notice that the calculation of EM depends upon both IAD and mean levels of measurements of the new and predicate product. The dependence of EM on the means may raise concerns that larger mean values in either or both of the new and predicate products would allow wider acceptance of analytical difference. Further research may be needed to further understand the impact of such selection of EM. Further research may also be needed to understand possible data transformation, such as log-transformation of the measurements, before performing t-test. Log-transforming the data will allow the calculation of EM depending only on IAD, not the means of the new and predicate products.

[ii] The majority of T-tables are calculated as either single or two-tailed distributions, but since TOST is two one-sided determinations, the correct look-up table is the 90% confidence Interval, which reflects a 95% confidence level for the decision of equivalence.

If the two-side 90% CIs are within the equivalence margins, the null hypothesis is rejected which leads to the conclusion of equivalence (illustrated by numbers 1, 2 and 8 in Figure 1).

It is important to understand that not rejecting the null hypothesis is not the same as accepting the null hypothesis. Only if the entire 2-sided 90% CI as illustrated by numbers 6 and 7 in Figure 1 is completely outside the two low and high limits, the alternative is rejected and the null is accepted, leading to the conclusion of inequivalence. If the 2-sided 90% CI is not completely outside the low or high limits, the alternative hypothesis is not rejected, but may be considered as not demonstrating equivalence and inconclusive for equivalence (illustrated by numbers 3, 4, and 5 in Figure 1).

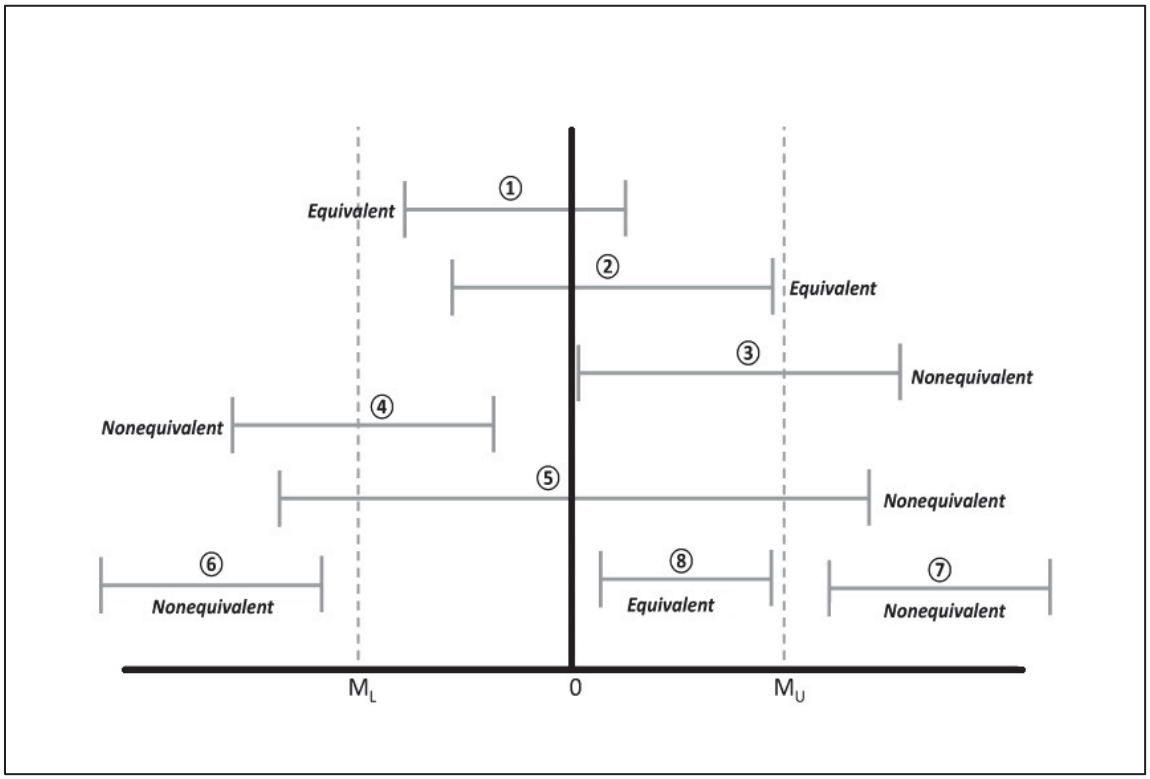


Figure 1. Determination of equivalence

To illustrate decision rules and their effect on determination of mean equivalence, consider a traditional American-blend cigarette that has been tested for TNCO and B[a]P using the ISO regimen:

Table 1. HPHC values for an example cigarette* - ISO regimen (Mean (mg/cig) w/ Std Dev.)

