THS 2.2 Comparability Strategy

AUTHOR(S):	Donatien Tafin Djoko, *Manager Product Science* (b) (6) 16-Mar-2020
SME(S):	Florence Vonmoos, Manager Product Safety & SSM 16-Mar-2020 (b) (6)
REVIEWER(S):	*See Review section.*
OWNER & APPROVER:	Marco Esposito, Director Product Stewardship & SSM 16-Mar-2020 (b) (6)

Table of Contents

1 Introduction 3
2 Scope 4
3 Definitions 4
4 Comparability Model Aerosol Chemical Composition 5
4.1 Analytical Variability – Head-to-Head Comparisons 5
4.1.1 Overview 5
4.1.2 Statistical Methods and Classification of Endpoints 5
4.1.3 Sample Size and Statistical Robustness 6
4.1.4 Normalization 7
4.1.5 Acceptance Criteria 7
4.1.6 Monitoring and Model Review Frequency 10
4.2 Agricultural and Analytical Variability – Product Performance Comparability 10
4.2.1 Overview 10
4.2.2 Statistical Methods and Classification of Endpoints 11
4.2.3 Sample Size and Statistical Robustness 12
4.2.4 Normalization 12
4.2.5 Acceptance Criteria 12
4.2.6 Product comparability acceptance criteria 15
4.2.7 Monitoring and Model Review Frequency 15
5 Comparability Model *In Vitro* Neutral Red Uptake Cytotoxicity 15
5.1 Product Performance Comparability 16
6 Reference Documents 17
7 Version History 17
8 Review 18

1 Introduction

Product performance comparability is a major step in the change management process[1]. It is employed to demonstrate an absence of meaningful impact of product changes on the composition of the product aerosol and, where applicable, on the biological activity of the aerosol fractions[1], specifically that changes made on the product or manufacturing process do not negatively impact the performance of the product beyond the precision of the assays and the natural product variation over time. The present approach to comparability is in alignment with relevant U. S. Food and Drug Administration (FDA) and International Conference for Harmonization (ICH) guidelines. It follows a tiered approach, starting from the analytical assessment of aerosol chemistry and when necessary involving *in vitro* assays.

The selection of aerosol constituents for quantification[3] took into account priority toxicants in tobacco smoke as listed by major regulatory bodies worldwide, constituents with ISO-standard methods for quantification or with established biomarkers of exposure in clinical trials, potentially harmful and quantifiable constituents which are predominantly formed below 400 °C and classified by the International Agency for Research on Cancer (IARC) as Group 1 carcinogens, potentially harmful constituents which are quantifiable and predomninantly formed above 400 °C and which are classified by the IARC as Group 1 carcinogens, and several product-specific analytes such as the aerosol-former glycerin. The constituents selected according to these criteria are referred to as the PMI-58.

Correspondingly, the first tier of the *in vitro* assays was selected to cover different toxicity endpoints, based on their regulatory acceptance and on the feasibility of their use as an efficient screening tool: the neutral red uptake cytotoxicity, and the Ames bacterial mutagenicity or Mouse Lymphoma assays were identified as first priority.

To provide a framework for result evaluation and interpretation, the comparability model takes into account long-term analytical variability of each assay, and product agricultural variability to define product performance ranges representative of natural, unavoidable variability of product performance over time. These ranges represent the normal ranges of product performance with the existing sourcing and manufacturing controls. In turn, some of the manufacturing controls, namely product performance and batch release specifications, are validated at the time when they are established for specific product variants against the comparability model[Error! Reference source not found.]. This cross-link ensures that once a product variant is deemed comparable, manufacturing controls will ensure that comparability is not compromised over time.

The statistical fundamentals necessary to address the long-term variability effects have been previously described[4] in the context of combustible cigarette assessments, and these considerations are equally applicable for reduced-risk heated tobacco platform evaluations.

Briefly, long-term analytical variability leads to the observation that the same batch of product assessed for the yields of aerosol constituents in the same laboratory with the same analytical methods but at two different points in time, is likely to produce statistically different results independent of the number of replicates taken. Agricultural variability leads to the same observation due to other factors, namely crop variability over time. Indeed, these factors are duly recognized for example in ISO methods for the few aerosol constituents for which they are established, e.g. mandated tolerances for printed pack values.

As a result, to provide meaningful comparative assessments, acceptance ranges need to be defined. This is the primary objective addressed within the scope of the comparability model.

2 Scope

The present document summarizes statistical methods and acceptance criteria defined for THS 2.2 (Platform 1, known under its commercial name *IQOS HEETS*) product performance comparability assessments in the case of:

- Aerosol constituent yields measured at Labstat International ULC (262 Manitou Drive, Kitchener, Ontario, N2C 1L3, Canada), an independent ISO 17025 accredited tobacco testing laboratory.
- *In vitro* cytotoxicity results obtained at PMI Product Testing Laboratories, Neuchâtel, Switzerland.

