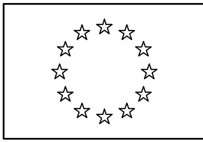


EUROPEAN COMMISSION
JOINT RESEARCH CENTRE

Institute for Health and Consumer Protection
In-Vitro Methods Unit
European Centre for the Validation of Alternative Methods (ECVAM)

Performance Standards for *In-Vitro* Skin Irritation Test Methods based on Reconstructed Human Epidermis (RhE)

European Centre for the Validation of Alternative Methods (ECVAM)

IMPORTANT NOTE FOR MODEL PRODUCERS SEEKING VALIDATION:

As outlined in section 1.3 of this document, only test methods that comply with Element 1 of the PS (the essential test method components, section 3.1) to an extent that they may be considered *similar* to the validated reference method qualify for an equivalence validation on the basis of this PS and are hereunder referred to as "RhE models". Such test methods are either new test methods that are found to be sufficiently similar to an already validated one ("me-too" methods) or modifications of validated methods. Criteria for similarity concern the biological relevance (including criteria for structural, functional, mechanistic similarity) as well as procedural similarity with regard to the validated reference method. It is therefore strongly recommended that test producers contact ECVAM with regard to an evaluation whether or not their test method qualifies for an equivalence validation study before embarking on external ring trials for equivalence validation testing in accordance with the standards and guidance provided in this document.

Moreover, while the Performance Standards outline defined test acceptance criteria, some of the parameters relating to these criteria may be adapted if sufficiently scientifically/statistically justified (e.g. the number of tissue replicates to be used, see **§ 12**). It is strongly recommended that test producers intending on adapting specific parameters, contact ECVAM with regard whether the adaptation is acceptable before embarking on an external ring trial.

Finally, it is explicitly noted that any test methods validated by ECVAM is expected to be available to all users without restrictions. To this end, ECVAM reserves the right to request declarations on commercial availability from test method producers seeking validation of their test method by ECVAM.

36 # TABLE OF CONTENTS

37

38 **ACKNOWLEDGEMENTS 3**

39 **INTRODUCTORY NOTE TO THIS DOCUMENT 4**

40 **1. BACKGROUND ON PERFORMANCE STANDARDS 6**

41 **1.1** PURPOSE OF PERFORMANCE STANDARDS 6
42 **1.2** GENERAL ELEMENTS OF PERFORMANCE STANDARDS (PS) 7
43 **1.3** STRUCTURE OF THE PERFORMANCE STANDARDS FOR *IN VITRO* SKIN IRRITATION TESTING BASED ON
44 RECONSTRUCTED HUMAN EPIDERMIS (RHE) 7

45 **2. *IN VITRO* SKIN IRRITATION TESTING BASED ON RHE TECHNOLOGY 9**

46 **2.1** USE AND LIMITATIONS OF *IN VITRO* RHE TEST METHODS FOR ASSESSING SKIN IRRITATION 9
47 ***2.1.1 Use of RhE-based test methods*** *9*
48 ***2.1.2 Known applications and limitations of the test methods*** *9*
49 **2.2** SCIENTIFIC BACKGROUND TO THE SKIN IRRITATION PERFORMANCE STANDARDS BASED ON RHE 10
50 **2.3** SCIENTIFIC PRINCIPLES OF *IN VITRO* RHE FOR SKIN IRRITATION TESTING 10
51 **2.4** PROCEDURE FOR SKIN IRRITATION TESTING USING RHE 11

52 **3. *IN VITRO* PERFORMANCE STANDARDS BASED ON RHE 11**

53 **3.1** ELEMENT 1: ESSENTIAL TEST METHOD COMPONENTS 11
54 ***3.1.1 General Model Conditions*** *11*
55 ***3.1.2 Functional Model Conditions*** *11*
56 ***3.1.3 Application of Test and Control Substances*** *12*
57 ***3.1.4 Cell Viability Measurements*** *13*
58 ***3.1.5 Organising and Conducting Tests*** *13*
59 ***3.1.6 Interpretation of Results and Prediction Model*** *14*
60 ***3.1.7 Test Acceptance Criteria*** *15*
61 **3.2** ELEMENT 2: LIST OF REFERENCE CHEMICALS 16
62 ***3.2.1 Selection Criteria*** *16*
63 ***3.2.2 Proposed Reference Chemicals*** *16*
64 **3.3** ELEMENT 3: TARGET VALUES FOR RELIABILITY AND PREDICTIVE CAPACITY (ACCURACY) 19
65 ***3.3.1 Study Acceptance Criteria for the Purpose of Equivalence Validation Studies Relevant for***
66 ***Assessment of Reliability and Relevance*** *19*
67 ***3.3.2 Rules for Calculation of Reliability (Reproducibility) and Predictive Capacity (Accuracy)*** *20*
68 ***3.3.3 Reliability*** *20*
69 ***3.3.4 Predictive Capacity (Accuracy)*** *21*
70 **3.4** ADDITIONAL GUIDANCE 23
71 ***3.4.1 Test Reporting*** *23*
72 ***3.4.2 Complementary Endpoints*** *24*

73 **ANNEXE I – CONDUCT OF EQUIVALENCE STUDIES 25**

74 **ANNEXE II – PROFICIENCY TESTING 27**

75 **ANNEXE III – GUIDANCE ON QUALIFICATION CRITERIA FOR TEST SUBSTANCES 29**

76

ACKNOWLEDGEMENTS

ECVAM would like to acknowledge all those who contributed to initial draft and the present update of the ECAVM Performance Standards (PS) for *in vitro* skin irritation testing based on Reconstructed human Epidermis. This includes all members of the Validation Management Group of the ECVAM skin irritation validation study (SIVS), conducted from 2003 to 2006 and, with regard to the updated version, all participants to the OECD expert meeting on the draft OECD Test Guideline, held from 22 to 23 October 2008 at the German Federal Institute for Risk Assessment (BfR) in Berlin, Germany, as well as the members of the ESAC working group (February to April 2009) dedicated to finalise this updated version.

In particular ECVAM would like to acknowledge the contributions of:

Original ECVAM Performance Standards (May 2007)

- Manfred Liebsch, ZEBET, BfR, Berlin, Germany
- Horst Spielmann, ZEBET, BfR, Berlin, Germany
- Helena Kandárová, MatTek Corporation, Ashland, USA
- Roland Roguet, L'Oréal, Paris, France
- José Cotovio, L'Oréal, Paris, France
- Bart de Wever, Phenion, Düsseldorf, Germany
- Sebastian Hoffmann, ECVAM, European Commission DG JRC, Ispra, Italy
- Julia Fentem, Unilever, Sharnbrook, UK
- Penny Jones, Unilever, Sharnbrook, UK
- Valérie Zuang, ECVAM, European Commission DG JRC, Ispra Italy

Adaptations of the ECVAM Performance Standards (May 2009)

<u>Commission staff:</u>

- Claudius Griesinger, ECVAM, European Commission DG JRC, Ispra, Italy
- João Barroso, ECVAM, European Commission DG JRC, Ispra, Italy
- Valérie Zuang, ECVAM, European Commission DG JRC, Ispra, Italy

<u>ESAC working group:</u>

- Manfred Liebsch, ZEBET, BfR, Berlin, Germany (chair)
- Helena Kandárová, MatTek Corporation, Ashland, USA (industry representative)
- José Cotovio, L'Oréal, Paris, France (industry representative)
- Nathalie Alépée, L'Oréal, Paris, France (industry representative)
- Patric Amcoff, OECD, Paris, France (acting as observer)
- William Stokes, NICEATM, Bethesda, USA (acting as observer)

112

113 # INTRODUCTORY NOTE TO THIS DOCUMENT

114 The present Performance Standards (PS) represent an updated version of the original ECVAM PS defined
115 in May 2007 [1] after completion of the ECVAM Skin Irritation Validation Study (SIVS) conducted
116 between December 2003 to August 2006 [2]. During the SIVS, the reliability, relevance and applicability
117 domain of two commercially available Reconstructed human Epidermis (RhE) models (EpiSkin^{TM} and
118 EpiDerm^{TM}) were analysed. The SIVS was designed and conducted prior to the adoption of the United
119 Nations (UN) Globally Harmonized System of Classification and Labelling of Chemicals (GHS) [3].
120 Consequently, the SIVS evaluated the test methods under scrutiny primarily with respect to the EU
121 classification system as described in the Dangerous Substances Directive (the “EU DSD” system) [4],
122 albeit considering the GHS classification system during selection of substances to be tested in the SIVS.
123 Thus, the original PS, including both the list of *Reference Chemicals* and the *Accuracy Target Values,*
124 were based on the EU DSD [4], which consists of two categories: *no label* (non-classified substances) and
125 *R38* (irritant substances) with a cut-off *in vivo* score of 2.0 (Fig. 1).

126 In December 2008 the EU adopted the UN GHS [3] and implements this by means of the Classification,
127 Labelling and Packaging (CLP) Regulation [5]. This regulation came into force on 20 January 2009 and
128 will replace, after a transitional period, the previous EU legislations [4] for the classification of substances
129 and mixtures (i.e. preparations). The EU classification system based on GHS (the "CLP" system) [5]
130 directly transposes the UN GHS system [3] which foresees one irritant category. The EU will not use an
131 additional optional category for mild irritants ("Category 3") that will apply only to some authorities (e.g.
132 pesticides) (UN GHS*). Therefore the CLP system continues to use two categories to distinguish *non-*
133 *classified* (No Category) from irritant (Category 2) substances. However, according to the new rules for
134 skin irritation classification and labelling (C&L) [5], the cut-off score to distinguish between No Category
135 and Category 2 substances was shifted to 2.3 (UN GHS or CLP) from a value of 2.0 (EU DSD).
136 Consequently substances with an *in vivo* score between 2.0 and 2.3 that were considered irritant under the
137 EU DSD are now non-classified under UN GHS, which does not use the optional Category 3 (Fig. 1).

138 This had practical consequences on the ECVAM PS:

139 (a) the set of Reference Chemicals (RC) was not balanced any more (three former R38 substances had
140 become not classified under UN GHS) [6] and although this can be regarded as reflecting the real
141 prevalence of irritants much better, it is good practice to have a balanced distribution of RC enabling
142 assessment of both classified (irritant) and non-classified substances on the basis of equal numbers of test
143 substances;

144 (b) the accuracy target values did not match the changed prevalence which results from the cut-off shift
145 [6]: with a higher cut-off, more substances will not be classified in the future and, inversely, the
146 prevalence of skin irritant substances will decrease.

147 Therefore, regulation EC 1272/2008 (CLP regulation) [5] made necessary an **update of the original**
148 **ECVAM PS** [1] in order to balance the set of RC and carefully adjust the accuracy target values [6].
149 Minor adaptations include more precise specifications concerning:

150 1) Recommendations regarding the training set for developing similar or modified test methods that
151 may qualify for PS-based equivalence validation studies, in particular limitations regarding the
152 use of RC for test development/optimisation purposes.

153 2) The number of times that invalid runs may be retested.