HPHC	New Product [Mean (SD)]	N	Predicate Product [Mean (SD)]	N
Tar (mg/cig)	13.4 (0.6)	20	14.3 (0.7)	8
Nicotine (mg/cig)	0.83 (0.04)	20	0.95 (0.12)	8
CO (mg/cig)	16.4 (0.9)	20	17.0 (0.8)	8
B[a]P (ng/cig)	10.8 (1.3)	7	10.7 (3.0)	8

*numbers are based on actual submitted measurements

For our example, the IADs are set to ±10% for tar and CO, ±15% for nicotine and 20% other HPHCs per Appendix B.

When the equivalence approach is applied to the Tar example (see the worked out calculation in Appendix C) above the MR is found to be -1.38 to -0.42 and the EM is ±1.39. Because both of the upper and lower ranges are within the EM, the tar values can be considered equivalent. When the same data is analyzed using the t-test, the tar values are found not to be the same (p=0.0035). This shows that data collected with good precision may lead to a potentially misleading indication of the results not being the same when they are similar enough for equivalence purposes.

The nicotine values present a different challenge. The t-test for nicotine also shows that the results are statistically significantly different (p=0.01), but the equivalence test results in an inconclusive result (MR = -0.05 to -0.19, EM = ±0.13). The MR overlaps the lower margin, resulting in our inability to reject the null hypothesis of non-equivalence, but it is not clearly inequivalent either. This would be a case of inconclusive equivalence. A larger sample size may potentially narrow MR and lead to equivalence conclusion.

CO is straight forward as both the t-test (p=0.1) and the equivalence test (MR = -0.01 to -1.19, EM = ±1.67) support the results as the same/equivalent result.

Finally, the B[a]P results are also inconclusive. Here, the large error in the measurement causes the t-test to indicate that the results are the same (p=0.9), however, the equivalence evaluation raises doubt about the confidence that should be placed in this data. The low number of replicates and large error lead to an MR of 2.17 to -1.97 with the EM of ±2.15. Again, this is a judgment call, but the equivalence test allows us to ask better questions.

There are many different of ways of calculating margins available in the literature for evaluating TOST, most are variants of the provided equation and are designed to address a specific application. One that may be useful for the interpretation of equivalence in an SE calculates the minimum EM value for a comparison based upon the average standard deviation and total number of measurements[3] (See Appendix A (A7)). This calculation does not provide an evaluation of the equivalence of the mean, but it does provide the check on the EM that is chosen for your comparison that can indicate whether the method used for the measurement is capable of meeting the statistical needs of the evaluation. In the B[a]P case above, the EM is ±2.15 while the minimum EM value is calculated as ±2.09, further supporting the need for more replicates or better SD for this determination.

Sample Size to analyze IAD

The equivalence test using a t-test is sensitive to the number of replicates measured. Please note that replicates need to be independent measurements of the sample (multiple solutions), not simply multiple injections of the same solution for the purposes described in this document. Increasing the number of measurements will improve the statistical power by improving the estimation of the true mean and improving the estimation of the standard deviation. The necessary size should be linked to the

confidence level desired for the decision, typical standard deviation and the EM. Because the EM may not be available before the testing is completed, as the means of the new and predicate products are unavailable before data collection, best guesses are appropriate. Here the rule of thumb is larger standard deviations need more replicates. An equation to provide approximate replicate needs is included in Appendix A (A6). The size of data needed is not onerous, as the *n* is factored into the decisions. All of the examples, except B[a]P had sufficient replicates for a good equivalence measure. The B[a]P measurement needs about 20 total replicates to increase our confidence.