3 Definitions

Long-term variability: Inherent variation in tobacco product assessment results obtained in studies conducted in the same laboratory using the same methods but at different points in time, manifested in statistically different results regardless of the number of replicates in a study[4]. Statistically, the average of measurements conducted on a single occasion does not converge to the population mean, which can only be estimated by considering multiple measurements conducted over a long period of time. Sources for long-term variability include analytical method including factors such as laboratory technician or aerosol generation machine impact, product variance stemming from crop-to-crop agricultural variability, and manufacturing variability which is the study objective of a dedicated recent CORESTA Task Force.

Product performance: Chemical characterization or *in vitro* biological activity assessment of the product aerosol. Chemical characterization refers to the measurement of the yields of a number of harmful and potentially harmful constituents in the product aerosol, typically the PMI-58 list of constituents. *In vitro* biological activity refers to the assessment of *in vitro* cytotoxicity, and mutagenicity of the product aerosol. These assays were selected for their regulatory acceptance and based on their feasibility to be employed as rapid screening tools. The measure of a difference in product performance implies the evaluation of the performance of a new product variant against a reference or control product variant for each endpoint, whereby an endpoint refers to an individual aerosol constituent in the case of aerosol characterization, or an individual *in vitro* assay endpoint in the case of biological aerosol activity assessments.

Reference product: Variant of THS 2.2 which underwent pivotal pre-clinical and clinical studies. Serves as the reference point for comparability assessments.

Control product: When considering e.g. product design changes, a product configuration prior to the change being assessed. Serves as the reference point for head-to-head comparisons.

Test product: Product configuration post-change, to be compared to the reference product by means of comparability, and/or to a control product by means of a head-to-head comparison.

Comparability model: Set of statistical methods and acceptance criteria enabling comparability, initial screening, and head-to-head comparisons to be performed. Each of these three assessment types is described further in the present document.

Endpoint: a single aerosol constituent yield in the case of aerosol chemistry characterization, or an *in vitro* endpoint.

THS 2.2 monitor: A single batch of THS 2.2 (or if unavailable, different batches of the same tobacco lot set) which is analysed in parallel with study samples, with each study. The results of the THS 2.2 monitor are used during follow-up investigations to evaluate laboratory method trends over time, verify contemporary validity of tolerance estimates, and provide any further supporting evidence for an investigation.

4 Comparability Model Aerosol Chemical Composition

4.1 Analytical Variability – Head-to-Head Comparisons

4.1.1 Overview

Head-to-Head comparisons are employed to assess the potential impact of product changes[1] by comparing the difference in the performance of a pre-change product variant (control product) to the performance of a post-change product variant (test product) constituent by constituent, and evaluating the magnitude of the differences against the corresponding long-term analytical variability[6] for each constituent, which essentially provides the inherent precision of the analytical method over time. Such an assessment is of a particular utility for assessing product changes where the experimental design allows a pre- and a post-change product manufactured with the same tobacco lot set, i.e. excluding the impact of agricultural product variability from the assessment.

This long-term analytical variability is evaluated empirically, through measurements conducted on the THS 2.2 monitor – a single batch of THS 2.2 assessed for aerosol constituent yields over an extensive period of time, ideally spanning a full year of at least five different studies. The variability estimates are re-evaluated annually. The data on which variability estimates are established at any given time is referred to as the THS 2.2 monitor data set.

The following values are derived from the THS 2.2 monitor data set for each Harmful and Potentially Harmful Constituent (HPHC):

- Monitor data set average – the mean value of all replicates in the THS 2.2 monitor data set;
- Tolerance based on long-term analytical variability – a value derived from the relative standard deviation of the THS 2.2 monitor data set, when all replicates are above the limit of quantification (LOQ);
- Threshold - a value derived from the value of the LOQ, and considering internally established exposure limit thresholds, when some replicates in the THS 2.2 monitor data set are below the LOQ.

4.1.2 Statistical Methods and Classification of Endpoints

Product performance is compared through pairwise comparisons of yields of individual aerosol constituents between the control and the test products. The yields are compared by means of the statistical equivalence test, unless they are below the LOQ in the THS 2.2 monitor data set, control, or test products. The application of the statistical equivalence test at a Type-I error rate of 5% is operationally equivalent to comparing a 90% confidence interval on the difference of two means, to a chosen equivalence range. The comparison can be presented graphically, by showing the difference between the yields of the test product and of the control product, expressed as a percentage of the monitor data set average. A 90% confidence interval of the difference of means expressed as a percentage of the monitor data set average is shown with error bars. The chosen tolerance (equivalence) ranges are shown with boxes, and are based on the variance of the reference value. **Figure 1** illustrates the graphical presentation and examples of interpretation of the results.