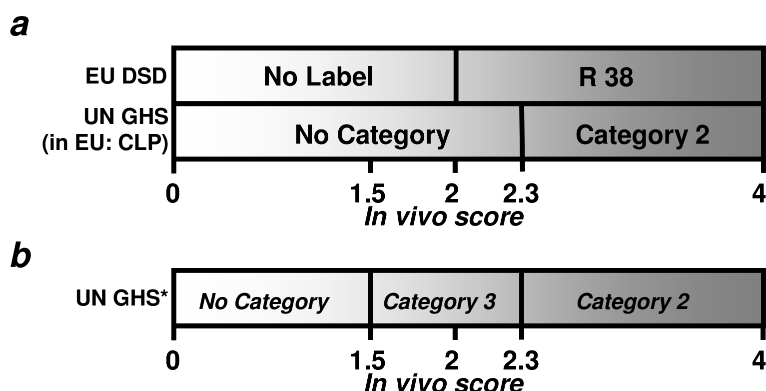
154 3) The number of invalid run sequences (i.e. absence of 3 valid independent runs in a single
155 laboratory) after retesting that are acceptable for the data set to be considered qualified for the
156 purpose of an equivalence validation study.

157 4) The calculation of Reliability (Reproducibility) and Predictive Capacity (Accuracy)

158 The rationale and details of the amendment of the PS are documented in an ECVAM Background
159 Document [6], which also provides an overview on the performance of the three ECVAM-validated full-
160 replacement *in vitro* skin irritation test methods (the EpiSkin™, the EpiDerm™ SIT and the SkinEthic™
161 RHE test methods) [7, 8] under the UN GHS/CLP system for skin irritation C&L [3, 5, 9].

162

163 ***Figure 1:** (**a**) Comparison of the threshold values for skin irritation classification and labelling in the EU*
164 *prior to adoption of the UN GHS system (abbreviated as EU DSD = European Union classification*
165 *system as described in the Dangerous Substances Directive) and after the adoption of the UN GHS system*
166 *as applicable to all authorities (UN GHS). In the EU the UN GHS system is implemented through the so-*
167 *called CLP regulation (CLP). While the cut-off score of the EU DSD was 2.0, the cut-off score of the UN*
168 *GHS (CLP) system is 2.3. Therefore, in the EU, all substances with scores from 0 to 2.3, previously*
169 *considered irritant, will be 'no category' under UN GHS (CLP) if tested in the in vivo test system. (**b**) The*
170 *UN GHS system as applicable to some authorities (UN GHS*) featuring an optional irritant category.*
171 *According to the provisions of the GHS health hazard document, this optional Category 3 is available for*
172 *"those authorities that want to have more than one skin irritant category" [3]. This optional additional*
173 *category is not implemented in the EU. A score of 0 means that the substance did not produce visible*
174 *signs of irritation on rabbit skin. A score of 4 means that the substance produced the strongest possible*
175 *visible response for one or two of the endpoints evaluated. This transition from non- irritant to irritant is*
176 *indicated by the gradient from white to grey.*



177

178

179

180

181

182 **N.B.**

183 ***Although PS should be used as a tool for an evidence-based evaluation of the potential equivalence of***
184 ***test methods in comparison to validated ones using pre-defined criteria, it should be noted that this***
185 ***approach is not intended to represent a mere "tick-box approach", but a tool for the transparent,***
186 ***conscientious and consistent assessment of potential equivalence. Consequently, it is conceivable that***
187 ***deviations from these standards may be justified by scientific reasons. It is required that any such***
188 ***deviations be recorded and explained in the summary report of the validation study.***

189

1. BACKGROUND ON PERFORMANCE STANDARDS

1.1 Purpose of Performance Standards

PS may be used for

1) PS-based equivalence validation studies concerning (a) similar and (b) modified test methods.
2) for the assessment of the performance of test methods without intending on formal validation

Equivalence validation studies

Prior to the acceptance of new test methods for regulatory testing, validation studies are conducted to assess the **reliability** (i.e., the extent of intra- and inter-laboratory reproducibility) and **relevance** of a test method (i.e., the method's ability to correctly predict or measure the biological effect of interest) [10, 11].

In this context PS are used to evaluate the potential **equivalence** of proposed test methods to previously validated methods through small-scale validation studies performed in explicit reference to PS. Such validation studies are referred to as **equivalence validation studies**. PS provide a set of pre-defined *evaluation* and *acceptance* criteria which should be met by the similar or modified test methods to be considered ***equivalent*** to the original validated method(s) on which the PS are based on. As indicated by the name, such studies intend to establish whether a proposed test method may be considered **equivalent** to an already validated test method, when considering its ability to predict possible adverse effects of xenobiotics on human health in a reliable manner.

Methods that may undergo such validation studies must be based on similar (i.e. sufficiently analogous) scientific principles as a reference method validated in a full prospective validation study and measure or predict the same biological or toxic effect.

Methods qualifying for such **equivalence validation studies** on the basis of PS are:

(a) **Similar methods:** test methods proven to be similar to the validated reference method(s) that were used to define PS. These are colloquially also referred to as "me-too" tests. The validation process is referred to as "catch-up validation study".

(b) **Modified methods:** modifications of test methods that were either validated in full prospective validation study before the definition of PS or modifications of similar methods. The validation process is referred to as "update validation study"

Assessment of methods without intending on formal validation

In addition, without intending on formal validation, PS may be used as a set of reference specifications and evaluation/acceptance criteria for the critical evaluation of the **performance** of a novel test method by several laboratories. Thus, reliability must comprise evaluations of both, within- and between laboratory reproducibility.

In summary, the purpose of Performance Standards (PS) is to communicate the basis by which proposed test methods, both proprietary (*i.e.*, copyrighted, trademarked, registered) and non-proprietary, can be determined to have sufficient reliability and relevance for specific testing purposes [11].

229

230 ## 1.2 General Elements of Performance Standards (PS)

231 The evaluation and acceptance criteria or PS are structured in three main areas that are referred to as PS
232 elements:

233 **Element 1. Essential Test Method Components:**
234 These consist of essential structural, functional, and procedural elements of a validated test method that
235 should be included in the protocol of a proposed, mechanistically and functionally similar test method.
236 These components include unique characteristics of the test method, critical procedural details, and
237 quality control measures. Adherence to essential test method components will help to assure that a
238 proposed test method is based on the same concepts as the corresponding validated test method.

239 **Element 2. List of Reference Chemicals:**
240 These are used to assess the accuracy and reliability of a proposed, mechanistically and functionally
241 similar test method. These chemicals are a representative subset of those used to demonstrate the
242 reliability and the accuracy of the validated test method. These Reference Chemicals (RC) are the
243 minimum that should be used to evaluate the performance of a proposed, mechanistically and functionally
244 similar test method. If any of the recommended chemicals is unavailable, other chemicals for which
245 adequate reference data are available could be used instead. To the extent possible, the substitute
246 chemical(s) should be of the same chemical class and activity as the original chemical(s). If desired,
247 additional chemicals representing other chemical or product classes and for which adequate reference data
248 are available can be used to more comprehensively evaluate the proposed test method. However, these
249 additional chemicals should not include any that have been used to develop the proposed test method.

250 **Element 3. Target Values for Reliability and Predictive Capacity (Accuracy):**
251 These are the performance requisites that should be achieved by the proposed test method when evaluated
252 using the minimum list of RC, i.e. reliability and predictive capacity that should be achieved by the
253 proposed test method when testing the RC.

254 ## 1.3 Structure of the Performance Standards for *In Vitro* Skin Irritation Testing Based
255 ## on Reconstructed Human Epidermis (RhE)

256 The PS are specifically based on the *in vitro* technology of *Reconstructed human Epidermis* (RhE).
257 Primary reference for these PS is the EpiSkin^{TM} test method that was validated as a full replacement
258 method during the ECVAM Skin Irritation Validation Study (SIVS) [2]. This method is hereunder
259 referred to as the Validated Reference Method (VRM). However, the provisions of these PS are to some
260 extent also based on the second method analysed during the SIVS, the EpiDerm^{TM} test method, which was
261 validated for identifying positives and for use within a testing strategy as outlined by the OECD TG 404
262 [12]. Thus, while Elements 1 (Essential Test Method Components) and 2 (List of Reference Chemicals)
263 are based on both test methods, Element 3 (Target Values for Reliability and Predictive Capacity) is based
264 mainly on the VRM.

265 In reference to these two test methods and in particular to the VRM, these PS provide:

266 a) A **test procedure** for *in vitro* skin irritation testing based on structural and functional specifications.
267 This procedure is listed in **§ 1 to 17** and thus is identical with the PS Element 1: Essential Test Method
268 Components.

269 b) **Evaluation and Acceptance Criteria** for the assessment and validation of proposed test methods and
270 following the test procedure.

271

272

273 These criteria are laid out in

274 **§ 1 to 17** PS Element 1: Essential test method components that the proposed test method
275 must show in order to *qualify* for a PS-based equivalence validation study.
276 **§ 18** PS Element 2: Reference Chemicals that must be used as test substances for the
277 evaluation of Reliability and Predictive Capacity (**§ 19 to 22**)
278 **§ 19 to 25** PS Element 3: Target values concerning Reliability and Predictive Capacity
279 (Accuracy) to be reached using the RC stipulated in **§ 18**.

280 Moreover, the following paragraphs provide further useful guidance on:

281 **§ 26 to 27** Test Reporting

282 **§ 28** Complementary endpoints

283

284

285 286 # 2. *IN VITRO* SKIN IRRITATION TESTING BASED ON RhE TECHNOLOGY

287 ## 2.1 Use and Limitations of *In Vitro* RhE Test Methods for Assessing Skin Irritation

288 ### 2.1.1 *Use of RhE-based test methods*

289 290 291 Skin irritation refers to the production of reversible damage to the skin following the application of a test substance for up to 4 hours [as defined by the United Nations (UN) Globally Harmonized System of Classification and Labelling of Chemicals (GHS)] [3].

292 293 294 295 * *In vivo* testing: skin irritation can be assessed *in vivo* by applying the test substance in a single dose to the skin of an experimental animal; untreated skin areas of the test animal serve as the control. The degree of irritation is read and scored at specified intervals in order to provide a complete evaluation of the effects.

296 297 298 299 * *In vitro* testing: skin irritation can be assessed *in vitro* by applying test substances to the surface of Reconstructed human Epidermis (RhE) models. Prevalidation and validation studies [2, 7-9, 13-20] have reported that such *in vitro* tests are able to reliably discriminate between classified (irritant) and non-classified substances.

300 301 302 303 For a full evaluation of local skin effects after single dermal exposure, it is recommended to follow the sequential testing strategy as appended to OECD Test Guideline 404 [12]. This testing strategy includes the conduct of *in vitro* tests for skin corrosion [21] and skin irritation (as described in this document) before considering the necessity of any exceptional or complimentary testing in living animals.

304

305 ### 2.1.2 *Known applications and limitations of the test methods*

306 307 308 Since 58 carefully selected substances representing a wide spectrum of chemical classes were included in the SIVS, RhE test methods are expected to be generally applicable across chemical classes. The following specifications and limitations of the applicability domain are known at present:

309 310
- RhE allows distinguishing classified (irritant) mono-, and multi-constituent substances from non-classified ones.

311 312
- It does not provide adequate information on skin corrosion, nor does it allow the sub-categorization of mild irritant substances as defined in the UN GHS (optional Category 3) [3].