Recommendation

Where data provided in an SE Report are sufficient to allow the use of the equivalence test (TOST), this approach is an important addition to the chemists' statistical toolkit. It is important, however, for the DPS to identify IADs for each analyte in cigarette smoke, tobacco filler, ENDS aerosols, and e-liquids prior to implementation of this approach. Appendix B provides a recommendation that is based on international public standards and recent FDA guidance. This recommendation provides a system that is flexible enough to apply to all of the HPHCs currently published in the CTP guidance and those that will be identified for newly deemed products. Further, basing the IADs upon the Horwitz-Thompson equation allows the use of the equivalence approach to be used for much higher concentration components used to constitute e-liquids. The recommended IAD is independent of the procedure used, but reflects the limitations of analytical chemical determinations. It is recommended that the TOST approach with IADs calculated using the Horwitz-Thompson equation be used as an additional statistical tool for the evaluation of SE Reports. This document provides an initial step of formulating the statistical tool which may only apply to submissions that have only one manufacturing run from each product. As more manufacturing runs are recommended in recent SE submission, the tool may need to be updated in the future to comply with the new trend of recommendation.

Appendix A - Equations

- **Equation A1 – t-test, common *n* and SD**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{2/n}}$$

Where $s_p(pooled\ standard\ deviation) = \sqrt{\frac{s_1^2 + s_2^2}{2}}$

s_1 *and* s_2 = standard deviation of data set 1 and 2

$\bar{X} = \text{mean}$

- **Equation A2 – t-test, different *n*, same SD**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where $s_p(pooled\ standard\ deviation) = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1 + n_2 - 2)}}$

s_1 *and* s_2 = standard deviation of data set 1 and 2

$\bar{X} = \text{mean}$

- **Equation A3 – t-test, no assumptions of *n* or SD**

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{calc}}$$

Where $s_{calc} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

s_1 *and* s_2 = standard deviation of data set 1 and 2

$\bar{X} = \text{mean}$

- **Equation A4 – Equivalence Margin**

M_U = average mean * IAD (as a fraction);
$M_L = M_U * -1$

Where M_U is the upper margin and M_L is the lower margin

- **Equation A5 – Mean Range**

$$\bar{X}_1 - \bar{X}_2 \pm t_{0.05,(n_1+n_2-2)}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where s_1 *and* s_2 = standard deviation of data set 1 and 2

$\bar{X} = \text{mean}$

$t_{(0.05,(n1+n2-2)}$ = value from a t-table at 2-sided 90% CI (0.05), and degrees of freedom

- **Equation A6 – Sample Size Needs (for each mean determination)**

$$n = \frac{2s^2(2z_{0.05})^2}{EM^2} + 1$$

Where s = standard deviation
z = z-statistic for a 95% CI
EM = Equivalence Margin

- **Equation A7 – Minimum Equivalence Margin**

$$\text{Minimum EM} = s * [t_{0.95,(n_1+n_2-2)} + t_{0.975,(n_1+n_2-2)}] * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Where s = standard deviation
$t_{(0.95,(n1+n2-2)}$ = value from a t-table at 95% CI, degrees of freedom

Appendix B. Important Analytical Difference (IAD) for Tobacco Measurements

The most important parameter in the evaluation of IAD is the analytical variability (i.e., precision) of a method. Analytical variability is the topic of a number of treatises, standards, books, and discussion forums. Many of the approaches are for very specific applications other than tobacco, but may be applied to the question of an IAD, others are broadly applicable to all analytical measurements, and some are specific for the measurement of tobacco products and their products (e.g., smoke, vapor, liquid extract (smokeless)). The later seem to be the most appropriate for this question, but many have limitations and specific applications that preclude their use in general applications such as those considered herein. This discussion will attempt to cover many of the most often common sources and their applicability to the question of IAD; however, it is by no means a comprehensive review of the topic.

National Metrology Institutes

Analytical variability is regularly discussed by national metrology institutions (e.g., Bureau International des Poids et Measures (BIPM) (global)[4], National Institute of Standards and Technology (NIST)(USA)[5], National Physical Laboratory (NPL)(UK)[6]) in relation to the standards that support national and international standards. These institutions often attempt to measure small components in a complex matrix which is similar to the challenge of measuring tobacco and smoke constituents. However, the focus of these institutes is largely on determining whether a measurement made by an instrument provides a "True" measure of the reference standard material[7]. NMI standard reference materials (SRMs) have been carefully measured and are generally unchanging, unlike the tobacco reference standard materials. Because of the stability of the SRM and the rigor of the analytical testing to provide very low measurement uncertainty, the NMIs are able to focus their acceptability criteria solely upon the bias of an analytical measurement. These limits are very tight and are linked to the minimum detectable difference, rather than identification of an IAD. These types of approaches are not applicable to the determination of the equivalence of two measurements of unknown content, as is necessary in an SE Review. However, their approaches can be applied to understand the lowest level of variability attainable with a specific procedure.