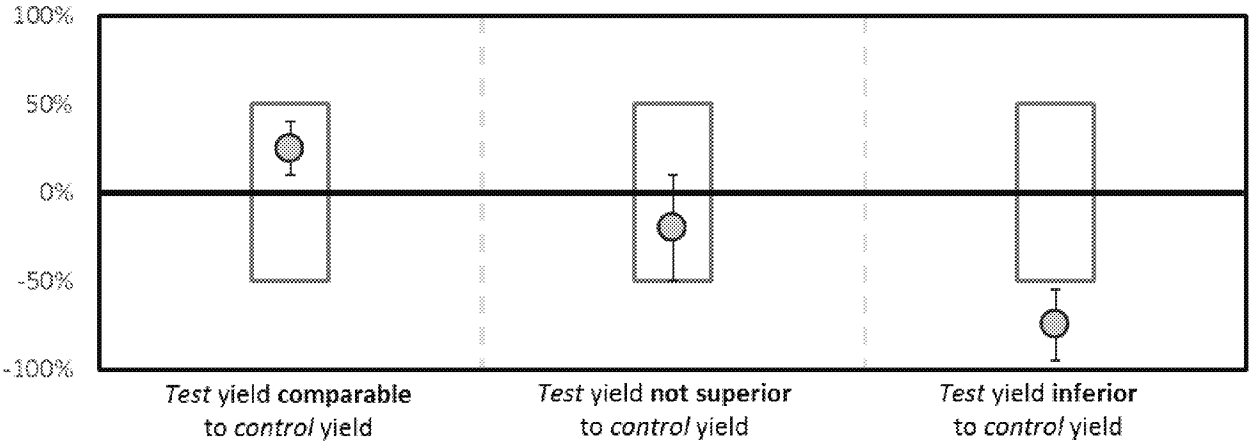


Figure 1: illustration of the graphical representation of the equivalence test, with examples of interpretation of results. Solid circles illustrate the difference in yields of the test item and the control item expressed as a percentage of the monitor data set average. Error bars illustrate the 90% confidence interval of the difference of the two yields, scaled by the monitor data set average. Solid boxes illustrate tolerance ranges.

If some of the yields in the monitor data set, or the yields of the control or test item in the study, are below the LOQ for a constituent, then the application of the equivalence test as described above is not possible. In this case:

- If the THS 2.2 monitor data set contains values below the LOQ, variability of the measurement cannot be calculated and consequently variability-based tolerances cannot be defined. A threshold-based approach is employed, whereby effectively it is considered that noise-level values were measured for THS 2.2, and consequently on the test and control items the yields should be also within noise level. To this end, the maximum measured value in the test item is assessed against a threshold value which is established based on the value of the LOQ and considering internally established exposure limit thresholds.
- Cases where the THS 2.2 monitor data set does not contain values below the LOQ, but the test or control items do contain values below the LOQ, indicate that overall the THS 2.2 yield for the constituent is very close to the LOQ value; such cases should be investigated during the annual tolerance limit review, as they may indicate limitations of the previously available THS 2.2 monitor data set. For the purpose of the assessment, in such cases the point estimate of the difference between the median yields of the test and the control items, expressed as a percentage of the THS 2.2 monitor data set average, is assessed against the defined tolerance.

4.1.3 Sample Size and Statistical Robustness

At least one batch each of control and test product should be assessed, with at least three replicates per constituent. Statistical robustness is improved if at least three batches each of control and test product can be compared; however, it should be noted that the head-to-head assessment methodology is only intended to provide an efficient screening methodology to enable targeted follow-up investigations as necessary.

Because the statistical equivalence test is employed, small sample sizes increase the probability of failing to demonstrate equivalence between products which are in fact equivalent.

4.1.4 Normalization

Head-to-head comparisons are performed based on the HPHC yields per stick, in particular to reduce the additional level of uncertainty brought by normalization of measured values by one another. Nicotine is included to avoid accepting changes which could have an impact on the nicotine yield.

4.1.5 Acceptance Criteria

For each of the endpoints above LOQ in the THS 2.2 monitor data set, the acceptance ranges are defined based on the empirical observation that an instantaneous average – i.e. mean of values measured on the product at a single point in time, are generally within the long-term average plus/minus three long-term standard deviations[4]. Based on this observation, the tolerances for differences of means of two single point in time measurements are defined to be:

$$\text{Tolerance of relative difference in means} = \sqrt{2} * 3 * \text{RSD},$$

where RDS stands for the relative standard deviation of the THS 2.2 monitor data set for the given constituent.