313 314 315
- RhE is applicable to solids, liquids, semi-solids and waxes. The liquids may be aqueous or non aqueous; solids may be soluble or insoluble in water. Whenever possible, solids should be ground to a fine powder before application; no other prior treatment of the sample is required.

316 317 318 319 320
- Gases, aerosols, formulations and preparations have not been assessed yet in a validation study. While it is conceivable that these can be tested using RhE technology, the current validated SOPs [22-24] do not allow testing of gases and aerosols. It should also be noted that highly coloured chemicals, *e.g.,* hair dye components, may interfere with the cell viability measurements and need the use of adapted controls for corrections.

321

322 ## 2.2 Scientific Background to the Skin Irritation Performance Standards based on RhE

323 After an ECVAM pre-validation study [15] and following test optimization phases [13, 17-19], the
324 ECVAM Skin Irritation Validation Study (SIVS) was conducted and completed between December 2003
325 and August 2006 [2]. From the five *in vitro* methods initially evaluated, only the EpiSkin™ and
326 EpiDerm™ skin irritation test methods went through the whole pre-validation and validation process.
327 Both test methods use RhE technology. A common test protocol [13, 18] was used for both RhE models,
328 and both test methods provided sufficient intra- and inter-laboratory reproducibility [2, 16]. In addition,
329 the EpiSkin™ test method showed sufficient sensitivity and specificity as a potential full replacement
330 method to reliably identify non-classified and classified (irritant) substances. The EpiDerm™ test method,
331 as validated in the SIVS [2], was proposed to be useful for the identification of skin irritants within the
332 framework of a tiered testing strategy as described in the annex of OECD TG 404 [12]. With the
333 completion of the validation study, the Performance Standards (PS) were defined.

334 Since completion of the PS in 2007 [1], further skin irritation test methods based on RhE have been
335 validated in November 2008 as full replacement methods on the basis of the original PS document:

336 1) The EpiDerm™ SIT test method (modification of the EpiDerm™ test method validated in
337 SIVS), validated in the context of an external update validation study [8]

338 2) The SkinEthic™ RHE test method (similar to VRM), validated in the context of an external
339 catch-up validation study [8]

340 Notably, all three test methods validated up to now as full replacements use in their SOPs [22-24] the
341 same post-treatment incubation period (42 hours) and deviate mainly with respect to two important
342 procedural aspects:

343 1) The **acute exposure time** ranges from 15 (EpiSkin™), over 42 (SkinEthic™ RHE) to 60 minutes
344 (EpiDerm™ SIT).

345 2) The exposure time **incubation temperature** is room temperature (~20°C) for EpiSkin™ and
346 SkinEthic™ RHE. For EpiDerm™ SIT, 35 minutes of the full 60 minutes contact time are performed at
347 37°C and the remaining time at room temperature.

348 These exposure conditions are optimised for each assay and are likely to reflect the different intrinsic
349 properties of the RhE models. The three validated full-replacement test methods mentioned above, i.e. the
350 EpiSkin™, the EpiDerm™ SIT and the SkinEthic™ RHE, are included in the EU Test Method B.46 as
351 well as in the OECD draft Test Guideline for skin irritation testing.

352 ## 2.3 Scientific Principles of *In Vitro* RhE for Skin Irritation Testing

353 The principle of an *in vitro* RhE irritation test method is based on the premise that irritant substances are
354 able to penetrate the *stratum corneum* by diffusion and are cytotoxic to the cells in the underlying layers.

355 The test substance is applied topically to a three-dimensional human RhE model, comprised of at least a
356 reconstructed epidermis with several epidermal cell layers and a functional *stratum corneum*. Optionally,
357 RhE models may contain, below the epidermal layer, an additional layer mimicking dermal properties.
358 For reconstructing the epidermal tissue, normal, human-derived epidermal keratinocytes are cultured to
359 form a multilayered, highly differentiated model of the human epidermis. This tissue consists of organized
360 basal, spinous and granular layers, and a multilayered *stratum corneum* containing intercellular lamellar
361 lipid layers arranged in patterns analogous to those found *in vivo*. Irritant substances are identified by their
362 ability to decrease cell viability below defined threshold levels (e.g. 50%).

363

364 ## 2.4 Procedure for Skin Irritation Testing Using RhE

365 Detailed standard operating procedures (SOPs) of ECVAM-validated test methods [22-24] can be
366 downloaded from the ECVAM website [http://ecvam.jrc.ec.europa.eu].

367 Essential components of the test procedure, irrespective of the individual test methods employed are
368 provided in **§ 1 to 17**.

369 # 3. *IN VITRO* PERFORMANCE STANDARDS BASED ON RhE

370 The PS are comprised of three elements: (i) Essential Test Method Components, (ii) List of Reference
371 Chemicals and (iii) Target Values for Reliability and Predictive Capacity (Accuracy). Moreover, the PS
372 are structured according to continuously numbered paragraphs (blue font) that stipulate
373 evaluation/acceptance criteria or provide rules/guidance for testing in the context of equivalence
374 validation studies.

375 ## 3.1 ELEMENT 1: Essential Test Method Components

376 RhE models are based on established technologies of tissue engineering [25-34] and can be obtained
377 commercially or be developed or constructed in a testing laboratory. Prior to routine use for safety
378 assessments according to GLP principles [35] or outside the framework of accepted quality assurance
379 schemes, any proposed test method employing a test system based on RhE should be validated and at least
380 comply with the following conditions:

381 ### 3.1.1 *General Model Conditions*

382 **§ 1** Cell source and tissue architecture: Normal human keratinocytes should be used to
383 reconstruct the epithelium. Multiple layers of viable epithelial cells (basal layer, *stratum spinosum*,
384 *stratum granulosum*) should be present under a functional *stratum corneum*. The *stratum corneum*
385 should be multilayered containing the essential lipid profile to produce a functional barrier with
386 robustness to resist rapid penetration of cytotoxic markers chemicals, e.g. sodium dodecyl
387 sulphate (SDS) or Triton X-100. The barrier function should be demonstrated and may be
388 assessed either by determination of the concentration at which an established cytotoxic marker
389 chemical reduces the viability of the tissues by 50% (IC_{50}) after a fixed exposure time, or by
390 determination of the exposure time required to reduce cell viability by 50% (ET_{50}) upon
391 application of the cytotoxic marker chemical at a specified, fixed concentration. The containment
392 properties of the RhE model should prevent the passage of material around the *stratum corneum*
393 to the viable tissue, which would lead to poor modelling of skin exposure. The RhE model should
394 be free of contamination by bacteria, viruses, mycoplasma, or fungi.

395 ### 3.1.2 *Functional Model Conditions*

396 **§ 2** Viability: The magnitude of viability is usually quantified by metabolically converted
397 vital dyes. The preferred assay for determining the magnitude of viability is the MTT [3-(4,5-
398 Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide, Thiazolyl blue; EINECS number 206-
399 069-5, CAS number 298-93-1] [36]. In the case of MTT, cell viability is measured by
400 dehydrogenase conversion of the vital dye into a blue formazan salt that is quantitatively
401 measured after extraction from tissues [36]. The optical density (OD) of the extracted

402 (solubilised) dye from the tissue treated with the negative control (NC) should be at least 20 fold
403 greater than the OD of the extraction solvent alone. It should be documented that the tissue treated
404 with NC is stable in culture (provide similar viability measurements) for the duration of the test
405 exposure period.

406 **§ 3** Barrier function: The *stratum corneum* and its lipid composition should be sufficient to
407 resist the rapid penetration of cytotoxic marker chemicals, e.g. SDS or Triton X-100, as estimated
408 by IC_{50} or ET_{50} (**§ 1**).

409 **§ 4** Morphology: RhE models should display human epidermis-like structure (including a
410 multilayered *stratum corneum*), as demonstrated by histological examination.

411 **§ 5** Reproducibility: The results of the RhE test method should demonstrate reproducibility
412 over time and between laboratories, preferably by an appropriate batch control (benchmark)
413 chemical producing a mid-range cell viability response of the tissue (between about 30 and 70%).

414 **§ 6** Quality control (QC): The RhE model developer/supplier should ensure that each batch of
415 the RhE model used meets defined production release criteria, among which those for *viability* (**§**
416 **2**), *barrier function* (**§ 3**) and *morphology* (**§ 4**) are the most relevant. These data should be
417 provided to the test method users, so that they are able to include this information in the test
418 report. An acceptability range (upper and lower limit) for the IC_{50} or the ET_{50} should be
419 established by the RhE model developer/supplier, or investigator when using an in-house model.
420 Only results produced with qualified tissues can be accepted for reliable prediction of skin
421 irritation classification. As an example, the acceptability ranges for the RhE test methods
422 validated in the ECVAM SIVS [2] are given in Table 1:

423

424 ***Table 1:*** *Examples of QC batch release criteria*

	Lower acceptance limit	**Mean of acceptance range**	**Upper acceptance limit**
EpiSkin™ (VRM) (18 hours treatment with SDS) [22]	IC_{50} = 1.0 mg/mL	IC_{50} = 2.3 mg/mL	IC_{50} = 3.0 mg/mL
EpiDerm™ (1% Triton X-100) [24]	ET_{50} = 4.8 hours	ET_{50} = 6.7 hours	ET_{50} = 8.7 hours

425 **3.1.3** ***Application of Test and Control Substances***

426 **§ 7** Application of Test Substances: A sufficient number of tissue replicates should be used
427 for each test and for the controls (usually three replicates per run, see **§ 12**). For liquid as well as
428 solid substances, sufficient amount of test substance should be applied to uniformly cover the
429 epidermis surface while avoiding an infinite dose, i.e. a minimum of 25 µL/cm^2 or (25 mg/cm^2)
430 should be used. For solid substances, the epidermis surface should be moistened with deionised or
431 distilled water before application, to improve contact between the test substance and the epidermis
432 surface. Whenever possible, solids should be ground and tested as a fine powder. At the end of the
433 exposure period, the test substance should be carefully washed from the epidermis surface with an
434 appropriate aqueous buffer, or 0.9% NaCl. Depending on the RhE test method used, the exposure
435 period may vary. For the RhE test methods validated by ECVAM at time of drafting, the exposure
436 time varies between 15, 42 and 60 minutes, and the incubation temperature between room
437 temperature (~20°C) and 37°C. These exposure times and temperatures are optimized for each

438 RhE test method and represent the different intrinsic properties of the test methods. For details,
439 see the SOPs [22-24].

440 **§ 8** Application of Control Substances: Concurrent negative (NC) and positive controls (PC)
441 should be used in each run to demonstrate that viability (NC), barrier function and resulting tissue
442 sensitivity (PC) of the tissues are within a defined historical acceptance range. The suggested PC
443 substance is 5% aqueous SDS. The suggested NC substances are water or phosphate buffered
444 saline (PBS).

445 3.1.4 *Cell Viability Measurements*

446 **§ 9** Post-treatment incubation period: The most important element of the test procedure is that
447 viability measurements are not performed immediately after the exposure to the test substances,
448 but after a sufficiently long post-treatment incubation period of the rinsed tissues in fresh medium.
449 This period allows both for recovery from weak cytotoxic effects and for appearance of clear
450 cytotoxic effects. The test optimisation phase [13, 17, 18] demonstrated that a 42 hours post-
451 treatment incubation period was optimal.