Standard Setting Organizations

In addition to the NMIs, there are a number of standard setting organizations (e.g., ISO, Eurachem[8], CORESTA, USP, etc.) that have made statements about analytical variability in a variety of applications, including in tobacco products. Most national standard-setting organizations are part of the International Organization for Standardization (ISO) and reflect the ISO standards and nomenclature. ISO has published a series of interrelated standards that speak to portions of IAD. The standard for measuring accuracy and precision (Trueness) is contained in ISO 5725-1, -2, and -4[9, 10, 11]. This standard is referenced by ISO/TR 22305[12], which describes the precision of the measurement of cigarette smoke for TNCOs. This standard is further referenced by ISO 8243[13], which describes sampling schemes and tolerances for confidence intervals relative to package

labeling. In these standards, the ISO standard states that the reproducibility of the TNCO measurements of 800 cigarettes support 15% - 25% difference between labeled content and measured content, depending on the sampling plan and the analyte. This tolerance could serve as the IAD; however, there are several restrictions in this standard that impact use of its tolerance as an IAD. The key restrictions are:

- Comparison is made for a tobacco product to a label statement, not between two tobacco products
- Applies to specific ISO analytical sampling and measurement procedures only [vii]
- Only applies to cigarettes
- ISO 22305:2006 states "The *r* and *R* values [repeatability and reproducibility values] from collaborative studies are thus essentially estimates of measurement variability on nearly identical samples. They cannot be used directly as a tolerance for compliance checks of cigarette brands where other sources of variability must be taken into account" [vii]

Although these standards are specific for tobacco, they have been crafted for a very specific purpose that differs from the application considered herein. These standards seek to describe an allowable level of variation from cigarette to cigarette in a single lot. It is neither designed to consider two products with different design or content, nor is it designed to provide an understanding of an acceptable level of variation in the measurement system. Further, these standards are specific for the ISO smoking system and TNCO measurements only.

Finally, this standard is intended to apply to all products on the marketplace and states that it has been broadened to be generally applicable (i.e., less discriminating). Data provided in ISO/TR 22305:2006[3] (CORESTA sourced) provides a demonstration of this point (see Table 2). The report presents TNCO measurements using 7 different types of smoking machines, 8 different cigarette brands, and as many as 39 different labs (not all labs did all tests). This report does not report the standard deviation of the measurements, but instead reports the mean, repeatability, and reproducibility calculated values or the way in which they were calculated. Regardless, the mean results calculated across all labs and machines resulted in relative repeatability (%*r*) of under 15% (6.01%-13.84), with most (21 of 24) measurements presenting about 10% RSD or less. While CORESTA[14] and ISO recommend using the relative reproducibility (%*R*), the rationale for that recommendation is not consistent with the definition of *R* in ISO 5725-1, which states that *R* is the "precision where the test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment." Further, the approach advocated by this standard is not intended to compare two products or even to compare the results of two measurements of a single product, but are instead designed to provide a recommendation on package labeling for TNCOs for all products across all marketplaces. Additionally, the standard does not provide any information or approaches to deal with the minor components (HPHCs) that are

[3] The data provided was included in Appendix F and is a reprint of the CORESTA 2003 collaborative study on TNCO repeatability.

also of concern in the evaluation of SE Reviews. Therefore, the ISO recommendations are not usable as a source of IAD.

Table 2. CORESTA 2003 Report and Calculated Repeatability and Reproducibility

Analyte	Brand	Mean (µg/cig)	*r*	*R*	% *r*	%R
CO	Camel	8.31	0.82	1.72	9.9	20.7
	CM4	13.23	0.87	2.2	6.6	16.6
	Ducados	10.99	1	2.11	9.1	19.2
	Marlboro	11.85	0.95	2.22	8.0	18.7
	Marlboro Lights	6.79	0.75	1.42	11.0	20.9
	Pall Mall 100	9.55	0.74	1.66	7.7	17.4
	PM Super Lights	4.12	0.52	1.08	12.6	26.2
	Regal	11.45	0.97	2.09	8.5	18.3
NFDPM	Camel	12.68	0.94	2.12	7.4	16.7
	CM4	14.06	0.74	1.73	5.3	12.3
	Ducados	10.07	0.66	1.24	6.6	12.3
	Marlboro	11.56	0.71	1.11	6.1	9.6
	Marlboro Lights	5.43	0.55	0.8	10.1	14.7
	Pall Mall 100	10.04	0.64	1.01	6.4	10.1
	PM Super Lights	3.54	0.49	0.75	13.8	21.2
	Regal	9.83	0.72	1.3	7.3	13.2
Nicotine	Camel	0.94	0.073	0.133	7.8	14.1
	CM4	1.298	0.078	0.145	6.0	11.2
	Ducados	0.817	0.068	0.123	8.3	15.1
	Marlboro	0.848	0.059	0.107	7.0	12.6
	Marlboro Lights	0.465	0.037	0.076	8.0	16.3
	Pall Mall 100	0.79	0.063	0.1	8.0	12.7
	PM Super Lights	0.334	0.035	0.073	10.5	21.9
	Regal	0.883	0.071	0.124	8.0	14.0