For endpoints where some replicates are below LOQ in the THS 2.2 monitor data set, thresholds are established as follows. When more than 50% of the replicates in the reference data set are below LOQ, the threshold is defined as 4*LOQ. When 50% or less of the replicates in the reference data set are below LOQ, the threshold is defined as 10*LOQ if the value of 10*LOQ guarantees a safety factor 10 as given by means of a calculation by an expert toxicologist; if 10*LOQ does not guarantee a safety factor 10, the limit of m*LOQ (m integer, but in any case no less than m=4) will be taken such that a safety factor of 10 or higher is achieved.

Following the annual empirical assessment of THS monitor test piece for re-assessing the continued applicability of product performance comparability criteria, Mercury[i] is re-classified as secondary endpoint above the limit of quantification, and a variance-based comparability limit is established using the approach described in 11.

The numerical values for the acceptance ranges are provided in **Table 1** below.

Table 1: Acceptance criteria determined based on robust estimation of the Long-term within analytical variability in Labstat, using the results determined from different THS monitor test pieces (P1M0, P1M1, and P1M2) between 2014 and 2018. The statistical evaluation methods and assessment details are described in reference 11.

HPHC		Units	Monitor Mean	Tolerance [%]	Threshold
Basic and product-specific endpoints	Total Particulate Matter (TPM)	mg/cig	**(b) (4)**		
	Nicotine	mg/cig			
	Glycerol	mg/cig			
	Carbon monoxide (CO)	mg/cig			

[i] A review of historical data on the yield of Mercury from THS monitor test piece highlights an increase trend in both the average and the within analytical variability, which is not expected from a product design perspective [12].

HPHC		Units	Monitor Mean	Tolerance [%]	Threshold
Aliphatic dienes	1,3-Butadiene	µg/cig	(b) (4)		
	Isoprene	µg/cig			
Carbonyls	Formaldehyde	µg/cig			
	Acetaldehyde	µg/cig			
	Acetone	µg/cig			
	Acrolein	µg/cig			
	Butyraldehyde	µg/cig			
	Crotonaldehyde	µg/cig			
	Propionaldehyde	µg/cig			
	Methyl ethyl ketone (MEK)	µg/cig			
Acid deriv.	Acetamide	µg/cig			
	Acrylamide	µg/cig			
	Acrylonitrile	µg/cig			
Epoxides	Ethylene oxide	µg/cig			
	Propylene oxide	ng/cig			
Nitro comp.	Nitro-benzene	µg/cig			
Aromatic amines	1-Aminonaphthalene	ng/cig			
	2-Aminonaphthalene	ng/cig			
	3-Aminobiphenyl	ng/cig			
	4-Aminobiphenyl	ng/cig			
	o-Toluidine	ng/cig			
	Benzidine	ng/cig			
N-heterocyclic aromatics	Pyridine	µg/cig			
	Quinoline	µg/cig			
Hal. comp.	Vinyl chloride	ng/cig			
	Nitric oxide (NO)	µg/cig			

HPHC		Units	Monitor Mean	Tolerance [%]	Threshold
Inorganic compounds	Nitrogen oxides (NOx)	µg/cig	**(b) (4)**		
	Hydrogen cyanide (HCN)	µg/cig			
	Ammonia (NH3)	µg/cig			
Monocyclic aromatic hydrocarbons	Benzene	µg/cig			
	Toluene	µg/cig			
	Styrene	µg/cig			
Tobacco-specific nitrosamines	N'-Nitrosoanabasine (NAB)	ng/cig			
	N'-Nitrosoanatabine (NAT)	ng/cig			
	N'-Nitrosonornicotine (NNN)	ng/cig			
	4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK)	ng/cig			
Phenols	Catechol	µg/cig			
	m-Cresol	µg/cig			
	p-Cresol	µg/cig			
	o-Cresol	µg/cig			
	Hydroquinone	µg/cig			
	Phenol	µg/cig			
	Resorcinol	µg/cig			
PAHs	Benzo[a]pyrene	ng/cig			
	Benz[a]anthracene	ng/cig			
	Dibenz[a,h]anthracene	ng/cig			
	Pyrene	ng/cig			
Elements	Arsenic	ng/cig			
	Cadmium	ng/cig			
	Chromium	ng/cig			
	Lead	ng/cig			
	Mercury	ng/cig			

HPHC		Units	Monitor Mean	Tolerance [%]	Threshold
	Nickel	ng/cig	(b) (4)		
	Selenium	ng/cig			

Performance of the test and control products can be deemed comparable if no follow-up investigation is required. A follow-up investigation in a head-to-head comparison is mandatory if:

- The upper limit of the confidence interval of more than 20% of the number of consistuents evaluated for equivalence (reflecting the expected error rate of 5% without multiple-testing corrections, see [5]) exceeds the upper tolerance threshold, or
- The confidence interval of any of the constituents evaluated for equivalence lies completely above the upper tolerance threshold, or
- The maximum measured value of any of the constituents evaluated against thresholds exceeds the established threshold.