452 **§ 10** Assay for determining cell viability: The MTT assay is a validated quantitative method
453 which should be used to measure cell viability. The MTT assay [36] has been shown to be
454 compatible with use in a three-dimensional tissue construct. The tissue sample is placed in MTT
455 solution of appropriate concentration (e.g. 0.3 – 1 mg/mL) for 3 hours. The precipitated blue
456 formazan product is then extracted from the tissue using a solvent (e.g. isopropanol or acidic
457 isopropanol), and the concentration of formazan is measured by determining the OD at 570 nm
458 using a bandpass filter of maximum ± 30 nm.

459 Optical properties of the test substance or its chemical action on the MTT may interfere with the
460 assay leading to a false estimate of viability (because the test substance may prevent or reverse the
461 colour generation as well as cause it). This may occur when a specific test substance is not
462 completely removed from the tissue by rinsing or when it penetrates the epidermis. If the test
463 substance acts directly on the MTT, is naturally coloured, or becomes coloured during tissue
464 treatment, additional controls should be used to detect and correct for test substance interference
465 with the viability measurement technique. Detailed description of how to correct direct MTT
466 reduction and interferences by colouring agents should be available in SOPs (for details see
467 references [22-24]). Non specific colour due to these interferences should not exceed 30% of NC
468 (for corrections). If non specific colouration is > 30% compared to NC value, the test substance is
469 considered as incompatible with the test method.

470 3.1.5 *Organising and Conducting Tests*

471 **§ 11** NC and PC: A negative control (NC) and a positive control (PC) must be tested
472 concurrently with the test substances to demonstrate that viability (NC), barrier function and
473 resulting tissue sensitivity (PC) of the tissues are within a defined historical acceptance range and
474 to be able to normalise the results obtained for the test substances to the NC.

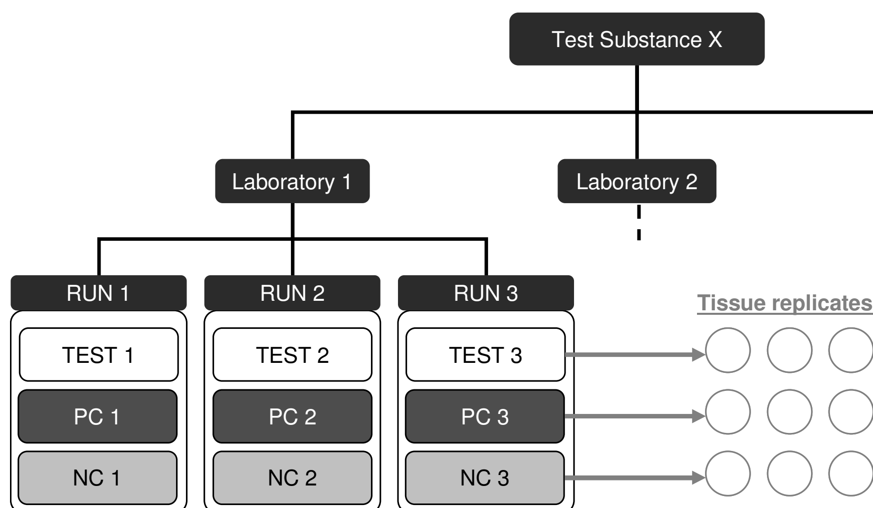
475 **§ 12** Test substances: In each run, each test substance as well as NC and PC are concurrently
476 tested in at least three tissues ("tissue replicates" or "replicate set"). However, the number of
477 concurrently tested tissue replicates recommended per run (all cases) may be reduced if
478 sufficiently statistically/scientifically justified (e.g. at least 90% of concurrently treated tissue
479 replicates during development/optimization of the test method showing a difference of viability

480 below 5%). A single replicate set for each NC and PC is sufficient, even if more than one test
481 substance is tested at the same time.

482 A set of x number of test substance concurrently tested in three tissue replicates each plus three
483 replicate tissues treated with NC and three replicate tissues treated with PC is called "run" (Figure
484 1). To interpret the results and make predictions, the results for each test substance must be
485 normalised to the NC which is arbitrarily set to 100% viability. This is done by calculating the
486 arithmetic mean of the tissue replicates of each test substances and normalising this value with
487 regard to the NC (=100%).

488

489 ***Figure 1:*** *Schematic depiction of test conduct for a given test substance. A test is the concurrent*
490 *testing of a test substance in a sufficient number of tissue replicates (§ 12) (grey circles). Any test*
491 *needs to be run concurrently with both a PC and a NC (both also tested in a sufficient number of*
492 *tissue replicates (§ 12) (grey circles). The total of test and concurrent PC and NC is hence called*
493 *a run. A run may be invalid if one of the three constituents is invalid: the test, the PC or the NC.*
494 *If more than one test substance is measured concurrently in one run, it is still sufficient to run one*
495 *PC and NC on the basis of a sufficient number of tissue replicates. A run may thus comprise one*
496 *to x number of tests concurrently executed with an appropriate number of tissue replicates of PC*
497 *and NC.*



498

499

500 **3.1.6 *Interpretation of Results and Prediction Model***

501 **§ 13** Normalisation of results to NC: The OD values obtained with each test substance can be
502 used to calculate the percentage of viability normalised to the NC, which is set to 100%. Test
503 acceptance criteria (section 3.1.7, **§ 15**) must be employed to ensure the quality of tests used for
504 the prediction of skin irritation.

505 **§ 14** Prediction model: The cut-off value of percentage cell viability distinguishing classified
506 (irritant) from non-classified test substances and the statistical procedure(s) used to evaluate the
507 results and identify irritant substances, should be clearly defined, documented, and proven to be
508 appropriate. The cut-off value (50% viability normalised to average NC viability value set at
509 100%) of the three ECVAM–validated full-replacement RhE test methods was established during
510 pre-validation and test optimisation studies [13, 15, 17-19] and confirmed in the ECVAM SIVS.
511 The prediction model for predicting the skin irritation potential of xenobiotics based on this cut-

512 off value used by all three validated full-replacement test methods is given below:

513 *Prediction model:*

514 *The test substance is considered to be irritant to skin in accordance with regulation*
515 *EC 1272/2008 (GHS Category 2) if the tissue viability after exposure and post-*
516 *treatment incubation is equal or lower (≤) than 50%.*

517 *The test substance may be considered as 'no-category' if the tissue viability after*
518 *exposure and post-treatment incubation is higher (>) than 50%.*
519

520 **3.1.7** ***Test Acceptance Criteria***

521 **§ 15** Acceptance criterion for test and control substances: A high Standard Deviation (SD)
522 associated with the arithmetic mean of the tissue replicates of a given test substance may indicate
523 defects in single tissues or may be due to inappropriate dosing, spreading or chemical instability
524 after contact with the tissue. Certain chemicals have intrinsic properties that may lead to high
525 variability. If variability is observed, storage and handling of the test substance at all steps should
526 be carefully reconsidered. A test is therefore only valid, if the SD obtained from three or more
527 concurrently tested tissue replicates is within the 95% viability confidence interval (e.g. SD ≤
528 18% for the VRM). If only two replicate tissues per test substance per run are performed during
529 the validation study (**§ 12**), the difference in viability between the two replicate tissues must be
530 equal or lower (≤) than 20%.

531 **If this condition is not met, the run must be repeated for the test substance (§ 24).**

532 **§ 16** Acceptance criteria for negative control (NC): A non-irritant (non-classified) NC (e.g.
533 PBS or water) must be tested concurrently with the test substance. The tissues treated with NC
534 should be stable in culture and provide similar viability measurements throughout the period
535 corresponding to both test substance exposure and post-treatment incubation. A minimum
536 viability (e.g. expressed as absolute OD of the vital dye based on the tissue replicates of the NC)
537 must be established as a test acceptance criterion, and control OD values should not be below
538 such historical established boundaries

539 **If this condition is not met, the full run must be repeated (§ 24).**

540 **§ 17** Acceptance criteria for positive control (PC): An appropriate PC should be used in the
541 assay (e.g. 5% aqueous SDS), typically specified in the SOP. The PC should not drive the test
542 method into saturation (i.e. kill all cells leading to a clear-cut but meaningless response) but allow
543 to assess whether tissues treated with PC are able to respond to an irritant substance (PC as
544 surrogate) within a dynamic response range and show the expected sensitivity for the specific test
545 method in the response to the PC. The range of responses to the PC must be developed and based
546 on data obtained from a sufficiently high number of independent experiments during the historical
547 test development/optimization phases. In each assay, the positive control must

548 - be correctly classified as irritant,
549 - be within the established range of responses
550 - and the SD of the three tissue replicates must be below a defined maximum

551 **If these conditions are not met, the full run must be repeated (§ 24).**

552 An example of typical ranges for the test methods validated in the ECVAM SIVS is given in
553 Table 2.

554

555 ***Table 2:*** *Example for range of model responses to PC (SDS, 5% w/v) issued from ECVAM formal*
556 *SIVS [2]*

	Viability	**Range (95% prediction interval)**	**SD**
EpiSkin™ (VRM)	< 40%	1.5 – 32.2 (1.3 – 41.6)	≤ 18%
EpiDerm™	< 20%	3.7 – 13.8 (4.7 – 13.6)	≤ 18%

557

558 3.2 ELEMENT 2: List of Reference Chemicals

559 3.2.1 *Selection Criteria*

560 The chemicals listed in Table 3 provide a representative distribution of the 58 chemicals used in the
561 validation study of the VRM with regard to chemical functionality and physical state [14]. Their selection
562 is based on the following criteria:

563 1. the chemicals are commercially available
564 2. they are representative of the full range of Draize irritancy scores (from non-irritant to strong
565 irritant)
566 3. they have a well-defined chemical structure
567 4. they are representative of the VRM's reproducibility and predictive capacity as determined in
568 the ECVAM validation study
569 5. they are representative of the chemical functionalities used in the validation process
570 6. they are not associated with an extremely toxic profile (e.g. carcinogenic or toxic to the
571 reproductive system) and they are not associated with prohibitive disposal costs.

572 3.2.2 *Proposed Reference Chemicals*

573 **§ 18** Reference chemicals: The 20 Reference Chemicals (RC) provided in Table 3 should only
574 be used to assess the performance of proposed test methods in the context of validation studies
575 intending to demonstrate equivalence of similar or modified test methods with respect to the
576 VRM. Importantly, these RC should not be used for the development of similar or modified test
577 methods (see Annex I). In the context of these PS, the RC listed in Table 3 represent the minimum
578 number of substances that should be used to determine if the reliability and relevance (accuracy)
579 of a proposed similar or modified test method, proven to be structurally and functionally
580 sufficiently similar to the VRM as specified in the PS, are comparable to those of the VRM [2,
581 16].

582 The RC listed in Table 3 include chemicals representing different chemical classes of interest, and
583 are representative of the full range of Draize irritancy scores (from non-irritant to strong irritant).
584 The chemicals included in this list comprise 10 UN GHS Category 2 chemicals and 10 Non-
585 Categorised chemicals, of which 3 are optional UN GHS Category 3 chemicals. In the EU, the
586 optional Category 3 is not implemented as a distinct category and is considered as No Category.