Horwitz Approach

Outside of the ISO standard, there are a couple of focused groups that have addressed the issue of IAD. These groups include FAO (Food and Agriculture Organization of the United Nations), EPA (Environmental Protection Agency), USP (United States Pharmacopeia, and FDA (Food and Drug Administration). The EPA and USP have advocated the use of "performance-based procedures"[15] which define the criteria needed to demonstrate a procedure is acceptable. These approaches, while similar to the IAD values that are sought herein, rely on a series of standard acceptance criteria that are not available in the tobacco industry. The *Codex Alimentarius*[16], FDA/OFVM[17], and other authors[xvi], recommend the approach developed by William Horwitz[18]. This approach links the concentration of the analyte in a matrix with an expected level of variability (*R*). This linkage was empirically derived as the Horwitz-Thompson equation:

%RSD = $2C^{-0.15}$, where C is the concentration fraction which is dimensionless weight fraction (see Table 3).

Because this value is based upon results from multi-center collaborative trials, a repeatability value is typically presented as approximately ½ that of the Horwitz value (except at values below ng/g levels)[19].

Table 3. Calculation of Concentration Fraction and Horwitz Values

Concentration of Analyte (%)	Concentration of analyte (ppm or ppb)	Concentration with w/w units (mg/g or µg/g)	Concentration Fraction	Horwitz–Thompson Value (%RSD)	Repeatability Value (% RSD)
10		1000 mg/g	0.1	2.8	1.4
1		10 mg/g	0.01	4	2
0.1	1000 ppm	1 mg/g	0.001	5.7	2.8
0.05	500 ppm	500 µg/g	0.0005	6	3
0.01	100 ppm	100 µg/g	0.0001	8	4
0.001	10 ppm	10 µg/g	0.00001	11	6
0.0001	1 ppm	1 µg/g	0.000001	16	8
0.00001	100 ppb	0.1 µg/g	0.0000001	22	11
<0.00001	< 100 ppb	< 0.1 µg/g	< 0.0000001	22	22

This approach was developed as a measure of the analytical system, includes variability associated with different labs and times, and is independent of individual analytical technique. This approach would need a calculation of the Horwitz value for every analyte relative to the expected typical concentration of that analyte. For example, if the data in Table 2 were used to calculate the Horwitz value, NFDPM (tar), nicotine, and carbon monoxide would define and acceptable precision of 11%, 16%, and 11%, respectively. These values are similar to the %*r* values in the Table 2. To use this approach, it is also important to know whether the results reported in an SE Review represent

products tested at different times or side-by-side. If tested in a close proximity in time, using the same instrumentation, then the repeatability values in the table should be used as the IAD.

Application of Horwitz to Tobacco Product Testing

The Horwitz approach is predicated on the observation that as the concentration of the analyte decreases, the amount of variability increases. The amount of variability to be expected can be calculated independently for each measurement. However, tobacco product contents and yields can be approximated as shown in Table 4. While the expected concentrations are approximations, they only need to correct to an order of magnitude for the calculation of a Horwitz value.