If a follow-up investigation is required, comparability of performance may be established if the deviations are determined to be due to analytical sources, or other laboratory or statistical reasons, i.e. if they can be proven to be unrelated to the product change under assessment.

4.1.6 Monitoring and Model Review Frequency

The tolerance ranges are reviewed annually. New tolerance ranges are calculated based on the THS 2.2 monitor data from the previous year. In case the THS 2.2 monitor batch was changed during the previous year, data for the THS 2.2 monitor assessed most extensively is used to determine the tolerances, and data for the other monitor batch(-es) is used to validate the results – such validation consists of a verification that the yields of the other monitor batch(-es) normalized to their respective all-batch mean are within the tolerance range. If validation is not successful, an investigation and separate report is required to finalize tolerances review.

A higher review frequency may be warranted in case of laboratory changes, e.g. changes in analytical methods. In this case, available monitor data on at least three batches of an THS 2.2 monitor assessed post laboratory changes is required to provide new tolerance ranges.

4.2 Agricultural and Analytical Variability – Product Performance Comparability

4.2.1 Overview

In addition to the analaytical method variability over time discussed above, a tobacco product manufactured over an extensive period of time will inevitably exhibit agricultural variability – variations in performance due to crop and manufacturing[ii] variability over time. To empirically evaluate the combined magnitude of analytical and agricultural product performance variability, results from the Blend Maintenance Study were employed. In this study, a total of 40 THS 2.2 batches were manufactured each with its own lot set of (b) (4) tobacco blend, simulating the impact

[ii] Recently, a CORESTA Task Force was established to investigate these sources of variability systematically.

of crop variability over time. The (b) (4) blend variant of THS 2.2 was used for the preclinical and many of the clinical studies, prior to the ERS clinical study.

Each of these 40 batches was assessed for the yields of the PMI-58 list of aerosol constituents. For the comparability model, the objective was to determine the natural range of product performance considering existing sourcing and manufacturing controls. Therefore, not all 40 batches were selected for the model. A subset of 26 batches which satisfied commercial release (performance specifications) criteria were selected as the basis for the comparability model.

The resulting product performance ranges provide acceptance criteria which ensure a direct link to the Investigational Product, i.e. the product variant which underwent the pivotal clinical studies to date.

Assessments against this product performance range are referred to as Product Performance Comparability or Initial Screening, depending on the robustness of the assessment (see sample size). Such assessments may be performed as stand-alone evaluations, particularly suitable for evaluating changes to the tobacco (e.g. blend changes), or in combination with head-to-head assessments to prevent product performance creep over time (i.e. by providing a central pivotal product performance range, to avoid successive head-to-head assessments leading to a systematic increase in the yields of some constituents over time whereby each individual increase would be not quantifiable within the precision of the method).

4.2.2 Statistical Methods and Classification of Endpoints

The performance of a test product is evaluated against the Product Performance Range (PPR). The PPR is defined by a historical reference data set, based on the empirical observation that mean values measured on the product at a single point in time are generally within the long-term average plus/minus three long-term standard deviations[4]. For each endpoint above limit of quantification (LOQ) in the reference data set, the PPR corresponds to the 99th percentile of the reference data set. For endpoints below LOQ, an empirical threshold of up to 10*LOQ is established.

Endpoints are grouped into primary and secondary endpoints. Except where reference data set values for an endpoint are below the limit of quantification, primary endpoints undergo statistical tests, and secondary endpoints are monitored. In turn, primary endpoints are tested either for equivalence (primary equivalent endpoints), or for inferiority (primary non-inferior endpoints). For both primary and secondary endpoints, the PPR defines tolerances – i.e. the maximum magnitude of differences beyond which product performance would no longer be considered comparable.

Classification of endpoints into primary equivalent, primary non-inferior and secondary groups are defined according to the criteria listed below.

4.2.2.1 Primary equivalent endpoints:

Aerosol attributes defined as critical factors determining the sensory properties of the aerosol: Total Particular Matter (TPM), nicotine, and glycerol.

4.2.2.2 Primary non-inferior endpoints[iii]:

Harmful and potentially harmful constituents (HPHCs) considered as carcinogenic or reprotoxic to humans according to harmonized classification under European regulation EC No 1272/2008 or as human carcinogens according to International Agency for Research on Cancer (IARC). In addition recognized mutagens are also considered.