587 In situations where a listed chemical is unavailable, other chemicals for which adequate *in vivo*
588 reference data are available could be used [13, 18]. If desired, additional chemicals representing
589 other chemical classes and for which adequate *in vivo* reference data are available may be added
590 to the minimum list of RC to further evaluate the accuracy of the proposed test method.

591

592 ***Table 3:*** *Reference Chemicals for determination of Reliability and Accuracy Values for Similar or*
593 *Modified RhE skin irritation Test Methods.*

NON-CLASSIFIED CHEMICALS (REGARDED AS NON-IRRITANT)					
Chemical	**CAS Number**	**Physical state**	***In vivo* score**[1]	**VRM *in vitro* Cat.**	**GHS *in vivo* Cat.**
1-bromo-4-chlorobutane	6940-78-9	Liquid	0	Cat. 2	No Cat.
diethyl phthalate	84-66-2	Liquid	0	No Cat.	No Cat.
naphthalene acetic acid	86-87-3	Solid	0	No Cat.	No Cat.
allyl phenoxy-acetate	7493-74-5	Liquid	0.3	No Cat.	No Cat.
isopropanol	67-63-0	Liquid	0.3	No Cat.	No Cat.
4-methyl-thio-benzaldehyde	3446-89-7	Liquid	1	Cat. 2	No Cat.
methyl stearate	112-61-8	Solid	1	No Cat.	No Cat.
heptyl butyrate	5870-93-9	Liquid	1.7	No Cat.	No Cat. *(Optional Cat. 3)*[2]
hexyl salicylate	6259-76-3	Liquid	2	No Cat.	No Cat. *(Optional Cat. 3)*[2]
cinnamaldehyde	104-55-2	Liquid	2	Cat. 2	No Cat. *(Optional Cat. 3)*[2]

594 Continued on next page

595

CLASSIFIED (IRRITANT) CHEMICALS					
Chemical	**CAS Number**	**Physical state**	***In vivo* score[1]**	**VRM *in vitro* Cat.**	**GHS *in vivo* Cat.**
1-decanol*	112-30-1	Liquid	2.3	Cat. 2	Cat. 2
cyclamen aldehyde	103-95-7	Liquid	2.3	Cat. 2	Cat. 2
1-bromohexane	111-25-1	Liquid	2.7	Cat. 2	Cat. 2
2-chloromethyl-3,5-dimethyl-4-methoxypyridine HCl	86604-75-3	Solid	2.7	Cat. 2	Cat. 2
di-n-propyl disulphide*	629-19-6	Liquid	3	No Cat.	Cat. 2
potassium hydroxide (5% aq.)	1310-58-3	Liquid	3	Cat. 2	Cat. 2
benzenethiol, 5-(1,1-dimethylethyl)-2-methyl	7340-90-1	Liquid	3.3	Cat. 2	Cat. 2
1-methyl-3-phenyl-1-piperazine	5271-27-2	Solid	3.3	Cat. 2	Cat. 2
heptanal	111-71-7	Liquid	3.4	Cat. 2	Cat. 2
tetrachloroethylene	127-18-4	Liquid	4	Cat. 2	Cat. 2

596 [1] *In vivo* score in accordance with the OECD Test Guideline 404 [12].
597 [2] Under these PS for RhE test methods, the UN GHS optional Category 3 [3, 5] is considered as no category.
598 * Chemicals that are identified as irritant in the rabbit test but for which there is reliable evidence that they are non-
599 irritant in humans [37-39]. See also § 22 on the Acceptance Criteria for the Predictive Capacity (Accuracy) of the
600 test methods and in particular the specific conditions/restrictions that apply to the sensitivity of a proposed test
601 method.

602

603

604

605 3.3 ELEMENT 3: Target Values for Reliability and Predictive Capacity (Accuracy)

606 When using the list of recommended RC (Table 3), the proposed test method should provide performance
607 characteristics (Reliability and Predictive Capacity) that are equal or better than the pre-defined
608 performance values derived from the performance of the VRM and specified below. Non-classified and
609 classified (GHS Category 2) [3] chemicals, ranging from non-irritant to strong irritant, and representing
610 relevant chemical classes are included, so that the Reliability and Accuracy (sensitivity, specificity and
611 overall accuracy) of the proposed test method can be determined and compared to those of the VRM. The
612 assessment of Reliability and Accuracy must include the use of different, independent production batches
613 of the RhE model so to assure that the performance is stable over time. The performance of a proposed
614 test method should be determined prior to its use for testing new substances.

615 3.3.1 *Study Acceptance Criteria for the Purpose of Equivalence Validation Studies Relevant*
616 *for Assessment of Reliability and Relevance*

617 **§ 19** Testing the 20 Reference Chemicals (RC) for purposes of validation: For purposes of
618 small scale validation studies to establish the reliability and relevance of proposed similar or
619 modified test methods, **all 20 RC should be tested in at least three laboratories**. In each
620 laboratory, **all 20 RC should be tested in three independent runs performed with different**
621 **tissue batches and at sufficiently spaced time points**.

622 Each run should consist of a minimum of **three concurrently tested tissue replicates** for each
623 included test substance, NC and PC (**§ 11**). However, the number of concurrently tested tissue
624 replicates recommended per run (all cases) may be reduced if statistically/scientifically
625 sufficiently justified (e.g. at least 90% of concurrently treated tissue replicates during
626 development/optimization of the test method showing a difference of viability below 5%).

627 **§ 20** Re-conducting runs ("retesting"): It is possible that one or several tests pertaining to one
628 or more test substances does/do not meet the test acceptance criteria specified (3.1.7, **§ 15 to 17**)
629 or is/are not acceptable for other reasons. To complement missing data, for each test substance a
630 maximum number of two additional tests is admissible ("retesting"). More precisely, since in case
631 of retesting also PC and NC have to be concurrently tested, a maximum number of two additional
632 runs may be conducted for each test substance.

633 **§ 21** Acceptance criteria with regard to the number of valid runs after retesting: It is
634 conceivable that even after retesting, the minimum number of three valid runs required for each
635 tested substance is not obtained for every RC in every participating laboratory, leading to an
636 incomplete data matrix (Figure 1-3 in Annex III). In such cases the following rules, connected by
637 a logical AND operator, apply with regard to the acceptability of datasets for purposes of
638 equivalence validation studies:

639 Rule 1. All 20 RC should have at least one COMPLETE RUN SEQUENCE (definition
640 below)
641 AND

642 Rule 2. In each of at least three participating laboratories, a minimum of 85% of the RUN
643 SEQUENCES need to be COMPLETE (for 20 chemicals: 3 invalid run sequences
644 allowed in a single laboratory).
645 AND

646 Rule 3. A minimum of 90% of all possible RUN SEQUENCES from at least three
647 laboratories need to be COMPLETE (for 20 chemicals tested in 3 laboratories: 6
648 invalid run sequences allowed in total).

649
650
651 **Definition of "Run Sequence" and "Complete Run Sequence" (see also Annex III):**

652 1) A **run sequence** consists of three independent runs from one laboratory concerning one test substance.

653 2) A **complete run sequence** is a run sequence where all three runs are valid. This means that any single
654 invalid run invalidates an entire run sequence of three runs.

655
656

657 ### 3.3.2 *Rules for Calculation of Reliability (Reproducibility) and Predictive Capacity*
658 *(Accuracy)*

659 **§ 22** Rules for the calculation of reliability and predictive capacity (accuracy) values: The
660 calculation of the reliability and accuracy values of the proposed test method should be done
661 according to the following rules, connected by a logical AND operator, which ensure that the
662 values describing reliability and relevance are calculated in a predefined and hence consistent
663 manner:

664 Rule 1. Only the data of runs from COMPLETE RUN SEQUENCES qualify for the
665 calculation of the test method within laboratory variability (WLV), between
666 laboratory variability (BLV), and predictive capacity (accuracy).
667 AND

668 Rule 2. The final classification for each RC in each participating laboratory should be
669 obtained by using the MEAN VALUE OF VIABILITY over the different runs.
670 AND

671 Rule 3. Only the data obtained for chemicals that have COMPLETE RUN SEQUENCES
672 IN ALL PARTICIPATING LABORATORIES qualify for the calculation of the
673 test method BLV.
674 AND

675 Rule 4. The calculation of the accuracy values should be done on the basis of the
676 INDIVIDUAL LABORATORY PREDICTIONS obtained for the 20 RC by the
677 different participating laboratories.

678 ### 3.3.3 *Reliability*

679 The reproducibility (within one laboratory over time and between laboratories) of the proposed test
680 method obtained with the set of RC should be at least comparable to that of the VRM.

681 **§ 23** Acceptance criterion for reproducibility within one laboratory (and over time): The
682 calculation of WLV of the test method, i.e. concordance of classifications between runs, should be
683 done independently for each participating laboratory on the basis of the RC tested during the
684 validation study and according to Rule 1 of **§ 22**. The concordance of classifications (No Category
685 / Category 2) for the 20 RC obtained in different, independent runs within a single laboratory
686 must be equal or higher (≥) than 90%.

687 *For example, applying the rules set out in § 22 to the data obtained in the ECVAM SIVS*
688 *[2]:*

- *With EpiSkin™, a 93-95% within-laboratory concordance of classifications was obtained for the 58 SIVS substances and a 93-100% within-laboratory concordance of classifications was obtained for the Reference Chemicals listed in Table 3*
- *With EpiDerm™, a 94-98% within-laboratory concordance of classifications was obtained for the 58 SIVS substances and a 87-95% within-laboratory concordance of classifications was obtained for the Reference Chemicals listed in Table 3*

§ 24 Acceptance criterion for reproducibility between laboratories: The calculation of BLV of the test method, i.e. concordance of classifications between laboratories, should be done on the basis of the RC tested during the validation study and according to Rules 1, 2 and 3 of **§ 22**. The concordance of final classifications (No Category / Category 2) for the 20 RC obtained by the different participating laboratories (minimum of three) must be equal or higher (≥) than 80%.

For example, applying the rules set out in **§ 22** *to the data obtained in the ECVAM SIVS [2]:*

- *With EpiSkin™, 6 of 50 test substances (88% concordance) and 1 of 13 Reference Chemicals (Table 3) (92% concordance) were classified discordantly across three laboratories.*
- *With EpiDerm™, 3 of 42 test substances (93% concordance) and 2 of 14 Reference Chemicals (Table 3) (86% concordance) were classified discordantly across three laboratories.*

3.3.4 *Predictive Capacity (Accuracy)*

§ 25 Acceptance criteria for the Predictive Capacity (Accuracy) of the test method: The calculation of the accuracy values of the test method should be done on the basis of the RC tested during the validation study and according to Rules 1, 2, and 4 of **§ 22**. The accuracy values (sensitivity, specificity, false negative rate, false positive rate and overall accuracy) of the proposed similar or modified test method should be comparable to those derived from the VRM, taking into consideration additional information relating to relevance in the species of interest (Table 4).

The **sensitivity** should be equal or higher (≥) than 80% [9]. However, a further specific restriction applies to the sensitivity of the proposed *in vitro* test method inasmuch as ***only*** two *in vivo* Category 2 substances, 1-decanol and di-n-propyl disulphide, may be misclassified as "No Category" by more than one participating laboratory.