Table 4. Tobacco Product Analyte Concentrations and associated Horwitz Values

Analyte	Medium	Analyte Concentration (approximate)	Horwitz-Thompson Value (% RSD)	Repeatability Value (% RSD)
Nicotine	Smoke	1 µg/g	16	8
Nicotine	E-liquid	10 mg/g	4	2
Nicotine	ENDS Aerosol	10 mg/g	4	2
Nicotine	Oral Extract	10 mg/g	4	2
Tar	Smoke	10 µg/g	11	6
CO	Smoke	10 µg/g	11	6
Propylene Glycol	E-liquid	10-90 mg/g	4	2
Glycerol	E-liquid	10-90 mg/g	4	2
Flavors	E-liquids	100 µg/g	8	4
Carbonyls	All	1 µg/mL (DNPH)	16	8
TSNAs	All	10 ng/g	22	22
PAHs	All	1 ng/g	22	22
Metals	Filler & E-liquids	100 ng/g	22	11

Please note that the approximate target concentrations in the table assume 1 gram of tobacco filler/cigarette, ISO protocols, and reflect the lowest concentrations expected, which gives the most generous expected variability. Remember that the values in the table do not reflect a mandatory precision of a procedure used to measure HPHCs, but rather is the margin beyond which two mean values of products being tested are evaluated for equivalence. Nicotine shows the largest variability in terms of target concentration and therefore in the expected measurement variability 4-16% RSD. Tar and Carbon Monoxide are measured at approximately the same level so each would be expected to have similar levels of analytical variation, approximately 11% RSD. The rest of the HPHCs are present at much lower levels (microgram-nanogram levels) and thus higher variability is expected (22% RSD). The Horwitz is an exponential equation and reaches the asymptote at 22 % RSD, however this value acceptability of this model is validated through its use in the FVM guidance

and concentration levels of 10X lower concentration in food (an equally complex matrix as tobacco products) than the lowest target concentration in tobacco. It may be appropriate to use Table 4 above as a reference during SE evaluations, for the ease of application, it is recommended to round the levels for use in IADs to 10% for tar and carbon monoxide, 15% for nicotine and carbonyls, and 20% for all other HPHCs. ENDS products and smokeless products have much higher concentrations of ingredients and extracts, respectively. For all of the listed constituents and extracts, an IAD of 4% RSD is indicated. While the Horwitz equation provides a convenient and consistent standard for the comparison of data, these values represent the beginning of the discussion rather than the end.

Recommendation

Based upon the correspondence of the %r values and the Horwitz values, and the long-standing application of the Horwitz approach to set IAD parameters in materials similar to tobacco, it is recommended to use 10%, 15%, and 20% as the IAD for tar and CO, nicotine and carbonyls, and all other HPHCs, respectively, for tobacco smoke. The acceptable values for larger components in ENDS and smokeless would be 4% for PG, VG, nicotine, all other ingredients, with HPHCs values mirroring the values for smoke. In the future, additional research into the most appropriate IADs may change this initial recommendation.

One concern with using the Horwitz approach is that the IADs would not directly reflect the variability of tobacco products over time and would only reflect the variability at a point in time. While this is factually correct, the IAD associated with a brand of cigarette are not generally provided in SE Reports. Where an applicant feels that an alternative IAD should be used and presents data to support their claim, then FDA should evaluate the IAD on a case-by-case basis.

Appendix C. Calculation of Example

Table - repeated. HPHC values for an example cigarette* - ISO regimen (Mean (mg/cig) w/ Std Dev.)

HPHC	New Product [Mean (SD)]	N	Predicate Product [Mean (SD)]	N
Tar (mg/cig)	13.4 (0.6)	20	14.3 (0.7)	8
Nicotine (mg/cig)	0.83 (0.04)	20	0.95 (0.12)	8
CO (mg/cig)	16.4 (0.9)	20	17.0 (0.8)	8
B[a]P (ng/cig)	10.8 (1.3)	7	10.7 (3.0)	8

*numbers are based on actual submitted measurements

For our example, the IADs are set to ±10% for tar and CO, ±15% for nicotine and 20% other HPHCs.

For Tar:

EM = ±Mean * IAD

M_L = -(13.4+14.3)/2 * 0.1 = -1.39

M_U = (13.4+14.3)/2 * 0.1 = 1.39

$$\mathrm{MR} = \bar{X}_1 - \bar{X}_2 \pm t_{0.05,(n_1+n_2-2)}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$MR_U = 13.4 - 14.3 + t_{0.05, 26} * Sqrt(0.6^2/20 + 0.7^2/8)$

= -0.9 + 1.71 * Sqrt (0.36/20 + 0.49/8)