The following HPHCs from the PMI-58 list fall under this category: 1,3-butadiene, Carbon monoxide (CO), Crotonaldehyde, Formaldehyde, Acetaldehyde, Acrolein, 4-(Methyl-nitrosamino)-1-(3-pyridyl)-1-butanone (NNK), N-Nitrosonornicotine (NNN), Acrylamide, Benzene, Benzo[a]pyrene, Propylene oxide, o-toluidine, Hydroquinone, Dibenz[a,h]anthracene, Ethylene oxide, Quinoline, Vinyl chloride, Arsenic, Cadmium, Chromium, Lead, Nickel, 2-Aminonaphthalene, and 4-Aminobiphenyl.

4.2.2.3 Secondary endpoints

All other constituents from the PMI-58.

4.2.3 Sample Size and Statistical Robustness

For the test product variant, the minimum extent of data which should be provided to support a product performance comparability assessment is three replicates on each of three stick batches. This is commonly referred to as a 3x3 testing matrix (3 batches in 3 replicates each), which results in a total of nine replicates. In case of device change assessments, the recommended protocol is nine replicates. For smaller numbers of replicates, the results of the assessment are considered less robust, and the assessment itself is referred to as initial screening.

Because the statistical equivalence test is employed, small sample sizes increase the probability of failing to demonstrate equivalence between products which are in fact equivalent.

4.2.4 Normalization

Assessments are performed based on the HPHC yields per stick, in particular to reduce the additional level of uncertainty brought by normalization of measured values by one another. Nicotine is included to avoid accepting products which could have a nicotine yield beyond the range observed in the reference product.

4.2.5 Acceptance Criteria

Statistical assessment principles depend both on endpoint classification as well as on measurability of the endpoints. The following acceptance criteria were defined for the individual endpoints:

4.2.5.1 Primary equivalent endpoints

The tolerance limits are defined as the 99% tolerance interval of the reference data set, taking into account the Orchard correction[6]. The acceptance criteria for a test product are that the 90% confidence interval on its mean yields are entirely contained within the tolerance limits. LTL – lower tolerance limit; UTL – upper tolerance limit.

Endpoint	Units	Target	LTL	UTL

[iii] Non-inferiority of performance refers in this case to non-superiority of yields; for historical reasons and consistency with early reports, the terminology is maintained.

Total Particulate Matter	mg/stk	(b) (4)
Nicotine	mg/stk	
Glycerol	mg/stk	

4.2.5.2 Primary non-inferior endpoints

For endpoints with reference data set values all above their limit of quantification, the upper tolerance limit is defined as the upper limit of the 99% tolerance interval of the reference data set, taking into account the Orchard correction [6]. The acceptance criteria for a test product endpoint is that the 95% confidence interval on its mean yield is entirely below the upper tolerance limit. UTL – upper tolerance limit.

Endpoint	Units	Target	UTL	Endpoint	Units	Target	UTL
Carbon monoxide (CO)	mg/stk	(b) (4)		Benzene	µg/stk	(b) (4)	
1,3-Butadiene	µg/stk			N-Nitrosonornicotine (NNN)	ng/stk		
Formaldehyde	µg/stk			4-(Methyl-nitrosamino)-1-(3-pyridyl)-1-butanone (NNK)	ng/stk		
Acetaldehyde	µg/stk			Benzo[a]pyrene	ng/stk		
Acrolein	µg/stk			Propylene oxide	ng/stk		
Acrylamide	µg/stk			o-Toluidine	ng/stk		
Hydroquinone	µg/stk						

For endpoints with reference data set values below the limit of quantification (LOQ), the acceptance criteria are such that the maximum measured value in the test item is below a threshold. The threshold has generally been empirically defined as four times the LOQ. Limitations of this approach have been identified when 50% or less of the values in the reference data set are below LOQ. Consequently, where more than 50% of the values in the reference data set are below LOQ, the threshold is retained at 4*LOQ; if 50% or less of the values in the reference data set are below LOQ, the threshold is established at 10*LOQ provided a toxicological safety factor of 10 is guaranteed, otherwise at n*LOQ where n is such that a toxicological safety factor of 10 is guaranteed, but in any case not lower than 4*LOQ. Subsequently the criteria for primary endpoints below LOQ are:

Endpoint	Units	Threshold	Endpoint	Units	Threshold
2-Aminonaphthalene	ng/stk	(b) (4)	Dibenz[a,h]anthracene	ng/stk	(b) (4)
4-Aminobiphenyl[7]	ng/stk		Arsenic	ng/stk	
Crotonaldehyde	µg/stk		Cadmium	ng/stk	
Ethylene oxide	µg/stk		Chromium	ng/stk	
Quinoline	µg/stk		Lead	ng/stk	
Vinyl chloride	ng/stk		Nickel	ng/stk	

4.2.5.3 Secondary endpoints

For endpoints with reference data set values all above their limit of quantification, the threshold is defined as the upper limit of the 99% tolerance interval of the reference data set, taking into account the Orchard correction [6]. The acceptance criteria for a test product are that the point estimate of the mean yield is below the threshold, or that at least one replicate is below LOQ.