The **specificity** should be equal or higher (≥) than 70% [9]. There is no further restriction with regard to the specificity of the proposed *in vitro* test method, i.e. any participating laboratory may misclassify any *in vivo* No Category substance as long as the final specificity of the test method is within the acceptable range.

Overall accuracy should be equal or higher (≥) than 75% [9].

Although the sensitivity of the VRM calculated for the 20 RC listed in Table 3 is equal to 90%, the defined minimum sensitivity value required for any similar or modified test method to be considered valid is set at 80% since both 1-decanol (a borderline chemical) and di-n-propyl disulphide (a false negative of the VRM, see Table 3) are known to be non-irritant to humans when tested in the 4h Human Patch Test [37-39], although being identified as irritants in the rabbit test. Since RhE models are based on cells of human origin, they may predict these chemicals as non-irritant (No Category).

735 Margins of tolerance of at maximum 2% of the pre-defined target values may be acceptable, but
736 should be scientifically justified.

737 ***Table 4.*** *Required accuracy values for any similar or modified test method to be considered*
738 *scientifically valid.*

PREDICTIVE CAPACITY		
Sensitivity	**Specificity**	**Overall Accuracy**
80%	70%	75%
FNR	**FPR**	
20%	30%	

739

740 ## 3.4 Additional Guidance

741 ### 3.4.1 *Test Reporting*

742 **§ 26** Requirements of data reporting: For each test, data from individual replicate tissues (*e.g.*
743 OD values and calculated percentage cell viability data for each test substance, including
744 classification) should be reported in tabular form, including data from repeat experiments as
745 appropriate. In addition, means ± standard deviation for each test should be reported. Observed
746 interactions with MTT reagent and coloured test substances should be reported for each tested
747 substance.

748 **§ 27** Suggested structure for the test report: The test report should include the following
749 information**:**

750 *1. Test and Control Substances*

751 - Chemical name(s) such as IUPAC or Chemical Abstract Services (CAS)
752 name and number, if known
753 - Purity and composition of the substance or – if employed for
754 preparations – of the preparation (in percentage(s) by weight)
755 - Physical-chemical properties relevant to the conduct of the study (e.g.
756 physical state, stability, volatility, pH, water solubility, if known)
757 - Treatment of the test/control substances prior to testing, if applicable (e.g.
758 warming, grinding)
759 - Storage conditions

760 *2. Justification of the RhE Model and Protocol Used*

761 *3. Test Conditions*

762 - Cell system used
763 - Calibration information for measuring device, and bandpass used for
764 measuring cell viability (e.g. spectrophotometer)
765 - Complete supporting information for the specific RhE model used, including
766 its performance. This should include, but is not limited to:
767 - o Viability
768 - o Barrier Function
769 - o Morphology (i.e. histological data, provided by the model
770 developer/supplier)
771 - o Reproducibility and Predictive Capacity
772 - o Quality controls (QC) of the model
773 - Details of the test procedure used
774 - Test doses used, duration of exposure and post-treatment incubation period
775 - Description of any modifications of the test procedure
776 - Reference to historical data of the model. This should include, but is not
777 limited to:
778 - o Acceptability of the QC data with reference to historical batch data

- Acceptability of NC and PC values with references to PC and NC means and ranges

- Description of evaluation criteria used including justification for the selection of the cut off point(s) for the prediction model

4. Results

- Tabulation of data from individual test substances
- Description of other effects observed

5. Discussion of the Results

6. Conclusion

3.4.2 *Complementary Endpoints*

§ 28 Although the inflammatory mediator IL-1α has undergone extensive evaluation during the ECVAM SIVS [2, 16] and showed some promise with regard to a possible improvement of the sensitivity of the MTT reduction endpoint [7], it was not found a scientifically valid endpoint mainly due to high within- and between-laboratory variability. Moreover, at present, there is no evidence from validation studies that would confirm the usefulness of any other complementary endpoint for predicting irritancy potential of xenobiotics or improving the accuracy of such predictions for classification and labelling.

It should be noted however, that some endpoints may have, due to their biological relevance, potential of being further developed/optimised for use as a predictive parameter in the future. In particular there is evidence that in response to physical or chemical stress, keratinocytes produce and release inflammatory cytokines interleukins [IL-1α, tumor necrosis factor α (TNF-a)], chemotactic cytokines [IL-8, interferon, e.g. induced protein 10 (IP-10)], growth-promoting factor [IL-6, IL-7, IL-15, granulocyte/macrophage colony-stimulating factor GM-CSF)], transforming growth factor [TGF], cytokines regulating humoral versus cellular immunity [IL-10, IL-12] and other signalling factors which rapidly generate cutaneous inflammation, suggesting that measurement of such keratinocyte responses may allow the evaluation of toxicological properties of chemicals in order to support the identification of irritants and/or sensitizers [40].

808 # ANNEXE I – Conduct of Equivalence Studies

809 ## Recommendations Regarding the Conduct of
810 ## Equivalence Validation Studies

811 ### 1. Recommendations for Test Development and Optimization

812 1) In agreement with international recommendations, a test method should never be developed and
813 validated with the same set of chemicals, i.e. the training set of chemicals must be sufficiently
814 different from the validation testing set. Therefore, also for the development of a similar or
815 modified test methods, the test substances used should be different from the 20 Reference
816 Chemicals (RC) specified in this PS document.

817 2) For this purpose, a variety of commercially available (existing) chemicals with well documented
818 *in vivo* data and *in vitro* data (for EpiSkin™ and EpiDerm™) are available in the published
819 literature [2, 13, 14, 18]. The set of RC's may then, in addition to the test substances used for
820 development and optimization, be tested before entering a blind equivalence inter-laboratory ring
821 trial.

822 3) Before entering an equivalence validation study according to this PS document, an SOP including
823 sufficiently developed test acceptance criteria must exist.

824 ### 2. Recommendations for Conducting Equivalence Validation Studies:

825 1) It is strongly recommended that test producers contact ECVAM with regard to an evaluation
826 whether or not their novel test method qualifies for an equivalence validation study before
827 embarking on external ring trials.

828 2) In case of positive evaluation, it is recommended that equivalence validation studies be conducted
829 in close consultation with an independent non-profit organization without specific interests and
830 experienced in validation (e.g. ECVAM, ICCVAM, JaCVAM, or ZEBET).

831 3) A Study plan containing management structure, training issues and timelines should be developed
832 and approved by the Validation Management Group.

833 4) The credibility of the information on reproducibility and robustness of the novel test under real
834 life conditions may be increased by involving a "naïve laboratory" which has never worked with
835 the specific test method before. Moreover, such a "naïve laboratory" should ideally have no or
836 only limited experience with RhE based test methods.

837 5) For each Reference Chemical three independent test runs in each of at least three laboratories
838 should be performed. A minimum of three concurrently tested tissue replicates should be used per
839 run. However, the number of concurrently tested tissue replicates recommended per run (all cases)
840 may be reduced if sufficiently statistically/scientifically justified (e.g. based on information
841 available from development/optimization phases or previous pre-validation/validation studies).
842 Test developers wishing to reduce the recommended number of tissue replicates are however
843 advised to contact ECVAM prior to embarking on external ring trials so that the reduction can be
844 confirmed or, should it be inappropriately justified, rejected by ECVAM.

845 6) Before starting a study, tissues which feature obvious, visible damage (e.g. infection with mould
846 or cuts in the epidermis) should be discarded and not used at all.

847 7) Excess production of data (e.g. through the use of a larger than required number of tissue
848 replicates) and subsequent data selection are regarded highly unethical. All tested tissues must be
849 reported.

850 8) Retesting should solely be based on test acceptance criteria not met, and should be performed
851 only after completing and submitting the required number of three independent runs.

852 9) In the final report, all non-qualified test runs should be reported together with the qualified test
853 runs.

854 10) Test substances should be coded "double blind", i.e. with different codes for each laboratory.
855 Preferably, coding and distribution should be performed by an independent organization.

856 11) Laboratories should not communicate between each other during the validation trial. If
857 communication is deemed necessary, the Validation Management Group must be consulted and
858 should mediate all contacts.

859 12) Results should be submitted in original, preferably password protected data spreadsheets,
860 immediately after each experimental run to either the independent biostatistician of the validation
861 study or any other independent organisation.

862 13) In a non GLP environment, in addition to the electronic raw data spreadsheets, electronic or paper
863 documentation allowing a retrospective complete audit of the study should be made and archived.

864 14) Disclosure of coded test substance identities should only happen after finalization of the study, i.e.
865 complete submission of data, and after participating laboratories have confirmed the correctness
866 of the coded data used for evaluation by an independent biostatistician.

867

868 # ANNEXE II – Proficiency Testing

869 ## Proficiency Testing and Chemicals

870 Prior to routine use of any of the three ECVAM-validated full-replacement test methods (EpiSkin™
871 (VRM), EpiDerm™ SIT and SkinEthicTM RHE) [7-9], laboratories may wish to demonstrate technical
872 proficiency, using the ten Proficiency Chemicals recommended in Table 1. For similar (me-too) test
873 methods that are structurally and functionally similar to the VRM, or for modifications of any of the three
874 validated test methods, the Performance Standards described above should be used to demonstrate
875 comparable reliability and accuracy of the proposed test method prior to its use for regulatory testing.

876 As part of the proficiency exercise, it is recommended that the user verifies the barrier properties of the
877 tissues after receipt, as specified by the RhE model producer. This is particularly important if tissues are
878 shipped over long distances/transit. Once a test method has been successfully established and proficiency
879 in its use has been demonstrated, such verification will not be necessary on a routine basis. However,
880 when using a test method routinely, it is recommended to continue to assess the barrier properties in
881 regular intervals, e.g. every six to twelve months.

882

883
884 ***Table 1:*** *Proficiency Chemicals*[1]

Chemical	CAS Number	*In vivo* score[2]	Physical state	GHS category
naphthalene acetic acid	86-87-3	0	Solid	No Cat.
isopropanol	67-63-0	0.3	Liquid	No Cat.
methyl stearate	112-61-8	1	Solid	No Cat.
heptyl butyrate	5870-93-9	1.7	Liquid	No Cat. *(Optional Cat. 3)*[3]
hexyl salicylate	6259-76-3	2	Liquid	No Cat. *(Optional Cat. 3)*[3]
cyclamen aldehyde	103-95-7	2.3	Liquid	Cat. 2
1-bromohexane	111-25-1	2.7	Liquid	Cat. 2
potassium hydroxide (5% aq.)	1310-58-3	3	Liquid	Cat. 2
1-methyl-3-phenyl-1-piperazine	5271-27-2	3.3	Solid	Cat. 2
heptanal	111-71-7	3.4	Liquid	Cat. 2

885 [1] The Proficiency Chemicals are a subset of the substances used in the validation study.
886 [2] *In vivo* score in accordance with the OECD Test Guideline 404 [12].
887 [3] Under these PS for RhE test methods, the UN GHS optional Category 3 [3, 5] is considered as no category.
888

889

890 # ANNEXE III – Guidance on Qualification Criteria for Test Substances

891 **Explanation of the Qualification Criteria for Test Substances**
892 **in the Context of Equivalence Validation Studies (c.f. § 21)**

893 ## 1. Overview

894 These PS feature three acceptance rules (described in **§ 21)** that ensure that there are sufficient data in the
895 validation data set to allow an appropriate evaluation of both ***reliability*** and ***relevance*** of the test method.
896 It is important to note that all three rules are connected by logical AND operators, i.e. all of the rules
897 apply in their entirety.