= -0.9 + 1.71 *0.285 = -0.42

$MR_L = 13.4 - 14.3 - t_{0.05, 26} * Sqrt(0.6^2/20 + 0.7^2/8)$

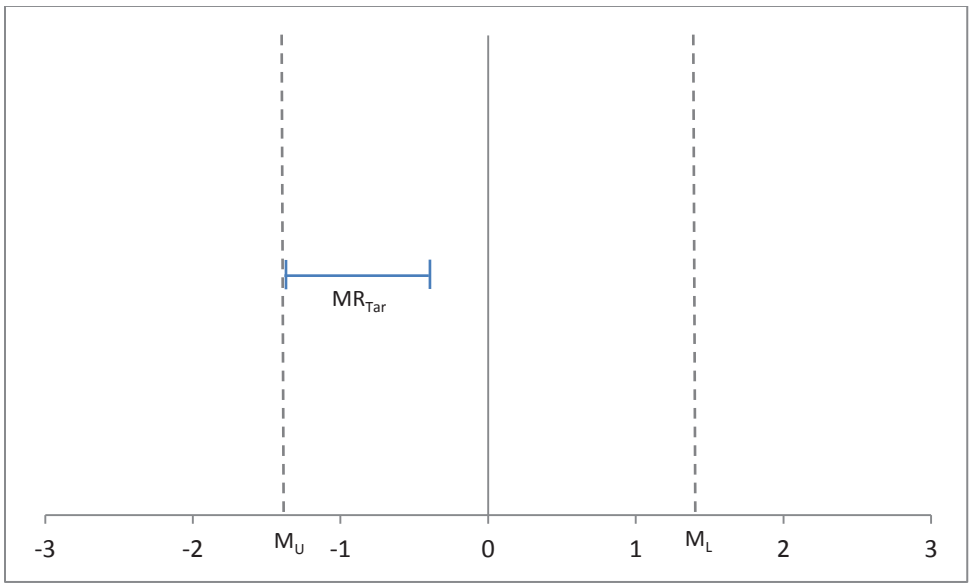
= -0.9 - 1.71 * Sqrt (0.36/20 + 0.49/8)

= -0.9 - 1.71 *0.285 = -1.38

[Note: the t values either come from a t-table or can be calculated in Excel using the T.INV.2T function]

A graphical representation of the results is included in Figure 2.

Figure 2. Illustration of Equivalence Example



References

[1] Ermer, J. and P. W. Nethercote (2014). Method Validation in Pharmaceutical Analysis. Singapore, Wiley-VCH.

[2] Chambers, D., et al. (2005). "Analytical Method Equivalency: An acceptable Analytical Practice." Pharmaceutical Technology (September 2005)

[3] Limentani, G. B., et al. (2005). "Beyond the t-test: statistical equivalence testing." Anal Chem **77**(11): 221A-226A.

[4] http://www.bipm.org

[5] https://www.nist.gov/

[6] http://www.npl.co.uk/

[7] Sharpless, K. E., et al. NIST Special Publication 260-181 The ABCs of Using Standard Reference Materials in the Analysis of Foods and Dietary Supplements: A Practical Guide.

[8] https://www.eurachem.org/

[9] ISO 5725-1:1993, Accuracy (trueness and precision) of measurement methods and results – Part1: General principles and definitions

[10] ISO 5725-2:1993, Accuracy (trueness and precision) of measurement methods and results – Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method

[11] ISO 5725-4:1993, Accuracy (trueness and precision) of measurement methods and results – Part 3: Basic methods for the determination of the trueness of a standard measurement method

[12] ISO/TR 22305:2006, Cigarettes – Measurement of nicotine-free dry particulate matter, nicotine, water, and carbon monoxide in cigarette smoke – Analysis of data from collaborative studies reporting relationships between repeatability, reproducibility, and tolerances

[13] ISO 8243:2013, Cigarette - Sampling

[14] CORESTA 2003, Collaborative Study Report, CORESTA study for the estimation of the repeatability and reproducibility of the measurement of nicotine-free particulate matter, nicotine, and CO in smoke using the ISP smoking methods

[15] Williams, R. et.al (2009). "Performance-based Monographs". Pharm Forum, **35**(3).

[16] *Codex Alimentarius*. Rome: Food and Agriculture Organization of the United Nations, 2010.

[17] FDA. Office of Food and Veterinary Medicine (2015). Guidelines for the Validation of Chemical Methods for the FDA FVM Program.

[18] Horwitz, W. (1997). "The variability of AOAC methods of analysis as used in analytical chemistry." J. Assoc. Off. Anal. Chem. **60**: 1355-1363.

[19] Massart, D. L., et al. (2005). "Benchmarking for analytical Methods: The Horwitz curve." Lc Gc Europe **18**(10): 528-+.