Endpoint	Units	Target	Threshold	Endpoint	Units	Target	Threshold
Butyraldehyde	µg/stk	(b) (4)		Styrene	µg/stk	(b) (4)	
Pyridine	µg/stk			N-nitrosoanabasine (NAB)	ng/stk		
Phenol	µg/stk			N-nitrosoanatabine (NAT)	ng/stk		
Acetone	µg/stk			Catechol	µg/stk		
Propionaldehyde	µg/stk			o-Cresol	µg/stk		
Nitric oxide (NO)	µg/stk			Ammonia	µg/stk		
Benz[a]anthracene	ng/stk			Toluene	µg/stk		
Pyrene	ng/stk			Acetamide	µg/stk		
Nitrogen oxides (NOx)	µg/stk			Mercury	ng/stk		
Isoprene	µg/stk						

For endpoints with reference data set values below the limit of quantification (LOQ), the acceptance criteria are such that the maximum measured value in the test item is below a threshold. The threshold has generally been empirically defined as four times the LOQ. Limitations of this approach have been identified when 50% or less of the values in the reference data set are below LOQ. Consequently, where more than 50% of the values in the reference data set are below LOQ, the threshold is retained at 4*LOQ; if 50% or less of the values in the reference data set are below LOQ, the threshold is established at 10*LOQ provided a toxicological safety factor of 10 is guaranteed, otherwise at n*LOQ where n is such that a toxicological safety factor of 10 is guaranteed, but in any case not lower than 4*LOQ. Subsequently the criteria for secondary endpoints below LOQ are:

Endpoint	Units	Threshold	Endpoint	Units	Threshold
1-Aminonaphthalene	ng/stk	(b) (4)	Resorcinol	µg/stk	(b) (4)
3-Aminobiphenyl[7]	ng/stk		Selenium	ng/stk	
Nitro benzene	µg/stk		Methyl-ethyl-ketone (MEK)	µg/stk	
Hydrogen cyanide	µg/stk		Acrylonitrile	µg/stk	
m-Cresol	µg/stk		p-Cresol	µg/stk	

4.2.6 Product comparability acceptance criteria

In a product performance comparability assessment, product performance is considered comparable if no follow-up investigation is required.

In an initial screening, product performance is considered more likely comparable than not if no follow-up investigation is required.

An investigation is required for a product performance comparability assessment if any evaluated endpoint does not satisfy the acceptance criteria.

An investigation is mandatory for an initial screening assessment if:

- The upper limit of the confidence interval of more than 20% of the number of consistuents evaluated for equivalence or non-inferiority (reflecting the expected error rate of 5% without multiple-testing corrections, see [5]) exceeds the upper tolerance threshold, or
- The confidence interval of any of the constituents evaluated for equivalence or non-inferiority lies completely above the upper tolerance threshold, or
- The maximum measured value of any of the constituents evaluated against thresholds exceeds the established threshold.

4.2.7 Monitoring and Model Review Frequency

The model or acceptance criteria are only reviewed if systematic THS 2.2 monitor trends are identified which demonstrate limitations of the established approach or evaluated magnitude of analytical variability, as part of a documented investigation.

In case of assessment requirement changes which require the establishment of new comparability models (e.g. change of laboratory, re-baselining to define a new reference product, etc), it is remarked that if data is collected on a test product variant before a new PPR is established, the test product data can be used retrospectively for comparability, as long as the calculation of PPR follows an identical procedure as previously, or an alternative set of criteria are documented prior to receipt of data on the test product variant.

5 Comparability Model *In Vitro* Neutral Red Uptake Cytotoxicity

The overall rationale for *in vitro* assessment follows a parallel testing design, between the test items under scrutiny and a control sample (combination of THS monitor test piece and the reference tobacco heating device (THD)). The *in vitro* data assessments consist of assessing, the ratio of average values between the control item and the test item, versus the upper tolerance limit.

The control samples are assessed in parallel with the test items, to ensure method performance adequate for comparability assessment during the project. THS monitor *in vitro* results shall be continuously assessed to characterize and monitor the analytical variability. Systematic trends in the control sample yields can be used to re-assess the continued applicability of product performance comparability criteria.

Currently, based on data availability, comparability criteria have been established for the neutral red uptake (NRU) cytotoxicity assay, total particulate matter (TPM) and gas-vapour phase (GVP) endpoints. Studies to assess the comparability criteria for the Mouse Lymphoma Assay (MLA) and Ames bacterial mutagenicity assay are ongoing, and the endpoint acceptance criteria will be established based on the same standard methodology for those *in vitro* endpoints for which a dose-response relationship is observed in the study.