898 ## 2. Explanation of the rules

899 These rules allow for a certain number of invalid runs and, as a consequence, incomplete run sequences,
900 since any single invalid run within a run sequence invalidates the entire run sequence with regard to the
901 dataset. More specifically the rules limit the number of incomplete run sequences to an acceptable
902 minimum with regard to:

903 a) The test substances (Reference Chemicals)

904 b) The data coming from one laboratory

905 c) The overall dataset

906 The rules are explained in the following example and Figures 1-3.

907 **Example:**

908 *In validation study XYZ testing has been performed on the basis of 20 RC in three laboratories each*
909 *producing 3 independent runs (=run sequences) for each test substance (Figures 1-3). Therefore,*
910 *the number of run sequences per laboratory is 20 and the total number of run sequences is 60.*

911 *Rule 1 specifies that all 20 RC need to have at least one complete run sequence, i.e. a run sequence*
912 *with 3 valid independent runs from at least one laboratory.*

913 *Furthermore, rule 2 specifies that at maximum 3 run sequences (15%) per laboratory may be*
914 *invalid, i.e. one particular laboratory can have invalid data for a maximum of 3 substances. That*
915 *does however not mean that EACH participating laboratory may have 3 invalid run sequences*
916 *(which would be a total of 9 invalid run sequences).*

917 *Here rule 3 applies by specifying that 90% of all run sequences of the complete dataset must be*
918 *valid. This means that in total at least 54 run sequences need to be valid or, inversely, that 6 run*
919 *sequences can be invalid. Moreover, the invalid run sequences must not be clustered in a single*
920 *laboratory.*

921 Only if these rules are fulfilled and the criteria hence fulfilled can the dataset be considered complete for
922 the purpose of the equivalence validation study.

Figure 1: *Illustration of the concept of a "complete run sequence". A complete run sequences is a sequence of at three consecutive runs (from one laboratory) that are all valid (blue frame). Therefore, one single invalid run within a sequence of adjacent runs, invalidates the entire run sequences (as shown for chemical seven). The table illustrates furthermore some conceivable constellations of how individual invalid runs can disqualify run sequences: these may be singular invalid runs within one run sequence or clusters of invalid runs. Invalid runs marked by a red X. Incomplete run sequences in grey. Non-qualified test substances in grey. Qualified test substances in white. Note that this dataset would be considered insufficient for an equivalence validation study since it violates rule 1 (all 20 RC should have at least one complete run sequence, not fulfilled for chemicals 5 and 7) and rule 3 (90% of all run sequences should be valid. Not fulfilled since 8 run sequences out of 60 are invalid. Note however, that rule 2 is not violated: all laboratories have indeed at least 85% of valid run sequences.*

	Laboratory 1			Laboratory 2			Laboratory 3		
Test substance	Run 1	Run 2	Run3	Run 1	Run 2	Run3	Run 1	Run 2	Run3
1									
2	complete run sequence								
3									
4									
5	X	X	X	X	X	X	X	X	X
6									
7		X		X					X
8									
9	X	X	X	X	X	X			
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									

Figure 2: *This figure illustrates an extreme case of a dataset that just about qualifies although showing 18 invalid runs. The invalid runs are clustered within run sequences in a manner that still allows compliance with all rules specified. Note that one single additional invalid run would disqualify the dataset. Invalid runs marked by a red X. Incomplete run sequences in grey. Qualified test substances in white.*

	Laboratory 1			Laboratory 2			Laboratory 3		
Test substance	**Run 1**	**Run 2**	**Run3**	**Run 1**	**Run 2**	**Run3**	**Run 1**	**Run 2**	**Run3**
1									
2									
3									
4									
5	X	X	X	X	X	X			
6									
7									
8	X	X	X				X	X	X
9									
10									
11	X	X	X	X	X	X			
12									
13									
14									
15									
16									
17									
18									
19									
20									

Figure 3: *This figure illustrates the other extreme case of a dataset that just about qualifies although showing only 6 invalid runs distributed in a manner that leads to 6 invalid run sequences. Note that one single additional invalid run in a different run sequence from the six already invalid would disqualify the dataset. Invalid runs marked by a red X. Incomplete run sequences in grey. Qualified test substances in white.*

	Laboratory 1			Laboratory 2			Laboratory 3		
Test substance	**Run 1**	**Run 2**	**Run3**	**Run 1**	**Run 2**	**Run3**	**Run 1**	**Run 2**	**Run3**
1									
2									
3									
4									
5		X			X				
6									
7									
8	X				X				
9									
10									
11		X						X	
12									
13									
14									
15									
16									
17									
18									
19									
20									

GLOSSARY

Accuracy: The closeness of agreement between test method results and accepted reference values. It is a measure of test method performance and one aspect of relevance. The term is often used interchangeably with “concordance” to mean the proportion of correct outcomes of a test method.

Batch control chemical: Benchmark chemical producing a mid-range cell viability response of the tissue.

Cell viability: Parameter measuring total activity of a cell population *e.g.,* as ability of cellular mitochondrial dehydrogenases to reduce the vital dye MTT ([3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide, Thiazolyl blue), which depending on the endpoint measured and the test design used, correlates with the total number and/or vitality of living cells.

Concordance: This is a measure of test method performance for test methods that give a categorical result, and is one aspect of relevance. The term is sometimes used interchangeably with accuracy, and is defined as the proportion of all chemicals tested that are correctly classified as positive or negative. Concordance is highly dependent on the prevalence of positives in the types of chemicals being examined.

ET_{50}: Can be estimated by determination of the exposure time required to reduce cell viability by 50% upon application of the marker chemical at a specified, fixed concentration, see also IC_{50}.

Formulation: Please see definition for “Preparation”.

GHS (Globally Harmonized System of Classification and Labelling of Chemicals by the United Nations (UN)): A system proposing the classification of chemicals (substances and mixtures) according to standardized types and levels of physical, health and environmental hazards, and addressing corresponding communication elements, such as pictograms, signal words, hazard statements, precautionary statements and safety data sheets, so that to convey information on their adverse effects with a view to protect people (including employers, workers, transporters, consumers and emergency responders) and the environment [3].

IC_{50}: Can be estimated by determination of the concentration at which a marker chemical reduces the viability of the tissues by 50% (IC_{50}) after a fixed exposure time, see also ET_{50}.

Infinite dose: Amount of test chemical applied to the *epidermis* exceeding the amount required to completely and uniformly cover the *epidermis* surface.

Me-too test: A colloquial expression for a test method that is structurally and functionally similar to a validated and accepted reference test method. Such a test method would be a candidate for catch-up validation. Interchangeably used with similar test method.

Mixture: Used in the context of the UN GHS [3] and the CLP Regulation [5] with the same meaning as “Preparation” (see below), but in the context of REACH [41] with the same meaning as “Multi-constituent substance” (see below).

Mono-constituent substance: An individual molecule present at >80% defines a single component substance, identified as that one molecule name only. See also definition for “Substance” below.

Multi-constituent substance: A reaction product obtained from a chemical process containing several individual molecules present in the range >10% to <80%. All these molecules are defined components of the reaction product (substance), which is identified with all component molecule names. Typically, multi-constituent substances would comprise isomers or generically related species as components. See also definition for “Substance” below.

Performance standards (PS): Standards, based on a validated test method, that provide a basis for evaluating the comparability of a proposed test method that is mechanistically and functionally similar. Included are; (i) essential test method components; (ii) a minimum list of Reference Chemicals selected

from among the chemicals used to demonstrate the acceptable performance of the validated test method; and (iii) the similar levels of accuracy and reliability, based on what was obtained for the validated test method, that the proposed test method should demonstrate when evaluated using the minimum list of Reference Chemicals.

Preparation: A blend of substances, integrated in measured proportions. Typically, preparations would include solvents or other substances combined as a formulation appropriate for storage, transport, marketing, use application, etc. “Preparation” has the same meaning as “Formulation”.

Relevance: Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of the accuracy (concordance) of a test method.

Reliability: Measures of the extent that a test method can be performed reproducibly within and between laboratories over time, when performed using the same protocol. It is assessed by calculating intra- and inter-laboratory reproducibility..

Replacement test: A test which is designed to substitute for a test that is in routine use and accepted for hazard identification and/or risk assessment, and which has been determined to provide equivalent or improved protection of human or animal health or the environment, as applicable, compared to t the accepted test, for all possible testing situations and chemicals.

Screening test: Often a rapid, simple test method conducted for the purpose of classifying chemicals into a general category of hazard. The results of a screening test generally are used for preliminary decision making in the context of a testing strategy (*i.e.,* to assess the need for additional and more definitive tests). Screening tests often have a truncated response range in that positive results may be considered adequate to determine if a chemical is in the highest category of a hazard classification system without the need for further testing, but are not usually adequate without additional information/tests to make decisions pertaining to lower levels of hazard or safety of the chemical

Sensitivity: The proportion of all chemicals considered positive/active that the test correctly classifies as such. It is a measure of accuracy for a test method that produces categorical results, and is an important consideration in assessing the relevance of a test method.

Skin irritation: The production of reversible damage to the skin following the application of a test chemical for up to 4 hours. Skin irritation is a locally arising inflammatory reaction, which appears shortly after stimulation [42]. Its main characteristic is the development of transient inflammatory reactions as evident by clinical signs of irritation: erythema (redness), oedema (swelling), itching and pain.

Specificity: The proportion of all chemicals considered negative/inactive that the test correctly classifies as such. It is a measure of accuracy for a test method that produces categorical results and is an important consideration in assessing the relevance of a test method.

Substance: A reaction product (frequently including multiple molecules as components and/or impurities) obtained from a chemical process. ELINCS defines substance components as molecules present at >10%. Impurities, defined as individual molecules present at <10%, are not listed in ELINCS unless significant contribution is made to the substance classification. Please see above for definitions for Mono-constituent substance and Multi-constituent substances.

Tiered testing strategy: Testing which uses test methods in a sequential manner; the test methods selected in each succeeding level are determined by the results in the previous level of testing.

1038

1039 # REFERENCES

1040-1041 1. EC-ECVAM (2007) Performance Standards for applying human skin models to *in vitro* skin irritation testing. Available under Validation Study Documents, at: [http://ecvam.jrc.ec.europa.eu/].

1042-1045 2. Spielmann H, Hoffmann S, Liebsch M, Botham P, Fentem JH, Eskes C, Roguet R, Cotovio J, Cole T, Worth A, et al. (2007) The ECVAM international validation study on *in vitro* tests for acute skin irritation: report on the validity of the EPISKIN and EpiDerm assays and on the Skin Integrity Function Test. *Altern Lab Anim* **35**, 559-601.

1046-1048 3. United Nations (UN) (2007) Globally Harmonized System of Classification and Labelling of Chemicals (GHS), 2nd revised edition. UN New York and Geneva, 2007. Available at: [http://www.unece.org/trans/danger/publi/ghs/ghs_rev02/02files_e.html/].