5.1 Product Performance Comparability

The NRU results shall be obtained for at least one aerosol collection on each of the three stick production batches of the test item. The control test piece shall be analysed in parallel within the same analytical sequence as the test item. The assessment steps are summarized as follow:

1. Calculate within day ratios of EC_{50}, mL/article (Control item/Test item), for the TPM and GVP fractions.
2. The average ratios of EC_{50}, mL/article (Control item/Test item) between days shall then be assessed versus on upper limits, established at 1.40 for the GVP fraction, and 1.20 for the TPM fraction. The test item is released if and only if the average EC_{50} is strictly below the upper tolerance limits for GVP and TPM fractions.

6 Reference Documents

1. *PMI-RRP-SOP-111979 - Change Management Board (RRPCE)*
2. *PMI-RD-SOP-112377 Product Comparability Assessment*
3. COF_REV_2011.00007 Chemical Assessment of Modified Tobacco Products Mainstream Aerosols (MMA)
4. PMI-RRP-SOP-111919 - RRP Specifications Management (RDNEU)
5. Considerations for comparative tobacco product assessments based on smoke constituent yields, M. Belushkin, G. Jaccard, A. Kondylis, Regulatory Toxicology and Pharmacology 73 (1), 105-113, 2015.
6. Orchard, T. (2006). "Setting acceptance criteria from statistics of the data." BioPharm International 19(11): 34-46.
7. Review of Platform 1 Comparability Limits for Aromatic Amines in Labstat Data, Product Technical Assessment, Version 1.0 issued 18th August 2015, electronic copy available at **(b) (4)**
8. Annual review of Platform 1 analytical variability acceptance criteria for PMI-58 list of constituents in Labstat Data, based on 18 sets of data of 9A026. Assessment permanent records are available **(b) (4)**
9. Annual review of Platform 1 analytical variability acceptance criteria for PMI-58 list of constituents measured in Labstat. Data based on 16 sets of data of THS2.2 monitor. Assessment permanent records are available **(b) (4)**
10. Annual review of Platform 1 analytical variability acceptance criteria for PMI-58 list of constituents measured in Labstat. Data based on 33 testing projects of THS2.2 monitor. Assessment report is available at **(b) (4)**
11. Review of THS Comparability Limits to accounting for the actual Long-term analytical variability in Labstat. Electronic copy available at **(b) (4)**
12. Interim Investigation Status Report - Mercury Out of Trend from Product Performance Comparability Assessment. Electronic copy available at **(b) (4)**

7 Version History

Version	Changes
1.1	Annual review and update of head-to-head acceptance criteria.
3.1	Annual review and update of head-to-head acceptance criteria.
4	Annual review and update of head-to-head acceptance criteria.
4.1	Review the NRU acceptance criteria.

5.0	Update scope 1: during the annual review of analytical variability, it was noted that for some endpoints the method variability exceeds the total variability in the comparability model. This confirmed that over the years since the initial blend maintenance study, laboratory performance has not remained exactly the same, and the updated method variability was incorporated into the acceptance criteria [11]. Update scope 2: the original classification of endpoints was based on measurability and statistical correlation. In order to better align with the assessment escalation process (aerosol chemistry to *in vitro*) the classification of endpoints was updated to be based on toxicologically relevant criteria. Primary non-inferior endpoints are substances harmonized classified as carcinogenic or mutagenic, and all other substances are secondary endpoints.
5.1	Minor update to correct a cross-reference typo error.

8 Review

Version 1.0 of the present white paper was reviewed by:

Name	Function
Maxim Belushkin, Manager Scientific & QA Governance	Author, Change Management Board appointed Statistician
Shelley Moores, Manager QA Assessment & Quality Systems	Subject Matter Expert
Marco Esposito, Director Product Stewardship	Owner & Approver

Version 1.0 of the present white paper summarized and clarified, without major content changes, statistical methods and acceptance criteria for aerosol chemistry assessments from the following documents:

- "Reduced Risk Product Heated Tobacco Platform Comparative Product Performance (Aerosol) Evaluations" version 1.0 approved 30th March 2015 (Product Technical Assessment), reviewed by the Scientific Council, Chair Change Management Board, Regulatory Operations, Toxicological & Preclinical Product Evaluation, Product Assessment and Scientific Substantiation.

- "Product Performance Comparability Criteria Platform 1" version 2.0 approved 19th August 2015 (Product Technical Assessment), reviewed by the Scientific Council, Toxicological Evaluation, Chair Change Management Board, Regulatory Operations, Clinical Safety.

- "Head-to-Head Comparisons for Evaluation of Potential Impact of RRP Changes on Aerosol Chemical Composition" version 3.0 approved 26th March 2015 (Product Technical Assessment), reviewed by Etienne Kaelin as statistics reviewer, Toxicological Evaluation, Scientific Council, Chair Change Management Board, Regulatory Operations, Clinical Safety.