1049-1052 4. EC (2001) Commission Directive 2001/59/EC of 6 August 2001 adapting to technical progress for the 28th time Council Directive 67/548/EEC on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances. *Official Journal of the European Union* **L225**, 1-333.

1053-1056 5. EC (2008) REGULATION (EC) No 1272/2008 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006. *Official Journal of the European Union* **L353**, 1-1355.

1057-1060 6. Griesinger C, Barroso J & Zuang V (2008) ECVAM background document on the recent adaptations of the ECVAM performance standards for *in vitro* skin irritation testing in the context of the drafting process of an EU Test Method and an OECD draft Test Guideline. Ispra, November 13, 2008. Available at: [http://ecvam.jrc.ec.europa.eu/].

1061-1063 7. EC-ECVAM (2007) Statement on the validity of *in-vitro* tests for skin irritation, issued by the ECVAM Scientific Advisory Committee (ESAC26), 27 April 2007. Available under Publications, at: [http://ecvam.jrc.ec.europa.eu/].

1064-1066 8. EC-ECVAM (2008) Statement on the scientific validity of *in-vitro* tests for skin irritation testing, issued by the ECVAM Scientific Advisory Committee (ESAC29), 5 November 2008. Available under Publications, at: [http://ecvam.jrc.ec.europa.eu/].

1067-1071 9. EC-ECVAM (2009) Statement on the performance under UN GHS of three *in-vitro* assays for skin irritation testing and the adaptation of the Reference Chemicals and Defined Accuracy Values of the ECVAM skin irritation Performance Standards, issued by the ECVAM Scientific Advisory Committee (ESAC30), 9 April 2009. Available under Publications, at: [http://ecvam.jrc.ec.europa.eu/].

1072-1074 10. Hartung T, Bremer S, Casati S, Coecke S, Corvi R, Fortaner S, Gribaldo L, Halder M, Hoffmann S, Roi AJ, et al. (2004) A modular approach to the ECVAM principles on test validity. *Altern Lab Anim* **32**, 467-472.

1075-1077 11. OECD (2005) OECD series on testing and assessment no. 34. Guidance Document on the validation and international acceptance of new or updated test methods for hazard assessment. Available at: [http://www.oecd.org/document/30/0,3343,en_2649_34377_1916638_1_1_1_1,00.html/].

1078-1080 12. OECD (2004) Test Guideline 404. OECD Guideline for the Testing of Chemicals. Acute dermal irritation/corrosion. Updated Guideline, adopted April 24, 2002. Available at: [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html/].

1081 13. Cotovio J, Grandidier M-H, Portes P, Roguet R & Rubinstenn G (2005) The *in vitro* acute skin
1082 irritation of chemicals: optimisation of the EPISKIN prediction model within the framework of the
1083 ECVAM validation process. *Altern Lab Anim* **33**, 329-349.

1084 14. Eskes C, Cole T, Hoffmann S, Worth A, Cockshott A, Gerner I & Zuang V (2007) The ECVAM
1085 international validation study on *in vitro* tests for acute skin irritation: selection of test chemicals.
1086 *Altern Lab Anim* **35**, 603-619.

1087 15. Fentem JH, Briggs D, Chesné C, Elliott GR, Harbell JW, Heylings JR, Portes P, Roguet R, van de
1088 Sandt JJ & Botham PA (2001) A prevalidation study on *in vitro* tests for acute skin irritation.
1089 Results and evaluation by the Management Team. *Toxicol In Vitro* **15**, 57-93.

1090 16. Hoffmann S (2006) ECVAM Skin Irritation Validation Study Phase II: Analysis of the primary
1091 endpoint MTT and the secondary endpoint IL-1α. 135 pp. + annexes. Available under Validation
1092 Study Documents, at: [http://ecvam.jrc.ec.europa.eu/].

1093 17. Kandárová H, Liebsch M, Genschow E, Gerner I, Traue D, Slawik B & Spielmann H (2004)
1094 Optimisation of the EpiDerm test protocol for the upcoming ECVAM validation study on *in vitro*
1095 skin irritation tests. *ALTEX* **21**, 107-114.

1096 18. Kandárová H, Liebsch M, Gerner I, Schmidt E, Genschow E, Traue D & Spielmann H (2005) The
1097 EpiDerm test protocol for the upcoming ECVAM validation study on *in vitro* skin irritation tests -
1098 An assessment of the performance of the optimised test. *Altern Lab Anim* **33**, 351-367.

1099 19. Portes P, Grandidier M-H, Cohen C & Roguet R (2002) Refinement of the Episkin protocol for the
1100 assessment of acute skin irritation of chemicals: follow-up to the ECVAM prevalidation study.
1101 *Toxicol In Vitro* **16**, 765-770.

1102 20. Zuang V, Balls M, Botham PA, Coquette A, Corsini E, Curren RD, Elliott GR, Fentem JH,
1103 Heylings JR, Liebsch M, et al. (2002) Follow-up to the ECVAM prevalidation study on *in vitro*
1104 tests for acute skin irritation. The European Centre for the Validation of Alternative Methods Skin
1105 Irritation Task Force report 2. *Altern Lab Anim* **30**, 109-129.

1106 21. OECD (2004) Test Guideline 431. OECD Guideline for the Testing of Chemicals. *In vitro* skin
1107 model. Adopted April 13, 2004. Available at:
1108 [http://www.oecd.org/document/22/0,2340,en_2649_34377_1916054_1_1_1_1,00.html/].

1109 22. EpiSkin™ SOP, Version 1.8 (February 2009) ECVAM Skin Irritation Validation Study: Validation
1110 of the EpiSkin™ test method [15 min - 42 hours] for the prediction of acute skin irritation of chemicals.
1111 Available under Validation Study Documents, at: [http://ecvam.jrc.ec.europa.eu/].

1112 23. SkinEthic™ RHE SOP, Version 2.0 (February 2009) SkinEthic skin irritation test[-42bis] test method
1113 for the prediction of acute skin irritation of chemicals: 42 minutes application + 42 hours post-
1114 incubation. Available under Validation Study Documents, at: [http://ecvam.jrc.ec.europa.eu/].

1115 24. EpiDerm™ SOP, Version 7.0 (Revised March 2009) Protocol for: *In vitro* EpiDerm™ skin
1116 irritation test (EPI-200-SIT). For use with MatTek Corporation's reconstructed human epidermal
1117 model EpiDerm (EPI-200). Available under Validation Study Documents, at:
1118 [http://ecvam.jrc.ec.europa.eu/].

1119 25. Cannon CL, Neal PJ, Southee JA, Kubilus J & Klausner M (1994) New epidermal model for dermal
1120 irritancy testing. *Toxicol In Vitro* **8**, 889-891.

1121 26. Parenteau NL, Bilbo P, Nolte CJ, Mason VS & Rosenberg M (1992) The organotypic culture of
1122 human skin keratinocytes and fibroblasts to achieve form and function. *Cytotechnology* **9**, 163-171.

1123 27. Ponec M (2002) Skin constructs for replacement of skin tissues for *in vitro* testing. *Adv Drug Deliv*
1124 *Rev* **54 Suppl 1**, S19-30.

1125 28. Ponec M, Boelsma E, Gibbs S & Mommaas M (2002) Characterization of reconstructed skin
1126 models. *Skin Pharmacol Appl Skin Physiol* **15 Suppl 1**, 4-17.

1127 29. Ponec M, Boelsma E, Weerheim A, Mulder A, Bouwstra J & Mommaas M (2000) Lipid and
1128 ultrastructural characterization of reconstructed skin models. *Int J Pharm* **203**, 211-225.

1129 30. Poumay Y, Boucher F, Leclercq-Smekens M, Degen A & Leloup R (1993) Basal cell adhesion to a
1130 culture substratum controls the polarized spatial organization of human epidermal keratinocytes into
1131 proliferating basal and terminally differentiating suprabasal populations. *Epithelial Cell Biol* **2**, 7-16.

1132 31. Pruniéras M, Regnier M & Woodley D (1983) Methods for cultivation of keratinocytes with an air-
1133 liquid interface. *J Invest Dermatol* **81**, 28s-33s.

1134 32. Rosdy M & Clauss LC (1990) Terminal epidermal differentiation of human keratinocytes grown in
1135 chemically defined medium on inert filter substrates at the air-liquid interface. *J Invest Dermatol* **95**,
1136 409-414.

1137 33. Tinois E, Gaetani Q, Gavraud B, Dupond D, Rougier A & Pouradier-Dupteil X (1994) The
1138 EPISKIN model: Successful reconstruction of human epidermis *in vitro*. In: *In Vitro Skin*
1139 *Toxicology* (Rougier A, Goldberg M & Maibach HI, eds), pp. 133-140. Liebert M. A. Inc.
1140 Publishers NY, New York.

1141 34. Wilkins LM, Watson SR, Prosky SJ, Meunier SF & Parenteau NL (1994) Development of a
1142 bilayered living skin construct for clinical applications. *Biotechnol Bioeng* **43**, 747-756.

1143 35. OECD (2004) OECD series on principles of Good Laboratory Practice and compliance monitoring
1144 number 14: Advisory document of the Working Group on Good Laboratory Practice. The
1145 application of the principles of GLP to *in vitro* studies. Available at:
1146 [http://www.olis.oecd.org/olis/2004doc.nsf/LinkTo/NT00008FEE/$FILE/JT00174939.PDF/].

1147 36. Mosmann T (1983) Rapid colorimetric assay for cellular growth and survival: application to
1148 proliferation and cytotoxicity assays. *J Immunol Methods* **65**, 55-63.

1149 37. Basketter DA, York M, McFadden JP & Robinson MK (2004) Determination of skin irritation
1150 potential in the human 4-h patch test. *Contact Dermatitis* **51**, 1-4.

1151 38. Hoffmann S, Saliner AG, Patlewicz G, Eskes C, Zuang V & Worth AP (2008) A feasibility study
1152 developing an integrated testing strategy assessing skin irritation potential of chemicals. *Toxicol*
1153 *Lett* **180**, 9-20.

1154 39. Jirova D, Liebsch M, Basketter DA, Spiller E, Kejlova K, Bendova H, Marriott M & Kandarova H
1155 (2007) Comparison of human skin irritation and photo-irritation patch test data with cellular *in vitro*
1156 assays and animal *in vivo* data. *AATEX, Special Issue-proceedings from WC6* **14**, 359-365.

1157 40. Williams IR & Kupper TS (1996) Immunity at the surface: homeostatic mechanisms of the skin
1158 immune system. *Life Sci* **58**, 1485-1507.

1159 41. EC (2006) REGULATION (EC) No 1907/2006 OF THE EUROPEAN PARLIAMENT AND OF
1160 THE COUNCIL of 18 December 2006 concerning the Registration, Evaluation, Authorisation and
1161 Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending
1162 Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission
1163 Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives
1164 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union*
1165 **L396/1**.

1166 42. Harvell JD, Lamminstausta K & Maibach HI (1995) Irritant contact dermatitis. In: *Practical*
1167 *Contact Dermatitis* (Guin JD, ed), pp. 7-18. Mc Graw-Hill, New York.